



ORBIT - Online Repository of Birkbeck Institutional Theses

Enabling Open Access to Birkbeck's Research Degree output

On the role of deduction in reasoning from uncertain premises

<https://eprints.bbk.ac.uk/id/eprint/40349/>

Version: Full Version

Citation: Cruz de Echeverria Loebell, Nicole (2018) On the role of deduction in reasoning from uncertain premises. [Thesis] (Unpublished)

© 2020 The Author(s)

All material available through ORBIT is protected by intellectual property law, including copyright law.

Any use made of the contents should comply with the relevant law.

[Deposit Guide](#)
Contact: [email](#)



École Pratique
des Hautes Études



On the role of deduction in reasoning from uncertain premises

Nicole Cruz de Echeverria Loebell
Ph.D./doctoral thesis

Written in the context of a co-tutorship/joint degree at Birkbeck, University of London, London, UK, and École Pratique des Hautes Études, PSL, Paris, France

Principal supervisor at Birkbeck, University of London: Prof. Mike Oaksford
Principal supervisor at EPHE, PSL: Maître de Conférences HDR Jean Baratgin
External supervisor: Prof. David Over

Examiners for Birkbeck, University of London:
Internal: Prof. Dorothy Edgington, Birkbeck, University of London
External: Prof. Nick Chater, University of Warwick

Examiners for EPHE, PSL:
Prof. Shira Elqayam, De Montfort University
Directeur de recherche du CNRS Wim De Neys, Université Paris Descartes

Abstract

(French version follows below)

The probabilistic approach to reasoning hypothesizes that most reasoning, both in everyday life and in science, takes place in contexts of uncertainty. The central deductive concepts of classical logic, consistency and validity, can be generalised to cover uncertain degrees of belief. Binary consistency can be generalised to coherence, where the probability judgments for two statements are coherent if and only if they respect the axioms of probability theory. Binary validity can be generalised to probabilistic validity (p-validity), where an inference is p-valid if and only if the uncertainty of its conclusion cannot be coherently greater than the sum of the uncertainties of its premises. But the fact that this generalisation is possible in formal logic does not imply that people will use deduction in a probabilistic way. The role of deduction in reasoning from uncertain premises was investigated across ten experiments and 23 inferences of differing complexity. The results provide evidence that coherence and p-validity are not just abstract formalisms, but that people follow the normative constraints set by them in their reasoning. It made no qualitative difference whether the premises were certain or uncertain, but certainty could be interpreted as the endpoint of a common scale for degrees of belief. The findings are evidence for the descriptive adequacy of coherence and p-validity as computational level principles for reasoning. They have implications for the interpretation of past findings on the roles of deduction and degrees of belief. And they offer a perspective for generating new research hypotheses in the interface between deductive and inductive reasoning.

Keywords: Reasoning; deduction; probabilistic approach; coherence; p-validity

Résumé

L'approche probabiliste du raisonnement émet l'hypothèse que la plupart des raisonnements, aussi bien dans la vie quotidienne qu'en science, se réalisent dans des contextes d'incertitude. Les concepts déductifs centraux de la logique classique, *consistance* et validité, peuvent être généralisés afin d'englober des degrés de croyance incertains. La *consistance* binaire peut être généralisée à travers la dénomination de cohérence, lorsque les jugements de probabilité à deux affirmations sont cohérents seulement s'ils respectent les axiomes de la théorie de la probabilité. La validité binaire peut se généraliser comme validité probabiliste (validité-p), lorsqu'une inférence est valide-p seulement si l'incertitude de sa conclusion ne peut être de façon cohérente plus grande que la somme des incertitudes de ses prémisses. Cependant le fait que cette généralisation soit possible dans une logique formelle n'implique pas le fait que les gens utilisent la déduction de manière probabiliste. Le rôle de la déduction dans le raisonnement à partir de prémisses incertaines a été étudié à travers dix expériences et 23 inférences de complexités différentes. Les résultats mettent en évidence le fait que la cohérence et la validité-p ne sont pas juste des formalismes abstraits, mais que les gens vont suivre les contraintes normatives établies par eux dans leur raisonnement. Que les prémisses soient certaines ou incertaines n'a pas créé de différence qualitative, mais la certitude pourrait être interprétée comme l'aboutissement d'une échelle commune de degrés de croyance. Les observations sont la preuve de la pertinence descriptive de la cohérence et de la validité-p comme principes de niveau de calcul pour le raisonnement. Ils ont des implications pour l'interprétation d'observations antérieures sur les rôles de la déduction et des degrés de croyance. Enfin, ils offrent une perspective pour générer de nouvelles hypothèses de recherche quant à l'interface entre raisonnement déductif et inductif.

Mots clés: Raisonnement; déduction; approche probabiliste; cohérence; validité-p

RÉSUMÉ LONG

(English version follows below)

Contexte théorique

L'approche probabiliste du raisonnement

L'approche probabiliste en psychologie du raisonnement (Evans, 2006; Evans & Stanovich, 2013; Oaksford & Chater, 2007, 2013; Over, 2016; Pfeifer & Kleiter, 2009, 2010; Politzer & Baratgin, 2016) est basée sur l'hypothèse fondamentale que la plupart des raisonnements, aussi bien dans la vie quotidienne qu'en science, se réalisent dans des contextes incertains. L'incertitude, ou degrés de croyance, dans les énoncés à partir desquels nous raisonnons ne peut être modélisée dans la logique classique, qui est binaire dans la mesure où elle représente seulement ce qui est assurément vrai ou faux. Cependant cette incertitude peut être modélisée dans la théorie des probabilités. Il y a une abondance de preuves empiriques qui démontrent que les gens tendent à raisonner à partir de prémisses incertaines même lorsqu'il leur ait donné l'instruction d'assumer que les prémisses sont certaines (Byrne, 1989; George, 1997; Liu, Lo, & Wu, 1996; Stevenson & Over, 1995; Thompson, 1994).

Dans la logique classique, un conditionnel, *si p alors q*, est interprété comme le *conditionnel matériel*, qui est équivalent à *non-p ou q*. Mais dans l'approche probabiliste cette vision est rejetée, et il est à la place proposée une interprétation probabiliste des conditionnels. La probabilité d'un conditionnel $P(\text{si } p \text{ alors } q)$, n'est pas proposée pour correspondre à $P(\text{non-}p \text{ ou } q)$, mais à la probabilité conditionnelle de *q considérant p*, $P(q|p)$. Cette relation, $P(\text{si } p \text{ alors } q) = P(q|p)$, est appelée l'Équation (Edgington, 1995). Avoir un degré de croyance dans un conditionnel est considéré différent que d'avoir un degré de croyance concernant un fait dans le monde. Le projet est que les gens arrivent à $P(q|p)$, non pas en ayant d'abord des degrés spécifiques de croyance en $P(p \ \& \ q)$ et en $P(p)$, prenant ensuite leur ratio, $P(q|p) = P(p \ \& \ q)/P(p)$, mais à travers un *test de Ramsey* (Ramsey, 1929/1990; Stalnaker, 1968). Il s'agit d'une simulation mentale dans laquelle les gens supposent hypothétiquement l'antécédent *p* du conditionnel, réalisent les changements nécessaires quant à leurs croyances pour préserver la *consistance*, et évaluent la probabilité de *q* selon cette supposition.

L'interprétation conditionnelle matérielle de conditionnels entraîne des conséquences contre-intuitives qui sont évitées par l'Equation. On peut prendre en exemple le conditionnel « Si l'on lance cette pièce non biaisée elle va atterrir côté face ». Intuitivement, la probabilité de ce conditionnel est .5, qui correspond à la probabilité conditionnelle pour la pièce non biaisée d'atterrir côté face si elle est lancée. Mais si l'on imagine que la pièce est dans une boîte en verre dans un musée, nous sommes alors presque certains que nous n'allons pas la lancer. Le conditionnel matériel est vrai dès que l'antécédent du conditionnel est faux. De ce fait, le plus

confiant l'on devient que nous n'allons pas jeter la pièce, le plus probable il sera que si on la lance elle atterrisse côté face, selon l'interprétation conditionnelle matérielle. Mais cela paraît absurde.

Un conditionnel basé sur l'Équation a été appelé le *conditionnel probabilité* (Adams, 1998), *conditionnel suppositionnel* (Edgington, 2014), et *événement conditionnel* (de Finetti, 1937/1980). Il y est fait référence dans cette thèse en tant que conditionnel probabilité. L'hypothèse selon laquelle l'interprétation modale des conditionnels qu'ont les gens correspond au conditionnel probabilité a reçu un large soutien empirique (Barrouillet & Gauffroy, 2015; Evans, Handley, & Over, 2003; Fugard, Pfeifer, Mayerhofer, & Kleiter, 2011; Oberauer & Wilhelm, 2003; Politzer, Over, & Baratgin, 2010).

Le rôle de la déduction dans le raisonnement à partir de prémisses incertaines

Avec l'avènement du paradigme probabiliste, la question du rôle, si elle en joue un, de la déduction dans le raisonnement, s'est posée à nouveau (Evans, 2012; Evans & Over, 2013). Cette thèse soutient que la réponse à cette question dépend en partie de la manière de définir la déduction. La définition de déduction dans la logique classique est binaire, et produit ce que de Finetti appelle « la logique de la certitude » (de Finetti, 1972; Elqayam & Over, 2013). Quand les prémisses sont présumées certaines, la conclusion d'une inférence valide classiquement doit être certaine également. Mais imaginons que les gens aient des degrés incertains de croyance dans les prémisses d'une inférence. Quel degré de croyance serait raisonnable qu'ils aient en la conclusion ? La logique classique ne peut pas répondre à cette question, du fait que cela ne peut s'appliquer au raisonnement en contexte d'incertitude. Si l'on retient sa définition binaire de la déduction, l'hypothèse centrale de l'approche probabiliste, selon laquelle la plupart des raisonnements se réalisent à partir de prémisses incertaines, implique que la déduction a seulement un petit rôle à jouer dans la plupart des situations de raisonnement (Oaksford & Chater, 2007).

Utilisant la définition classique de la déduction, un certain nombre d'études empiriques ont trouvé un effet de la déduction dans les tâches de raisonnement, en plus de ce qui a été appelé un effet de « croyance » (Evans, Handley, Neilens, & Over, 2010; Klauer, Beller, & Hütter, 2010; Markovits, Brisson, & Chantal, 2015; Rips, 2001; Singmann & Klauer, 2011; Thompson, 1994; Trippas, Thompson, & Handley, 2017). En particulier, les gens acceptent plus souvent la conclusion d'inférences valides qu'invalides, en plus d'accepter des conclusions vraisemblables plus fréquemment que celles invraisemblables. Ce dernier effet cité a été appelé *biais de la croyance*, constituant un écart par rapport à la réponse normative selon les instructions logiques binaires.

Les observations d'effets à la fois de « logique » et de « croyance » dans les raisonnements a contribué au développement de théories à deux composantes (Evans, 2006; Klauer et al., 2010; Markovits et al., 2015; Verschueren, Schaeken, & d'Ydewalle, 2005).

Quand ces théories adoptent un processus binaire logique qui est contrasté par un processus « basé sur la croyance », le processus « logique » interroge l'hypothèse basique de l'approche probabiliste selon laquelle la plupart des raisonnements dans le monde réel se réalisent à partir de prémisses incertaines.

La thèse ci-présente plaide pour une manière alternative d'expliquer les observations sur un effet de la déduction qui n'interroge pas la base de l'approche probabiliste. Plutôt que d'intégrer une composante logique binaire d'un côté et une composante probabiliste basée sur la croyance de l'autre, au sein d'une perspective à double composante, elle explore l'intégration de la logique et de la probabilité elle-même.

Les concepts déductifs centraux de la logique classique, *consistance* et validité, peuvent être généralisés pour englober les degrés de croyance. La *consistance* binaire peut être généralisée à travers la dénomination de *cohérence*, et la validité binaire à travers la validité probabiliste, ou validité-p (Adams, 1998; Coletti & Scozzafava, 2002; Gilio, 2002). Deux énoncés sont cohérents seulement s'ils peuvent être vrais tous les deux au même moment. De manière similaire, les jugements de probabilité sont cohérents seulement s'ils respectent les axiomes de la théorie de la probabilité. Par exemple, si l'on croit qu'il y a 80% de chances qu'il pleuve aujourd'hui, alors pour que notre supposition soit cohérente il faudrait aussi que l'on soit prêt à croire qu'il y a 20% de chances qu'il ne pleuve pas aujourd'hui, sinon les probabilités n'attendraient pas 1 en s'additionnant. Une inférence est valide seulement s'il est incohérent pour ses prémisses d'être vraies mais ses conclusions fausses. Cela signifie qu'une inférence valide maintient la vérité des prémisses aux conclusions. Pour comprendre comment cette définition peut être généralisée aux degrés de croyance, l'on prend le cas simple d'une inférence à une prémisse. Une inférence à une prémisse est valide-p seulement s'il est incohérent pour la probabilité de sa conclusion d'être inférieure à la probabilité de sa prémisse. De ce fait une inférence valide-p maintient la probabilité, tout comme une inférence binaire valide maintient la vérité. Les généralisations ci-dessus rendent possible l'étude de la déduction à partir de prémisses incertaines (Stevenson & Over, 1995), de manière à ce qu'il n'y ait pas besoin de qualifier la base de l'approche probabiliste.

Avec cette notion généralisée de la déduction, beaucoup d'observations sur le biais de croyance peuvent être réinterprétées comme effets de cohérence. Les participants d'une expérience peuvent violer les *instructions* déductives classiques qui assument que les prémisses sont certaines et toutefois raisonner déductivement, en déterminant un degré de croyance cohérent dans la conclusion d'une inférence sur la base de sa forme logique et de leurs degrés incertains de croyance en les prémisses. Ne pas suivre des instructions binaires peut être vu comme une faute, mais ce n'est pas nécessairement une faute logique.

Question de recherche

Le fait que la définition de la déduction puisse être généralisée pour englober les degrés de croyance n'implique pas que les gens utilisent la déduction de manière probabiliste. Il est possible que quand les gens réalisent une déduction ils le fassent selon la voie binaire classique, et que raisonner à partir de prémisses incertaines soit inductif en pratique même si cela n'a pas à être le cas dans la théorie. Le fait que les gens soient sensibles ou non aux contraintes de la cohérence et à la validité-p est une question empirique qui a seulement récemment commencé à être étudiée, mais qui était l'axe principal de cette thèse.

Cette thèse étudiait le rôle de la cohérence et de la validité-p dans le raisonnement incertain à travers dix expériences. Des études antérieures sur la cohérence des jugements de probabilité de conclusion s'étaient centrées sur les syllogismes conditionnels (Evans, Thompson, & Over, 2015; Singmann, Klauer, & Over, 2014). Des réponses à ces inférences à deux prémisses se sont avérées être plus cohérentes que le niveau aléatoire, surtout pour l'inférence de *modus ponens* (MP), et moins solidement pour l'inférence de *déni de l'antécédent* (DA). Cependant, les taux de chance dans ces études étaient souvent très élevés (e.g. Singmann et al., 2014, Figure 3) et variaient entre les différentes inférences, rendant plus difficile de détecter la cohérence supérieure au niveau aléatoire quand elle se présente, et de comparer cette cohérence entre inférences.

Les observations obtenues à travers les expériences

Les expériences ci-présentes ont élargi ces observations, utilisant différentes méthodologies, à travers 23 formes d'inférence. Ces inférences différaient dans la complexité de leur structure. Elles incluaient des inférences valides (déductives) et invalides (inductives), et elles présentaient des négations, des conjonctions, des disjonctions et des conditionnels. La plupart des inférences comprenant des conditionnels les considéraient interprétables comme conditionnels de probabilité, se basant sur l'Équation $P(\text{si } p \text{ alors } q) = P(q|p)$ (Adams, 1998; Edgington, 1995; Jeffrey, 1991), mais certaines expériences ont inclus des comparaisons de différentes interprétations de conditionnels.

Réponses cohérentes au MT

Pour ce qui est des syllogismes conditionnels, les résultats confirmaient l'observation antérieure de cohérence supérieure au niveau aléatoire pour le MP, mais ont fait le constat d'une cohérence fortement supérieure au niveau aléatoire également pour l'inférence de *modus tollens* (MT). Des six expériences à travers lesquelles les deux inférences ont été étudiées, la cohérence globale était supérieure au niveau aléatoire dans tous les cas pour le MP. Pour le MT, la cohérence était supérieure au niveau aléatoire dans tous les cas si ce n'est

dans deux d'entre eux. Le premier était l'Expérience 3, dans laquelle elle était au niveau aléatoire dans la tâche des énoncés mais supérieure à ce niveau dans la tâche des inférences. Dans l'Expérience 4, qui répliquait l'Expérience de lab 3 sur internet, la cohérence était supérieure au niveau aléatoire dans certaines conditions. Le deuxième cas se trouvait dans l'Expérience 7. La cohérence pour le MT était au niveau aléatoire quand il était donné aux participants des instructions de paradigme binaire pour présumer du caractère véridique des prémisses, et ensuite juger si la conclusion aussi se doit de l'être. Cela était malgré le fait que la cohérence pour le MT était supérieure au niveau aléatoire pour les mêmes éléments quand été données des instructions probabilistes (Expérience 6).

Ces découvertes constituent la preuve solide d'une sensibilité de base aux contraintes de la cohérence pour le MT. Elles fournissent aussi directement une preuve de l'adéquation descriptive de la proposition selon laquelle les contraintes de la déduction ne sont pas limitées au cas spécial dans lequel il est présumé que les prémisses sont certaines.

La méthodologie des Expériences 5 à 7 a permis, non seulement d'évaluer si les jugements de probabilité de conclusion sont cohérents au-dessus du niveau aléatoire, mais aussi de réaliser des comparaisons quantitatives de cohérence supérieure au niveau aléatoire entre inférences. Cela a été réalisé à travers un design qui assimilait le taux de hasard de la cohérence à travers les inférences et les conditions. Les expériences ont révélé que malgré le fait que les réponses au MT étaient généralement cohérentes au-dessus du niveau aléatoire, la cohérence était plus faible pour le MT que pour le MP, conformément à la littérature utilisant les instructions de paradigme binaire. Dans l'approche probabiliste, la différence dans l'acceptation des deux inférences peut être expliquée comme étant le résultat d'un raisonnement dynamique (Oaksford & Chater, 2013). Plus spécifiquement, le MT peut parfois être vu comme une instance d'une inférence *reductio ad absurdum*. La prémisse catégorique *non-q* nie un élément de la prémisse conditionnelle *si p alors q*, avec le résultat selon lequel un fort degré de croyance dans les deux prémisses est incompatible avec un fort degré de croyance dans l'élément *p* de la conclusion. Mais le conflit entre *si p alors q*, *non-q*, et *p* peut être résolu de différentes manières. Dépendant du contexte personnel de croyances, une personne pourrait voir *si p alors q* et *non-q* comme une raison pour ne pas croire *p*, conformément au MT, alors qu'une autre personne pourrait voir *non-q* et *p* comme une raison pour ne pas croire *si p alors q*, se livrant au raisonnement dynamique. La logique elle-même ne nous dit pas quel énoncé devrait être écarté, de sorte que sans instructions pour présumer que les deux prémisses sont certaines, le choix peut être fait sur des fondements inductifs. Cette perspective permet d'expliquer l'asymétrie entre le MP et le MT de manière rationnelle.

Réponses changeantes à l'AC et à la NA

Les résultats de cohérence pour les syllogismes conditionnels *affirmation du conséquent* (AC) et *négation de l'antécédent* (NA) étaient plus équivoques. Dans l'Expérience 3, la cohérence

pour les deux inférences était supérieure au niveau aléatoire dans la tâche des énoncés, mais pas dans la tâche des inférences. Dans l'Expérience 4 la cohérence était supérieure au niveau aléatoire pour les deux inférences ainsi que pour les deux conditions de tâche. Toutefois, la cohérence pour les deux inférences était inférieure au niveau aléatoire dans l'Expérience 5, et au niveau aléatoire dans les Expériences 6 et 7. Dans l'Expérience 8, la cohérence était supérieure au niveau aléatoire pour la NA (cette expérience n'incluait pas l'AC). Laissant de côté les tâches d'énoncés des Expériences 3 et 4, cela signifie que la cohérence pour l'AC était supérieure au niveau aléatoire dans une expérience sur cinq, et la cohérence pour la NA était supérieure au niveau aléatoire dans deux expériences sur six. Un manque de cohérence supérieure au niveau aléatoire pour ces deux inférences est difficile à interpréter, mais il pourrait être expliqué par une interprétation biconditionnelle de la prémisse conditionnelle. Plus spécifiquement, les matériaux utilisés peuvent avoir suggéré aux participants que l'antécédent et le conséquent du conditionnel étaient corrélés positivement. Conformément à cela, une analyse des données de l'Expérience 3 a montré que la cohérence était supérieure au niveau aléatoire dans l'hypothèse d'une interprétation biconditionnelle pour les mêmes réponses et que la cohérence était au niveau aléatoire dans l'hypothèse d'une interprétation conditionnelle. Pour pouvoir interpréter les résultats de cohérence pour ces inférences, les deux interprétations devraient être dissociées, contrôlant explicitement la corrélation entre antécédent et conséquent. Bien qu'une telle expérience ait été hors de la portée de cette thèse, le fait qu'il soit nécessaire d'interpréter les observations sur l'AC et la NA est néanmoins une perspective plus précise et moins pessimiste que la suggestion d'études antérieures selon laquelle les réponses sont généralement incohérentes pour ces inférences (e.g. Singmann et al., 2014).

Des études supplémentaires sur l'AC et la NA, mais aussi sur le MP, le MT, et d'autres inférences à deux prémisses, pourraient évaluer dans quelle mesure les réponses incohérentes sont des réponses qui surestiment ou sous-estiment la probabilité de la conclusion, au vu des probabilités attribuées aux prémisses, et à quel point cela dépend des risques et bénéfices suggérés par les matériaux (Oberauer & Wilhelm, 2003). L'hypothèse selon laquelle il peut y avoir une confiance excessive en la conclusion d'une inférence valide ne pouvait même pas être formulée dans le paradigme binaire. Mais cette hypothèse peut prolonger l'étude des effets de suppression, rendant nécessaire la distinction entre la suppression de la conclusion d'une inférence, et la suppression de l'inférence elle-même (Over & Cruz, 2018).

Conditionnels, introduction du ou, et biais de la conjonction

Les Expériences 1 et 2 n'ont pas seulement prolongé les résultats de cohérence à d'autres inférences, mais ont aussi fourni des informations sur les sens des énoncés composants, et sur les facteurs impliqués dans le raisonnement avec eux. Une analyse de l'inférence *ou-à-si* a démontré que les réponses des gens étaient cohérentes au-dessus du niveau aléatoire dans

l'hypothèse où ils interprètent le conditionnel dans la conclusion comme le conditionnel probabilité, mais ils étaient cohérentes au-dessous du niveau aléatoire dans l'hypothèse où ils interprètent le conditionnel dans la conclusion comme le conditionnel matériel, qui est équivalent à *non-p ou q*. Cela constitue une forme de preuve novatrice pour l'interprétation des conditionnels en termes de l'Équation.

La cohérence était sérieusement supérieure au niveau aléatoire à travers quatre variantes de l'inférence d'introduction du *ou*, suggérant que malgré le fait que cela puisse être pragmatiquement malheureux, dans le cadre d'instructions de paradigme binaire, d'indiquer *p ou q* quand l'on pourrait être plus informatif et précis en indiquant *p*, les gens considèrent aisément l'inférence comme étant valide quand ils sont directement interrogés sur leur degrés de croyance quant à la prémisse et la conclusion. Cela est en conformité avec l'idée que les plus bas taux d'adhésion trouvés pour l'inférence dans le cadre d'instructions de paradigme binaire sont dus à des effets pragmatiques (Cruz, Over, & Oaksford, 2017; Orenes & Johnson-Laird, 2012), contrairement à la récente proposition d'une révision de la théorie des modèles mentaux où l'inférence est en fait invalide (Johnson-Laird, Khemlani, & Goodwin, 2015).

De plus, l'Expérience 2 a montré que malgré le fait que les réponses étaient cohérentes pour *et-élimination* lorsque été utilisés des matériaux neutres (voir aussi Politzer & Baratgin, 2016, pour une concordance des preuves), la cohérence pour la même inférence s'effondrait lorsque été utilisés les matériaux connus pour causer le biais de la conjonction (Tversky & Kahneman, 1983). Cela en dépit de la tâche visiblement transparente de déduire la probabilité de *p* de la probabilité de *p & q*. Les observations soulignent la force, mais également la portée limitée, de l'erreur.

Comparer la cohérence supérieure au niveau aléatoire entre inférences

Il a été souligné dans la thèse qu'il est difficile de faire des comparaisons quantitatives de cohérence supérieure au niveau aléatoire entre inférences quand les taux de hasard des inférences ne sont pas équivalents. Cela est dû au fait que le taux de hasard est soustrait au taux observé de cohérence pour obtenir la mesure de la cohérence supérieure au niveau aléatoire. Plus le taux de hasard est élevé, plus la probabilité de détecter la cohérence supérieure au niveau aléatoire quand elle est présente est faible, c'est-à-dire plus faible est la sensibilité du test pour la cohérence supérieure au niveau aléatoire. Le taux de hasard correspond à la largeur de l'intervalle de cohérence, qui à son tour dépend de la forme de l'inférence, de la validité ou non de l'inférence, et des probabilités de la prémisse. Il est possible d'étudier s'il y a ou non cohérence supérieure au niveau aléatoire pour différentes formes d'inférences sans maintenir constant le taux de hasard (Cruz, Baratgin, Oaksford, & Over, 2015; Evans et al., 2015; Singmann et al., 2014), mais des comparaisons supplémentaires paraissent difficiles à interpréter.

Dans les Expériences 5 à 7, la cohérence supérieure au niveau aléatoire s'est révélée plus comparable, à travers 12 inférences, en utilisant deux méthodes. Dans la première, il a été donné aux participants une série de probabilités de prémisse, chacune d'entre elle présentée avec un nombre de probabilités de conclusion. La tâche à réaliser était de donner une réponse binaire quant à la possibilité d'avoir une probabilité de conclusion binaire, au vu des probabilités de prémisse. Dans la deuxième méthode, il a été donné à nouveau aux participants une série de probabilités de prémisse, et ils étaient interrogés sur le fait que la probabilité de la conclusion puisse ou non être supérieure, ainsi que sur le fait qu'elle puisse ou non être inférieure, à la probabilité de prémisse (dans le cas d'inférences à une prémisse) ou à 50% (dans le cas d'inférences à deux prémisses). Ces formats de réponse binaire ont établi le taux de hasard d'une réponse cohérente à 50% à travers toutes les inférences.

En utilisant ces méthodes, il a été découvert que l'inférence pour laquelle la cohérence supérieure au niveau aléatoire était la plus élevée était la contradiction *non de morgan*, suivie de près par le MP. Excepté pour l'AC et la Na dont nous avons parlé, la cohérence pour le reste des inférences était moins élevée mais toujours supérieure au niveau aléatoire dans l'ensemble. Deux intrus ont été les inférences à partir de *ou-à-si* et à partir de *si-à-ou*. La cohérence était au niveau aléatoire pour *ou-à-si* dans l'Expérience 5, pour *si-à-ou* dans l'Expérience 6, et pour les deux dans l'Expérience 7 (qui utilisait des instructions de paradigme binaire). Si l'on laisse de côté les tâches d'énoncés des Expériences 3 et 4, et que l'on considère seulement les tâches d'inférences dans ces expériences, cela signifie que la cohérence pour *ou-à-si* et *si-à-ou* était supérieure au niveau aléatoire dans quatre expériences sur six. Ces inférences comprennent des négations au début de la prémisse ou de la conclusion, ayant pu les rendre plus difficiles à traiter. Les réponses aux deux inférences étaient supérieures au niveau aléatoire quand elles étaient mesurées à travers les positions de la négation dans l'Expérience 1, mais des études supplémentaires seraient nécessaire pour établir la raison pour laquelle la cohérence était quelque peu moins fiable pour ces inférences.

Une étape supplémentaire serait aussi d'examiner plus en détail ce qui fait que les inférences de non de morgan et de MP se distinguent en termes de taux élevés de réponses cohérentes. Par exemple, il pourrait être testé si les contradictions en général sont détectées plus aisément que les autres relations logiques, ou s'il s'agit de quelque chose spécifique à la négation de non de morgan.

De manière générale, l'observation que, à travers l'étude des inférences, la cohérence était supérieure au niveau aléatoire dans la grande majorité des cas, et que les échecs de cohérence étaient l'exception plus que la règle, apporte un appui supplémentaire solide à l'hypothèse que les gens sont sensibles, au même degré, aux contraintes de la cohérence plus qu'aux contraintes de la *consistance*.

L'effet d'une tâche d'inférence explicite et la mémoire de travail

Les expériences 3 et 4 ont prolongé les résultats d' Evans et al. (2015) sur le rôle d'une tâche d'inférence explicite pour la cohérence. De manière assez surprenante, la cohérence était déjà supérieure au niveau aléatoire dans la grande majorité des cas dans les tâches d'énoncés : quand il était donné aux gens les énoncés qui composaient les inférences dans un ordre aléatoire une par une sur l'écran. Une tâche d'inférence explicite, dans laquelle les énoncés étaient disposés dans des inférences, et les inférences présentées une par une sur l'écran, a eu tendance à augmenter la cohérence dans le petit nombre de cas pour lesquels elle n'était pas déjà supérieure au niveau aléatoire. Toutefois, une exception s'est présentée dans l'Expérience 3 pour l'AC et la NA, où la cohérence était supérieure au niveau aléatoire dans les tâches d'énoncés mais au niveau aléatoire dans les tâches d'inférences. Cela peut être dû au fait que dans certains cas, les facteurs pragmatiques liés à l'assertabilité d'un énoncé comme conclusion tirée d'autres énoncés, ou à la relation entre énoncés qui sont difficiles à intégrer du fait de la présence de négations, peut donner lieu à des réductions de cohérence qui ne se manifesteraient pas quand les énoncés sont considérés isolément. Une autre possibilité est liée au fait que mettre les conditionnels dans une tâche d'inférence explicite AC ou NA a tendance à renforcer l'interprétation biconditionnelle, puisque sinon les inférences sont invalides. Mais l'observation réalisée pourrait aussi simplement refléter le fait que différents groupes étaient présents dans les tâches d'énoncés et d'inférences, et certains peuvent avoir interprété les matériaux pour l'AC et la NA comme impliquant une corrélation entre antécédant et conséquent, alors que d'autres non. D'autres répliques seraient nécessaires pour juger de la fiabilité de cette observation.

Les Expériences 3 et 4 ont aussi inclus une tâche d'inférence avec une charge de mémoire travail. Les réponses en cette condition ont généralement peu différencié de celles de la tâche d'inférence sans charge de mémoire travail. Mais là où elles différaient, la condition de charge était associée à des taux de cohérence plus faibles, suggérant que la différence entre la tâche d'énoncés et celle d'inférences peut être due en partie à la différence de demandes qu'elles présentent quant à la mémoire travail pour calibrer les croyances à travers les énoncés. Toutefois, le faible effet de la condition de charge est conforme à l'observation faite que la cohérence était dans la plupart des cas déjà supérieure au niveau aléatoire dans la tâche d'énoncés, de sorte que la tâche d'inférences avait peu à ajouter qui puisse être perturbé.

La tendance générale suggère que les gens ont peut-être une tendance implicite, spontanée, à établir une cohérence entre croyances, mais que dans des situations dans lesquelles cela échoue, une tâche d'inférence explicite, dans laquelle tous les renseignements pertinents sont disponibles simultanément sur l'écran, et les gens peuvent concentrer leur attention directement sur les relations entre eux, peut se révéler utile. Dans tous les cas, l'inférence explicite est nécessaire lorsque s'établissent des relations entre des matériaux nouveaux pour lesquels aucune croyance n'est encore disponible.

Prémises certaines vs. incertaines, instructions probabilistes vs. de paradigme binaire

Les Expériences 6 et 7 ont permis de comparer la cohérence supérieure au niveau aléatoire pour des inférences aux prémisses certaines vs. aux prémisses incertaines, et pour des inférences avec des instructions probabilistes vs. avec des instructions de paradigme binaire, en utilisant les mêmes inférences, matériaux, et formes de réponses. Il n'y a pas eu de preuve que la cohérence est plus faible quand les prémisses sont incertaines, ou que la cohérence est plus faible quand sont utilisées des instructions probabilistes plutôt que des instructions de paradigme binaire. Cela apporte une preuve nouvelle et solide quant au fait que la déduction à partir de prémisses incertaines est possible, et n'est pas limitée au raisonnement à partir de la certitude. Une vérité certaine et une fausseté certaine n'ont pas semblé être différentes qualitativement des degrés de croyance incertains, mais semblent plutôt être les extrémités sur une échelle commune.

Facteurs sans effet systématique sur la cohérence supérieure au niveau aléatoire

Outre l'observation que la cohérence ne différait pas entre prémisses certaines et incertaines, ou entre instructions probabilistes et de paradigme binaire, les Expériences 3 et 4 n'ont trouvé aucune preuve d'une différence systématique entre les réponses des gens aux tâches de raisonnement étudiées sur l'internet et dans une configuration de labo, rendant plus aisée la généralisation entre elles, ainsi qu'entre les expériences conduites dans cette thèse et les résultats antérieurs d'Evans et al. (2015). De plus, l'Expérience 5 n'a trouvé aucune preuve d'une différence dans la cohérence de réponse en fonction de s'il était demandé aux gens de juger si la conclusion tombait ou non dans l'intervalle de cohérence. A travers les expériences, il n'a également pas semblé avoir de différence systématique dans la cohérence des réponses entre les inférences à une ou à deux prémisses. Les différences de cohérence entre inférences ont plutôt paru être basées sur des facteurs plus spécifiques, comme le fait qu'elles contiennent ou non des négations ou qu'elles puissent être interprétées de différentes manières. Enfin, les expériences n'ont donné aucune preuve du fait que la cohérence diffère entre inférences valides et invalides, i.e. entre inférences déductives et inductives. Ce résultat est sensé puisque les contraintes de cohérence sont valables pour les deux types d'inférences, et les inférences déductives ont seulement des contraintes plus fortes pour les bords inférieurs de leurs limites d'intervalle. Les résultats négatifs mentionnés ci-dessus peuvent aider à interpréter et à préciser les observations positives de ces expériences.

La précision des degrés de croyance des gens

Les intervalles de cohérences sont généralement mesurés en utilisant des probabilités de point, mais il n'y avait aucune preuve que les degrés de croyance des gens ne soient pas si rigoureux. L'expérience 3 a mesuré la cohérence supérieure au niveau aléatoire en utilisant les intervalles

à point exact, et l'a comparé avec la cohérence supérieure au niveau aléatoire pour laquelle les limites d'intervalles avaient été élargies de 5% et 10%, accroissant donc le taux de hasard de la cohérence du même montant. Cela a rendu l'échelle de mesure plus grossière sans la rendre nécessairement plus accommodante. La cohérence supérieure au niveau aléatoire augmentait quand s'élargissait l'échelle de 5%, i.e. quand le nombre de points sur l'échelle se réduisait de 101 à 10, principalement pour l'équivalence de *de morgan* et la contradiction de *non de morgan*, pour lesquels l'intervalle de cohérence de la conclusion est une valeur de point. Cela a eu peu d'effet sur les autres inférences dont les intervalles de cohérence étaient déjà plus grands dès le début. Augmenter la grossièreté de 10% n'a pas eu d'effet supplémentaire. Dans l'Expérience 5, la question de la précision des degrés de croyance des gens a été évaluée de manière différente, en comparant la cohérence de réponse pour les probabilités de conclusion qui étaient clairement à l'intérieur ou à l'extérieur de l'intervalle, avec les probabilités de conclusion qui étaient à la limite de l'intervalle. La cohérence supérieure au niveau aléatoire était plus grande pour les probabilités de conclusion clairement d'un côté de l'intervalle, et cet effet n'était pas limité à *de morgan* et *non de morgan* mais présent généralement à travers les inférences.

Il paraît sensé que les degrés de croyance soient généralement plus grossiers que les probabilités de point, étant donné la nature incertaine d'une grande part de l'information que l'on reçoit dans les situations quotidiennes, et les limites de notre mémoire travail quant aux événements passés (c.f. Sanborn & Chater, 2016). La thèse ci-présente a proposé deux méthodes pour quantifier cette précision, ou imprécision, dans les croyances des gens. Cette précision va sûrement varier dans différentes situations et domaines d'expertise. Mais la capacité à la mesurer pour un contexte donné, utilisant les outils de la théorie de la probabilité, peut être utile pour interpréter les observations expérimentales, et paraît bloquer d'un des arguments amené par les partisans des systèmes de niveau de calcul qui sont eux-mêmes plus grossiers que la théorie de la probabilité, comme la théorie du classement, ou l'usage d'expressions verbales, de probabilité qualitative (Khemlani, Lotstein, & Johnson-Laird, 2014; Politzer & Baratgin, 2016; Spohn, 2013). De telles échelles de mesure alternatives ont un degré de grossièreté intégré, fixe, décidé a priori, son usage rendant impossible la mesure empirique de la grossièreté réelle des degrés de croyance.

La variance des distributions de croyances

En plus d'évaluer la sensibilité des gens à la position des intervalles de cohérence, les Expériences 3, 4, 8 et 9 ont examiné les intuitions des gens quant à la largeur des intervalles. Les Expériences 3 et 4 ont inclus une évaluation quant à la présence ou non d'une variance des réponses plus grande quand l'intervalle de cohérence était grand que quand il était réduit, en utilisant l'information de probabilité de prémisse pour estimer la largeur d'intervalle. L'hypothèse était que la variance de réponse serait plus grande quand l'intervalle serait plus

grand, mais aucune relation n'a été trouvée entre les deux. L'Expérience 8 évaluait si la confiance des gens en l'exactitude de leurs jugements de probabilité de conclusion (Thompson & Johnson, 2014) variait en fonction de la largeur d'intervalle. Si la confiance était plus faible pour les intervalles plus grands, cela pourrait suggérer que les gens cherchent une seule réponse optimale parmi une distribution, e.g. correspondant au moyen de distribution, qui est plus difficile à trouver quand il y a beaucoup d'options. Si la confiance était plus grande pour les intervalles plus grands, cela pourrait suggérer que les gens se concentrent sur la tâche de rendre leurs réponses cohérentes, ce qui est plus aisé quand le nombre d'options de réponses cohérentes est plus élevé. Mais encore une fois aucune relation n'a été trouvée entre les deux.

L'Expérience 9 a aidé à interpréter les résultats de l'Expérience 8, en suggérant que l'absence de relation entre la confiance de réponse et la largeur d'intervalle n'était pas due à un manque de sensibilité pour les paramètres déterminant la variance de distribution. A la place, il semble que les gens, dans un premier temps, suivent la contrainte déductive de la cohérence, essayant de donner des réponses qui se trouvent dans l'intervalle ; mais que si l'intervalle est assez large, alors des considérations inductives peuvent ou non réduire davantage le choix de réponse. Cette interprétation avait aussi été suggérée par une inspection de la distribution de réponses pour chaque inférence. Quand l'intervalle était réduit, la distribution de réponses était elle-aussi réduite et paraissait suivre de près l'emplacement de l'intervalle. Quand l'intervalle était large, la distribution de réponses était uniforme dans certains cas, suggérant que les gens essayaient surtout d'être cohérents, ne réduisant pas davantage leurs réponses de façon particulière. Mais dans d'autres cas la distribution de réponses était fortement biaisée vers la limite d'un intervalle, ou même multimodale, suggérant que des critères inductifs supplémentaires jouaient un rôle important dans la réduction des réponses supplémentaires de différentes manières. Les distributions de réponses calculées dans l'Expérience 10 ont généré des impressions similaires. De manière générale, ces observations ont éclairé les rôles complémentaires de la déduction et l'induction dans le raisonnement à partir de prémisses incertaines.

La validité-p importe au-delà de la cohérence

Il peut être difficile d'évaluer le rôle de la validité-p au-delà du rôle de la cohérence dans le raisonnement, parce que les contraintes normatives pertinentes se basent sur la cohérence dans les deux cas. Dans cette thèse il a été proposé de décrire la validité-p, i.e. le maintien de probabilité, comme une caractéristique des intervalles de cohérence. La validité-p peut être utilisée pour catégoriser les inférences en deux groupes (déductives et inductives) selon si leurs intervalles de cohérence maintiennent ou non la probabilité des prémisses à la conclusion. A travers cette caractérisation, la question n'est pas de savoir si les gens respectent ou non les contraintes normatives de la validité-p dans leurs jugements de probabilité de conclusion, parce que ces contraintes normatives sont établies par la cohérence. La question

est plutôt d'établir dans quelle mesure la distinction marquée par la validité-p entre les deux groupes d'inférences importe aux gens.

A travers les expériences, il n'y a pas eu de preuve que les gens distinguent les inférences valides-p (déductives) des inférences invalides-p (inductives) en termes de l'effort qu'ils investissent pour les élaborer, car la cohérence supérieure au niveau aléatoire ne différait pas systématiquement entre inférences valides-p et invalides-p. Cependant l'Expérience 10 a montré que les gens distinguaient entre inférences déductives et inductives dans leurs jugements de qualité d'inférence. Les inférences déductives qui maintenaient la probabilité étaient jugées plus correctes que les inférences inductives qui ne le faisaient pas. De plus, la validité-p était considérée comme spéciale parmi les différents niveaux de maintien de probabilité étudiés, avec des formes de maintien de probabilité qui étaient plus strictes que la validité-p n'ayant qu'un impact négligeable sur les jugements de qualité. Cela corroborait empiriquement le traitement spécial donné depuis longtemps à la distinction entre déduction et induction dans la littérature philosophique.

L'Expérience 10 a aussi amené une distinction, pour les inférences inductives, entre les cas suivants. Les inférences dont l'intervalle de cohérence est l'intervalle d'unité non-informatif (comme les paradoxes du conditionnel matériel); les inférences avec un intervalle de cohérence qui ne maintient pas fortement la probabilité mais est contraint de manière différente par les prémisses (comme l'AC); et les inférences avec une conclusion qui est la négation de la conclusion d'une inférence valide, de sorte que la conclusion est impossible quand les prémisses sont certaines, et la conclusion est très improbable quand les prémisses sont très probables. Il serait intéressant d'approfondir dans quelle mesure ces distinctions plus rigoureuses jouent un rôle dans les évaluations de qualité d'inférence des gens.

Il vaudrait aussi la peine de développer d'autres manières d'évaluer dans quelle mesure, et dans quels contextes, les gens traitent différemment les inférences déductives et inductives (c.f. Trippas, Handley, Verde, & Morsanyi, 2016). De manière générale on peut s'attendre à ce que la différence importe dans certains contextes, mais pas dans d'autres. Le maintien de probabilité ajoute de la fiabilité à la probabilité de conclusion d'une inférence à travers les cas individuels. Cette fiabilité peut être importante dans les situations où, comme dans les matériaux expérimentaux, l'enjeu est grand et un examen approfondi est nécessaire pour ne pas tirer trop vite des conclusions. Mais dans d'autres contextes il serait peut-être plus utile de répondre rapidement, sans hésiter de tirer des conclusions, e.g. parce que seule une réponse approximative est nécessaire ou possible au vu de l'information disponible, et le raisonneur doit avancer pour aborder la tâche suivante. Si l'on s'appuyait uniquement sur la déduction pour raisonner quotidiennement, même si cela est probabiliste, l'on pourrait régulièrement être bloqué en l'absence de critères suffisants pour tirer des conclusions. De plus, comme traité en relation avec les Expériences 8 et 9, la déduction et l'induction paraissent souvent travailler de concert. Ainsi, au lieu d'interroger dans quels contextes la déduction est pertinente, il serait

peut-être plus utile d'interroger comment les différentes contributions de la déduction et de l'induction peuvent être mesurées dans des contextes de raisonnement où elles jouent toutes deux un rôle.

Conclusions

Les notions de déduction binaires de la logique classique logique, *consistance* et validité, peuvent être généralisées pour englober les degrés de croyance. La *consistance* peut être généralisée comme cohérence, et la validité comme validité-p. Mais le fait que cette généralisation soit possible dans la logique formelle n'implique pas que les gens utilisent la déduction de manière probabiliste. La recherche présentée dans cette thèse étudiait le rôle de la déduction dans le raisonnement à partir de prémisses incertaines à travers dix expériences. Elle a trouvé des preuves que la cohérence et la validité-p ne sont pas juste des formalismes abstraits, mais que les gens suivent les contraintes normatives qu'ils fixent dans leur raisonnement. Cela fournit des preuves quant à l'adéquation descriptive de la cohérence et de la validité-p comme principes de niveau de calcul définissant les tâches à accomplir au sein d'un raisonnement. Cela a des implications quant à l'interprétation d'observations antérieures dans la littérature sur les rôles de la déduction et des degrés de croyance, et offre une perspective pour générer de nouvelles hypothèses de recherche sur l'interface entre raisonnement déductif et inductif.

SUMMARY

Theoretical background

The probabilistic approach to reasoning

The probabilistic approach in the psychology of reasoning (Evans, 2006; Evans & Stanovich, 2013; Oaksford & Chater, 2007, 2013; Over, 2016; Pfeifer & Kleiter, 2009, 2010; Politzer & Baratgin, 2016) is based on the fundamental hypothesis that most reasoning, in both everyday life and in science, takes place in contexts of uncertainty. The uncertainty, or degrees of belief, in the statements from which we reason cannot be modelled in classical logic, which is binary in that it represents only what is definitely true or false. But this uncertainty can be modelled in probability theory. There is ample empirical evidence that people tend to reason from uncertain premises even when instructed to assume the premises to be certain (Byrne, 1989; George, 1997; Liu, Lo, & Wu, 1996; Stevenson & Over, 1995; Thompson, 1994).

In classical logic, a conditional, *if p then q*, is interpreted as the *material conditional*, which is equivalent to *not-p or q*. But in the probabilistic approach this view is rejected, and it is instead proposed that conditionals are interpreted probabilistically. The probability of a conditional, $P(\text{if } p \text{ then } q)$, is proposed to correspond, not to $P(\text{not-}p \text{ or } q)$, but instead to the conditional probability of q given p , $P(q|p)$. This relationship, $P(\text{if } p \text{ then } q) = P(q|p)$, is called the Equation (Edgington, 1995). To have a degree of belief in a conditional is considered different from having a degree of belief about a matter of fact in the world. The proposal is that people arrive at $P(q|p)$, not by first having some specific degrees of belief in $P(p \ \& \ q)$ and in $P(p)$, and then taking their ratio, $P(q|p) = P(p \ \& \ q)/P(p)$, but instead through a *Ramsey test* (Ramsey, 1929/1990; Stalnaker, 1968). This is a mental simulation in which people hypothetically suppose the antecedent p of the conditional, make any necessary changes to their beliefs to preserve consistency, and assess the probability of q under this supposition.

The material conditional interpretation of conditionals leads to counterintuitive consequences that are avoided by the Equation. For example, consider the conditional "If we toss this fair coin, it will land heads." Intuitively, the probability of this conditional is .5, which corresponds to the conditional probability of the fair coin landing heads, given that it is tossed. But suppose the coin is in a glass box in a museum, and we are almost certain that we are not going to toss it. The material conditional is true whenever the antecedent of the conditional is false. As a result, the more confident we become that we will not toss the coin, the more probable it will be that if we do toss it, it will land heads, according to the material conditional interpretation. But this seems absurd.

A conditional based on the Equation has been called the *probability conditional* (Adams, 1998), *suppositional conditional* (Edgington, 2014), and *conditional event* (de Finetti,

1937/1980). It is referred to as the probability conditional in this thesis. The hypothesis that people's modal interpretation of conditionals corresponds to the probability conditional has received wide empirical support (Barrouillet & Gauffroy, 2015; Evans, Handley, & Over, 2003; Fugard, Pfeifer, Mayerhofer, & Kleiter, 2011; Oberauer & Wilhelm, 2003; Politzer, Over, & Baratgin, 2010).

The role of deduction in reasoning from uncertain premises

With the advent of the probabilistic paradigm, the question of what role, if any, deduction plays in reasoning, arose anew (Evans, 2012; Evans & Over, 2013). The present thesis argues that the answer to this question depends in part on how deduction is defined. The definition of deduction from classical logic is binary, and produces what de Finetti called "the logic of certainty" (de Finetti, 1972; Elqayam & Over, 2013). When the premises are assumed to be certain, the conclusion of a classically valid inference must be certain as well. But suppose people have uncertain degrees of belief in the premises of an inference. What degree of belief would be reasonable for them to have in the conclusion? Classical logic cannot answer this question, for it cannot be applied to reasoning under uncertainty. If we retain its binary definition of deduction, the central hypothesis of the probabilistic approach, that most reasoning takes place from uncertain premises, implies that deduction has only a small role to play in most reasoning situations (Oaksford & Chater, 2007).

Using the classical definition of deduction, a number of empirical studies have found an effect of deduction in reasoning tasks, in addition to what has been called an effect of "belief" (Evans, Handley, Neilens, & Over, 2010; Klauer, Beller, & Hütter, 2010; Markovits, Brisson, & Chantal, 2015; Rips, 2001; Singmann & Klauer, 2011; Thompson, 1994; Trippas, Thompson, & Handley, 2017). In particular, people more often accept the conclusion of valid than of invalid inferences, in addition to accepting believable conclusions more often than unbelievable ones. The latter effect has been called *belief bias*, as it constitutes a departure from the normative response under binary logical instructions.

The findings of effects of both "logic" and "belief" in reasoning have contributed to the development of dual-component theories (Evans, 2006; Klauer et al., 2010; Markovits et al., 2015; Verschueren, Schaeken, & d'Ydewalle, 2005). When these theories assume a binary logical process that is contrasted with a "belief based" process, the "logical" process questions the basic hypothesis in the probabilistic approach that most real world reasoning is from uncertain premises.

The present thesis argues for an alternative way of accounting for the findings on an effect of deduction that does not question the basis of the probabilistic approach. Rather than integrating a binary logical component on the one side, and a probabilistic, belief based component on the other, into a dual-component framework, it explores the integration of logic and probability itself.

The central deductive concepts of classical logic, consistency and validity, can be generalised to cover degrees of belief. Binary consistency can be generalised to *coherence*, and binary validity to probabilistic validity, or *p-validity* for short (Adams, 1998; Coletti & Scozzafava, 2002; Gilio, 2002). Two statements are consistent if and only if they can both be true at the same time. Similarly, probability judgments are coherent if and only if they respect the axioms of probability theory. For example, if we believe it is 80% likely to rain today, then for our beliefs to be coherent, we would also have to be willing to believe it 20% likely not to rain today, otherwise the probabilities would not sum to 1. An inference is valid if and only if it would be inconsistent for its premises to be true but its conclusion false. This means that a valid inference preserves truth from premises to conclusion. To see how this definition can be generalised to degrees of belief, consider the simple case of a one-premise inference. A one-premise inference is p-valid if and only if it would be incoherent for the probability of its conclusion to be lower than the probability of its premise. Hence a p-valid inference is probability preserving just like a binary valid inference is truth preserving. The above generalisations make it possible to study deduction from uncertain premises (Stevenson & Over, 1995), so that there is no need to qualify the basis of the probabilistic approach.

With this generalised notion of deduction, many findings on belief bias can be reinterpreted as effects of coherence. Participants in an experiment may be violating the classical deductive *instructions* to assume the premises to be certain, and nonetheless be reasoning deductively, by estimating a coherent degree of belief in the conclusion of an inference on the basis of its logical form and of their uncertain degrees of belief in the premises. Not to follow binary instructions can be said to be a fault, but it is not necessarily a logical one.

Research question

The fact that the definition of deduction can be generalised to cover degrees of belief does not imply that people actually use deduction in a probabilistic way. It could still be that when people engage in deduction, they do so in the classical binary way, and that reasoning from uncertain premises is inductive in practice even when it does not have to be so in theory. Whether people are sensitive to the constraints of coherence and of p-validity is an empirical question that has only recently started to be investigated, but which was the main focus of this thesis.

The thesis investigated the role of coherence and p-validity in uncertain reasoning through ten experiments. Previous studies of the coherence of people's conclusion probability judgments had focussed on conditional syllogisms (Evans, Thompson, & Over, 2015; Singmann, Klauer, & Over, 2014). Responses to these two-premise inferences were found to be coherent above chance levels mainly for the inference of *modus ponens* (MP), and less reliably for *denial of the antecedent* (DA) inference. However, the chance rates in these

studies were often very high (e.g. Singmann et al., 2014, Figure 3) and varied between inferences, making it more difficult to detect above-chance coherence when it is there, and to compare this coherence between inferences.

The findings obtained across experiments

The present experiments extended these findings, using different methodologies, across 23 inference forms. These inferences differed in their structural complexity. They included valid (deductive) and invalid (inductive) inferences, and they featured negations, conjunctions, disjunctions, and conditionals. Most of the inferences containing conditionals assumed them to be interpreted as probability conditionals, based on the Equation $P(\text{if } p \text{ then } q) = P(q|p)$ (Adams, 1998; Edgington, 1995; Jeffrey, 1991), but some experiments included comparisons of different interpretations of conditionals.

Coherent responses to MT

For conditional syllogisms, the results corroborated the earlier finding of above-chance coherence for MP, but found coherence to be reliably above chance levels for the inference of modus tollens (MT) as well. Of the six experiments in which both inferences were investigated, overall coherence was above chance in all cases for MP. For MT, coherence was above chance levels in all but two instances. The first was in Experiment 3, where it was at chance levels in the statements task but above chance levels in the inferences task. In Experiment 4, which replicated lab Experiment 3 on the internet, coherence was above chance levels across conditions. The second instance was in Experiment 7. Coherence for MT was at chance levels when participants were given binary paradigm instructions to assume the premises to be true, and then judge whether the conclusion also has to be true. This was in spite of the fact that coherence for MT was above chance levels for the same materials when given probabilistic instructions (Experiment 6).

These findings constitute strong evidence for a basic sensitivity to coherence constraints for MT. They also provide direct evidence for the descriptive adequacy of the proposal that the constraints of deduction are not limited to the special case in which premises are assumed to be certain.

The methodology of Experiments 5 to 7 made it possible, not only to assess whether or not conclusion probability judgments are coherent above chance levels, but also to make quantitative comparisons of above-chance coherence between inferences. This was accomplished by a design that equated the chance rate of coherence across inferences and conditions. The experiments revealed that although responses to MT were generally coherent above chance levels, coherence was lower for MT than for MP, in line with the literature using

binary paradigm instructions. In the probabilistic approach, the difference in the acceptance of the two inferences can be explained as a result of dynamic reasoning (Oaksford & Chater, 2013). Specifically, MT can sometimes be viewed as an instance of a *reductio ad absurdum* inference. The categorical premise *not-q* negates an element of the conditional premise *if p then q*, with the result that a high degree of belief in both premises is incompatible with a high degree of belief in the element *p* of the conclusion. But the conflict between *if p then q*, *not-q*, and *p* can be resolved in different ways. Depending on background beliefs, one person could see *if p then q* and *not-q* as a reason to disbelieve *p*, in line with MT, while another person could see *not-q* and *p* as a reason to disbelieve *if p then q*, engaging in dynamic reasoning. Logic does not itself tell us which statement should give way, so that without instructions to assume that both premises are certain, the choice can be made on inductive grounds. This perspective makes it possible to account for the asymmetry between MP and MT in a rational way.

Changing responses to AC and DA

The coherence findings for the conditional syllogisms *affirmation of the consequent* (AC) and *denial of the antecedent* (DA) were more equivocal. In Experiment 3, coherence for both inferences was above chance levels in the statements task, but not in the inferences task. In Experiment 4 coherence was above chance levels for both inferences and both task conditions. However, coherence for both inferences was below chance in Experiment 5, and at chance in Experiments 6 and 7. In Experiment 8, coherence was above chance for DA (this experiment did not include AC). Leaving aside the statements tasks of Experiments 3 and 4, this means that coherence for AC was above chance in one out of five experiments, and coherence for DA was above chance in two out of six experiments. A lack of above-chance coherence for these two inferences is difficult to interpret, but it could be explained by a biconditional interpretation of the conditional premise. More specifically, the materials used may have suggested to participants that the antecedent and consequent of the conditional were positively correlated. In line with this, an analysis of the data in Experiment 3 showed that coherence was above chance levels under the assumption of a biconditional interpretation for the same responses that coherence was at chance levels under the assumption of a conditional interpretation. To be able to interpret the coherence results for these inferences, the two interpretations would therefore have to be disentangled, by controlling explicitly for the correlation between antecedent and consequent. Although such an experiment was beyond the scope of this thesis, the fact that it is needed to interpret the findings on AC and DA is nonetheless a more precise, and less pessimistic, standpoint than the suggestion from previous studies that responses are generally incoherent for these inferences (e.g. Singmann et al., 2014).

Further studies of AC and DA, but also of MP, MT, and further two-premise inferences, could assess to what extent incoherent responses are responses that overestimate or underestimate the probability of the conclusion, given the probabilities assigned to the premises, and to what extent this depends on the risks and benefits suggested by the materials (Oberauer & Wilhelm, 2003). The hypothesis that there can be overconfidence in the conclusion of a valid inference could not even be formulated in the binary paradigm. But this hypothesis can extend the study of suppression effects, making it necessary to distinguish between the suppression of the conclusion of an inference, and the suppression of the inference itself (Over & Cruz, 2018).

Conditionals, or-introduction, and the conjunction fallacy

Experiments 1 and 2 not only extended the coherence results to further inferences, but also provided information about the meanings of the component statements, and of factors involved in reasoning with them. An analysis of the *or-to-if* inference showed that people's responses were coherent above chance levels under the assumption that they interpret the conditional in the conclusion as the probability conditional, whereas they were incoherent below chance levels under the assumption that they interpret the conditional as the material conditional, which is equivalent to *not-p or q*. This constitutes a novel form of evidence for the interpretation of conditionals in terms of the Equation.

Coherence was reliably above chance levels across four variants of the inference of *or-introduction*, suggesting that although it may be pragmatically infelicitous, under binary paradigm instructions, to state *p or q* when one could be more informative and precise by stating *p*, people readily treat the inference as valid when asked directly about their degrees of belief in premise and conclusion. This is in accordance with the view that the lower endorsement rates found for the inference under binary paradigm instructions are due to pragmatic effects (Cruz, Over, & Oaksford, 2017; Orenes & Johnson-Laird, 2012), contrary to the recent proposal in a revision of mental model theory that the inference is in fact invalid (Johnson-Laird, Khemlani, & Goodwin, 2015).

Further, Experiment 2 showed that although responses were reliably coherent for *and-elimination* when using neutral materials (see also Politzer & Baratgin, 2016, for converging evidence), coherence for the same inference broke down when using the materials known to cause the conjunction fallacy (Tversky & Kahneman, 1983). This was in spite of the apparently transparent task of explicitly inferring the probability of *p* from the probability of *p & q*. The finding underlines the strength, and at the same time the limited scope, of the fallacy.

Comparing above-chance coherence between inferences

It was pointed out in the thesis that it is difficult to make quantitative comparisons of above-chance coherence between inferences when the chance rates of the inferences are not equal.

The reason is that the chance rate is subtracted from the observed coherence rate to obtain the measure of above-chance coherence. The larger the chance rate, the lower the probability of detecting above-chance coherence when it is there; that is, the lower the sensitivity of the test for above-chance coherence. The chance rate corresponds to the width of the coherence interval, which in turn depends on the form of the inference, on whether the inference is valid or invalid, and on the premise probabilities. It is possible to study whether there is above-chance coherence for various inference forms without holding the chance rate constant (Cruz, Baratgin, Oaksford, & Over, 2015; Evans et al., 2015; Singmann et al., 2014), but further comparisons seem difficult to interpret.

In Experiments 5 to 7, above-chance coherence was made more comparable, across 12 inferences, using two methods. In the first, participants were given a set of premise probabilities, each of which was presented with a number of conclusion probabilities. The task was to give a binary response as to whether a given conclusion probability was possible or not, given the premise probabilities. In the second method, participants were again given a set of premise probabilities, and were asked whether the probability of the conclusion could be higher, and whether it could be lower, than the probability of the premise (in the case of one-premise inferences) or than 50% (in the case of two-premise inferences). These binary response formats rendered the chance rate of a coherent response 50% across all inferences.

Using these methods, it was found that the inference for which above-chance coherence was highest was the contradiction *not de morgan*, followed closely by MP. Except for AC and DA discussed above, coherence for the remaining inferences was lower but still generally above chance. Two oddballs were the inferences from *or-to-if* and from *if-to-or*. Coherence was at chance levels for *or-to-if* in Experiment 5, for *if-to-or* in Experiment 6, and for both in Experiment 7 (which used binary paradigm instructions). Leaving aside the statements task of Experiments 3 and 4, and considering only the inferences task in those experiments, this means that coherence for *or-to-if* and *if-to-or* was above chance in four out of six experiments. These inferences contain negations at the start of the premise or of the conclusion, and this could have made them more difficult to process. Responses to both inferences were above chance when measured across positions of the negation in Experiment 1, but further studies would be necessary to establish why coherence was somewhat less reliable for these inferences.

A further step would also be to investigate in more detail what makes the inferences of not de morgan and MP stand out in terms of the high rates of coherent responses to them. For example, one could test whether contradictions in general are detected more easily than other logical relations, or whether it is something specific to the negation of de morgan that is at play.

Generally, the finding that, across the inferences investigated, coherence was above chance levels in the great majority of cases, and failures of coherence were the exception

rather than the rule, provides strong additional support for the hypothesis that people are sensitive, at some level, to the constraints of coherence over and above the binary constraints of consistency.

The effect of an explicit inference task and working memory

Experiments 3 and 4 extended the results of Evans et al. (2015) on the role of an explicit inference task for coherence. Somewhat surprisingly, coherence was already above chance levels in the large majority of cases in the statements task: when people were given the statements that made up the inferences in random order one at a time on the screen. An explicit inference task, in which the statements were arranged into inferences, and each inference presented one at a time on the screen, tended to increase coherence in the few cases in which it was not already above chance. However, there was an exception in Experiment 3 for AC and DA, where coherence was above chance in the statements task but at chance levels in the inferences task. It may be that in some cases, pragmatic factors related to the assertability of a statement as a conclusion drawn from other statements, or to the relation between statements that are difficult to integrate due to the presence of negations, can lead to reductions in coherence that would not arise when the statements are considered in isolation. Another possibility is that putting the conditionals in an explicit AC or DA inference task tends to increase the biconditional interpretation, since otherwise the inferences are invalid. But the finding could also simply reflect the fact that different groups were in the statements and inferences task, and some may have interpreted the materials for AC and DA as implying a correlation between antecedent and consequent, and others not. Further replications would be needed to establish the reliability of this finding.

Experiments 3 and 4 also included an inferences task with working memory load. Responses in this condition generally differed little from those in the inferences task without working memory load. But where they differed, the load condition was associated with lower coherence rates, suggesting that the difference between the statements and the inferences task may be due in part to the differing demands they pose on working memory for calibrating beliefs across statements. However, the weak effect of the load condition is in line with the finding that coherence was in most cases already above chance in the statements task, so that the inferences task had only little to add that could be disrupted.

The overall pattern suggests that people may have an implicit, spontaneous tendency to establish coherence between beliefs, but that in situations in which this fails, an explicit inference task, in which all relevant pieces of information are available simultaneously on the screen, and people can focus their attention directly on the relations between them, tends to be helpful. In any case, explicit inference is necessary when establishing relations between novel materials for which no beliefs are yet available.

Certain vs. uncertain premises, probabilistic vs. binary paradigm instructions

Experiments 6 and 7 made it possible to compare above-chance coherence for inferences with certain vs. uncertain premises, and for inferences with probabilistic vs. binary paradigm instructions, using the same inferences, materials, and response format. There was no evidence that coherence is lower when the premises are uncertain, nor that coherence is lower when probabilistic rather than binary paradigm instructions are used. This provides a novel, strong form of evidence that deduction from uncertain premises is possible, and is not restricted to reasoning from certainty. Certain truth and certain falsity did not appear qualitatively different from uncertain degrees of belief, but rather as endpoints on a common scale.

Factors with no systematic effect on above-chance coherence

In addition to the finding that coherence did not differ between certain and uncertain premises, nor between probabilistic and binary paradigm instructions, Experiments 3 and 4 found no evidence of a systematic difference in people's responses to the reasoning tasks studied in an internet and in a lab setting, making it easier to generalise results between them, as well as between the experiments conducted in this thesis and the earlier lab results from Evans et al. (2015). In addition, Experiment 5 found no evidence for a difference in response coherence as a function of whether people were asked to judge whether a conclusion fell inside or outside the coherence interval. Across experiments, there also seemed to be no systematic difference in response coherence between one- and two-premise inferences. The differences in coherence between inferences rather appeared to be based on more specific factors, such as whether they contained negations or could be interpreted in alternative ways. Finally, across experiments there was no evidence that coherence differed between valid and invalid, i.e. between deductive and inductive, inferences. This result makes sense given that the constraints of coherence hold for both inference types, and deductive inferences merely have stronger constraints on the lower limits of their interval boundaries. The above negative results can help interpret and add precision to the positive findings observed in these experiments.

The precision of people's degrees of belief

Coherence intervals are usually measured using point probabilities, but there was evidence that people's degrees of belief are not that fine grained. Experiment 3 measured above-chance coherence using the exact point intervals, and compared this with above-chance coherence in which the interval boundaries were widened by 5% and by 10%, thereby widening the chance rate of coherence by a corresponding amount. This made the measurement scale coarser without making it necessarily more lenient. Above-chance coherence increased when widening the scale by 5%, i. e. when the number of points on the scale was reduced from 101 to 10, mainly for the equivalence of *de morgan* and the contradiction of *not de morgan*, for which the conclusion coherence interval is a point value. It had only little effect on the other

inferences whose coherence intervals were already wider from the beginning. Increasing the coarseness by 10% had no incremental effect. In Experiment 5, the question of the precision of people's degrees of belief was assessed in a different way, comparing response coherence for conclusion probabilities that were clearly inside or outside the interval, with conclusion probabilities that were at the interval edge. Above-chance coherence was higher for conclusion probabilities clearly on one side of the interval, and this effect was not restricted to de Morgan and not de Morgan but held more generally across inferences.

It seems to make sense for degrees of belief to be generally coarser than point probabilities, given the uncertain nature of much of the information we receive in everyday situations, and the limits of our working memory for past instances of an event (c.f. Sanborn & Chater, 2016). The present thesis proposed two methods of quantifying this precision, or fuzziness, in people's beliefs. This precision will likely vary across content domains and domain expertise. But the ability to measure it for a given context, using the tools of probability theory, can be useful for interpreting experimental findings, and seems to disable one of the arguments brought forward by advocates of computational level systems that are themselves coarser than probability theory, like ranking theory, or the use of verbal, qualitative probability expressions (Khemlani, Lotstein, & Johnson-Laird, 2015; Politzer & Baratgin, 2016; Spohn, 2013). Such alternative measurement scales have a built-in, fixed degree of coarseness that is decided a priori, the use of which makes it impossible to measure the actual coarseness of degrees of belief empirically.

The variance of belief distributions

In addition to assessing people's sensitivity to the location of coherence intervals, Experiments 3, 4, 8, and 9 examined people's intuitions about interval width. Experiments 3 and 4 included an assessment of whether the variance of responses was larger when the coherence interval was wide than when it was narrow, using premise probability information to estimate interval width. The hypothesis was that response variance would be higher when the interval was wider, but no relation was found between the two. Experiment 8 assessed whether people's confidence in the correctness of their conclusion probability judgments (Thompson & Johnson, 2014) varied as a function of interval width. If confidence was lower for wider intervals, this might suggest that people are looking for a single optimal response within a distribution, e.g. corresponding to the distribution mean, which is more difficult to find when there are many options. If confidence was higher for wider intervals, this might suggest that people are focussing on the task of rendering their responses coherent, which is easier when the number of coherent response options is larger. But again no relation was found between the two.

Experiment 9 helped interpret the results of Experiment 8, by suggesting that the absence of a relation between response confidence and interval width was not due to a lack of

sensitivity for parameters determining distribution variance. Instead, it seems as if people, in the first instance, follow the deductive constraint of coherence, trying to give responses that fall within the interval; but that if the interval is wide enough, then inductive considerations may or may not narrow down the choice of response further. This interpretation was also suggested by an inspection of the distribution of responses for each inference. When the interval was narrow, the distribution of responses was also narrow and seemed to follow the location of the interval closely. When the interval was wide, the distribution of responses was flat in some cases, suggesting that people were mainly trying to be coherent, without narrowing down their responses further in any specific way. But in other cases the distribution of responses was strongly skewed towards one interval edge, or even multimodal, suggesting that additional inductive criteria were playing a strong role in narrowing down people's responses further in various ways. The response distributions computed in Experiment 10 led to similar impressions. Generally, these findings shed further light on the complementary roles of deduction and induction in reasoning from uncertain premises.

P-validity matters over and above coherence

It can be difficult to assess the role of p-validity over and above the role of coherence in reasoning, because the relevant normative constraints are based on coherence in both cases. In this thesis it was proposed to describe p-validity, i.e. probability preservation, as a feature of coherence intervals. P-validity can be used to categorise inferences into two groups (deductive and inductive) according on whether or not their coherence intervals preserve probability from premises to conclusion. With this characterisation, the question is not whether people respect the normative constraints of p-validity in their conclusion probability judgments, because these normative constraints are set by coherence. The question is rather to what extent the distinction marked by p-validity between the two groups of inferences matters to people.

Across experiments, there was no evidence that people distinguish between p-valid (deductive) and p-invalid (inductive) inferences in terms of the effort they invest in drawing them, because above-chance coherence did not differ systematically between p-valid and p-invalid inferences. But Experiment 10 showed that people did distinguish between deductive and inductive inferences in their judgments of inference quality. Deductive inferences that preserved probability were judged more correct than inductive inferences that did not. Further, p-validity was treated as special among the different levels of probability preservation studied, with forms of probability preservation that were stricter than p-validity having only a negligible further impact on quality judgments. This corroborated empirically the special treatment long given to the distinction between deduction and induction in the philosophical literature.

Experiment 10 also drew a distinction, for the inductive inferences, between the following cases. Inferences whose coherence interval is the uninformative unit interval (like

the paradoxes of the material conditional); inferences with a coherence interval that is not high probability preserving but is constrained in a different way by the premises (such as AC); and inferences with a conclusion that is the negation of the conclusion of a valid inference, so that the conclusion is impossible when the premises are certain, and the conclusion is very improbable when the premises are very probable. It would be interesting to investigate further to what extent these more fine-grained distinctions play a role in people's evaluations of inference quality.

It would also be worth developing further ways of assessing to what extent, and in which contexts, people treat deductive and inductive inferences differently (c.f. Trippas, Handley, Verde, & Morsanyi, 2016). In general one can expect the difference to matter in some contexts, but not in others. Probability preservation adds reliability to the conclusion probability of an inference across individual instances. This reliability may be important in situations when, as in some of the experimental materials, much is at stake and careful consideration is called for to avoid jumping to conclusions. But in other contexts it may be more helpful to respond quickly, without hesitating to jump to conclusions, e.g. because only an approximate answer is needed or possible given the available information, and the reasoner must move on to address the next task. If we relied only on deduction in everyday reasoning, even if it is probabilistic, we might regularly freeze in the absence of sufficient criteria for drawing any conclusion. Moreover, as discussed in relation to Experiments 8 and 9, deduction and induction often seem to work hand in hand. Thus, instead of asking in which contexts deduction is relevant, it may be more useful to ask how the different contributions of deduction and induction can be measured in reasoning contexts in which they both play a role.

Conclusions

The binary deductive notions of classical logical logic, consistency and validity, can be generalised to cover degrees of belief. Consistency can be generalised to coherence, and validity to p-validity. But the fact that this generalisation is possible in formal logic does not imply that people will actually use deduction in a probabilistic way. The research presented in this thesis investigated the role of deduction in reasoning from uncertain premises through ten experiments. It found evidence that coherence and p-validity are not just abstract formalisms, but that people follow the normative constraints set by them in their reasoning. This is evidence for the descriptive adequacy of coherence and p-validity as computational level principles defining the tasks people set out to accomplish when reasoning. It has implications for the interpretation of past findings in the literature on the roles of deduction and degrees of belief, and it offers a perspective for generating new research hypotheses in the interface between deductive and inductive reasoning.

References

- Adams, E. (1998). *A primer of probability logic*. Stanford, US: CLSI publications.
- Barrouillet, P., & Gauffroy, C. (2015). Probability in reasoning: A developmental test on conditionals. *Cognition*, *137*, 22-39.
- Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61-83.
- Coletti, G., & Scozzafava, R. (2002). *Probabilistic Logic in a Coherent Setting*. Dordrecht, NL: Kluwer.
- Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6*, 192.
- Cruz, N., Over, D., & Oaksford, M. (2017). The elusive oddness of or-introduction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *The 39th Annual Meeting of the Cognitive Science Society*. London, UK: Cognitive Science Society.
- De Finetti, B. (1937/1980). Foresight: its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokier (Eds.), *Studies in subjective probability* (55-118). New York, US: Wiley.
- De Finetti, B. (1972). *Probability, induction, and statistics*. London, UK: Wiley.
- Edgington, D. (1995). On conditionals. *Mind*, *104*, 235-329.
- Edgington, D. (2014, Winter). Indicative conditionals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. URL = [<https://plato.stanford.edu/archives/win2014/entries/conditionals/>](https://plato.stanford.edu/archives/win2014/entries/conditionals/).
- Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by S. Elqayam, J.F. Bonnefon, & D. E. Over. *Thinking & Reasoning*, *19*, 249-265.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378-395.
- Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*(1), 5-31.
- Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology*, *63*(5), 892-909.
- Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 321-335.
- Evans, J. St. B. T., & Over, D. E. (2013). Reasoning to and from belief: Deduction and induction are still distinct. *Thinking & Reasoning*, *19*, 268-283.
- Evans, J. St. B. T., Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.
- Evans, J. St. B. T., Thompson, V., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology*, *6*, 398.

- Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., & Kleiter, G. D. (2011). How people interpret conditionals: Shifts toward the conditional event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 635-648.
- George, C. (1997). Reasoning from uncertain premises. *Thinking and Reasoning*, *3*, 161-190.
- Gilio, A. (2002). Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence*, *34*, 5-34.
- Jeffrey, R. C. (1991). Matter of fact conditionals. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *65*, 161-183.
- Johnson-Laird, P. N., Khemlani, S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201-214.
- Khemlani, S. S., Lotstein, M., & Johnson-Laird, P. N. (2014). Naive probability: Model-based estimates of unique events. *Cognitive Science*, *39*(6), 1216-1258.
- Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 298-323.
- Liu, I., Lo, K., & Wu, J. (1996). A probabilistic interpretation of "if-then". *Quarterly Journal of Experimental Psychology*, *49A*, 828-844.
- Markovits, H., Brisson, J., & Chantal, P.-L. (2015). Additional evidence for a dual-strategy model of reasoning: Probabilistic reasoning is more invariant than reasoning about logical validity. *Memory & Cognition*, *43*, 1208-1215.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.
- Oaksford M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*, 346-379.
- Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 680-693.
- Orenes, I., & Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. *Mind & language*, *27*(4), 357-377.
- Over, D. E. (2016). The paradigm shift in the psychology of reasoning: The debate. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive unconscious and human rationality* (pp. 79-97). Cambridge US: MIT Press.
- Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In L. J. Ball, & V. A. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 434-450). Hove, UK: Psychology Press.
- Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206-217.

- Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford, & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 153-173). Oxford, UK: Oxford University Press.
- Politzer, G., & Baratgin, J. (2016). Deductive schemas with uncertain premises using qualitative probability expressions. *Thinking & Reasoning*, 22, 78-98.
- Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking & Reasoning*, 16, 172-197.
- Ramsey, F. P. (1929/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145-163). Cambridge, UK: Cambridge University Press.
- Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, 12(2), 129-134.
- Sanborn, A. N. & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883-893.
- Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, 17(3), 247-281.
- Singmann, H., & Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, 5, 316.
- Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, 37(6), 1074-1106.
- Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98-112). Oxford, UK: Blackwell.
- Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Journal of Experimental Psychology*, 48A(3), 613-643.
- Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition*, 22, 742-758.
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20(2), 215-244.
- Trippas, D., Handley, S. J., Verde, M., F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: learning, Memory, and Cognition*, 42(9), 1448-1457.
- Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, 45(4), 539-552.
- Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, 11(3), 239-278.

Acknowledgements

I am deeply grateful to a number of people for their encouragement and support in connection with the research presented in this thesis, only some part of which I undertake to report here. First I would like to thank my three supervisors, Mike Oaksford, Jean Baratgin, and David Over, from which I have learned so much, and who I have been extremely lucky to have had the chance to work with. I would like to thank Karl Christoph Klauer, Keith Stenning, Björn Meder, Rick Cooper, Carlos Gershenson García, and Sören Krach, who took the time to write references for an application process that ended with the award of scholarships for my Ph.D. studies, without which the present work would not have been possible in the given time frame. I am grateful to Klaus Oberauer and Phil Johnson-Laird, for very good research discussions and an ongoing (adversarial) collaboration. I would like to thank Ulrike Hahn and Stephan Hartmann for the opportunity to visit their lab at the Munich Center for Mathematical Philosophy, and further very good research discussions that broadened my perspective of the field. I would also like to thank Guy Politzer and Igor Douven in Paris for very good research discussions, and for introducing me to their theoretical positions. I am indebted to several researchers with whom I have grown to be friends over the years, even though we practically only meet at conferences. I am grateful to have had the chance to share the Merlin lab with my fellow Ph.D. students, for their company, support and interesting conversations, including but not limited to research. I would further like to thank my friends and family. Finally, but not least, I am grateful to DAAD and CONACYT for the scholarships that enabled me to pursue my studies.

CONTENTS

Overview

List of tables and figures	40
Part 1. Introduction	45
Chapter 1. Introduction	46
Part 2. Theoretical background	54
Chapter 2. Binary theories of reasoning and their accounts of conditionals	55
Chapter 3. Probabilistic theories of reasoning and probability conditionals.....	69
Chapter 4. Alternatives to the probabilistic approach in psychology	91
Part 3. Experiments	105
Chapter 5. Experiments 1 to 4: Coherence above chance levels	106
Chapter 6. Experiments 5 to 7: Quantitative comparisons of degrees of coherence	168
Chapter 7. Experiments 8 and 9: Response variance	223
Chapter 8. Experiment 10: Probability preservation properties	246
Part 4. General discussion	260
Chapter 9. General discussion.....	261
References	280
Appendices	299

Detailed view

List of tables and figures	40
Part 1. Introduction	45
Chapter 1. Introduction	46
1.1 Types of reasoning.....	47
1.2 Types of statements.....	48
1.3 Research questions.....	51
1.4 Outline of the thesis	52
Part 2. Theoretical background	54
Chapter 2. Binary theories of reasoning and their accounts of conditionals	55
2.1 Classical logic	56
2.2 The material conditional	56
2.3 The truth conditions of the material conditional plus conditions of assertability: Grice and Jackson.....	58

2.3.1 Grice.....	58
2.3.2 Jackson.....	60
2.4 Possible world semantics: Stalnaker and Lewis.....	62
2.4.1 Stalnaker	62
2.4.2 Lewis.....	65
2.5 The triviality results	66
Chapter 3. Probabilistic theories of reasoning and probability conditionals	69
3.1 Why represent degrees of belief with probabilities?.....	70
3.2 Which interpretation of probability?.....	71
3.2.1 The frequentist interpretation.....	71
3.2.2 The logical interpretation.....	72
3.2.3 The subjectivist interpretation.....	72
3.3 How can we measure degrees of belief, and why would we want them to be coherent?.....	74
3.3.1 Measuring beliefs by measuring actions.....	74
3.3.2 Dutch book arguments	76
3.4 Coherence and p-validity: Deduction from uncertain premises.....	77
3.5 Probability conditionals	80
3.6 Conditionals and validity	83
3.7 Uncertain reasoning beyond deduction: Dynamic reasoning.....	84
3.8 Empirical evidence for the probabilistic approach.....	86
3.8.1 Evidence for reasoning from uncertain premises.....	86
3.8.2 Evidence for the probability conditional.....	89
Chapter 4. Alternatives to the probabilistic approach in psychology	91
4.1 Mental model theory (MMT).....	92
4.1.1 Conditionals in MMT	93
4.1.2 Reasoning with conditional syllogisms in MMT	95
4.1.3 Mental models and probabilities.....	96
4.1.4 New MMT	97
4.2 Dual-component theories	98
4.2.1 "Logic" vs. "belief" in dual-component theories	99
4.2.2 Breaking the association of "logic" to type 2 and "belief" to type 1 processes	101
4.2.3 Breaking the "logic" vs. "belief" dichotomy itself.....	102
4.3 Research question	104
Part 3. Experiments	105
Chapter 5. Experiments 1 to 4: Coherence above chance levels.....	106

5.1 Methodological points relevant across experiments	107
5.1.1 Above-chance coherence	107
5.1.2 Linear mixed models.....	107
5.2 Experiment 1: Ifs and ors	111
5.2.1 Method	115
5.2.2 Results and discussion	117
5.2.3 General discussion	120
5.3 Experiment 2: Ifs, ands, and the conjunction fallacy	123
5.3.1 Overview of the conjunction fallacy	123
5.3.2 Ifs and ands	126
5.3.3 Method	127
5.3.4 Results and discussion	129
5.3.5 General discussion	130
5.4 Experiments 3 and 4: Intuition, reflection, and working memory	132
5.4.1 Experiment 3.....	133
5.4.2 Experiment 4.....	160
5.4.3 General discussion	167
Chapter 6. Experiments 5 to 7: Quantitative comparisons of degrees of belief	168
6.1 Experiment 5: At the edge vs. the centre of the coherence interval	169
6.1.1 Method	170
6.1.2 Results and discussion	176
6.1.3 General discussion	193
6.2 Experiment 6: Higher vs. lower than the premise probabilities.....	196
6.2.1 Method.....	196
6.2.2 Results and discussion	199
6.2.3 General discussion	209
6.3 Experiment 7: Certain premises and binary paradigm instructions	212
6.3.1 Method.....	213
6.3.2 Results and discussion	214
6.3.3 General discussion	220
Chapter 7. Experiments 8 and 9: Response variance	223
7.1 Experiment 8: Coherence interval width and response confidence	224
7.1.1 Varying location and width of coherence intervals	225
7.1.2 Measuring people's sensitivity to location and width	226
7.1.3 Method.....	229
7.1.4 Results and discussion	231
7.1.5 General discussion	239

7.2 Experiment 9: Sensitivity to the variance of distributions	240
7.2.1 Method	241
7.2.2 Results and discussion	243
7.2.3 General discussion	244
Chapter 8. Experiment 10: Probability preservation properties.....	246
8.1 Method	251
8.2 Results and discussion	253
8.3 General discussion	258
Part 4. General discussion	260
Chapter 9. General discussion	261
9.1 The findings obtained across experiments	263
9.1.1 Coherent responses to MT	263
9.1.2 Changing responses to AC and DA	266
9.1.3 Conditionals, or-introduction, and the conjunction fallacy.....	266
9.1.4 Comparing above-chance coherence between inferences.....	267
9.1.5 The effect of an explicit inference task and working memory.....	268
9.1.6 Certain vs. uncertain premises, probabilistic vs. binary paradigm instructions	269
9.1.7 Factors with no systematic effect on above-chance coherence.....	269
9.1.8 The precision of people's degrees of belief.....	270
9.1.9 The variance of belief distributions	271
9.1.10 P-validity matters over and above coherence	272
9.2 Conclusions.....	273
9.3 Implications for belief bias and dual-component theories	273
9.4 Limits of deduction and dynamic reasoning	275
9.5 Where next?	276
9.5.1 Dynamic reasoning	276
9.5.2 Counterfactuals, generals, and universals	277
9.5.3 Coherence and rationality	279
References.....	280
Appendices.....	299
Appendix A. Jeffrey tables for the probability biconditional.....	299
Appendix B. The materials used in Experiment 1.	300
Appendix C. The materials used in Experiments 3 and 4	311

Appendix D. The materials used in Experiment 5	330
Appendix E. The materials used in Experiments 6 and 7	332
Appendix F. The materials used in Experiment 9.....	335
Appendix G. The materials used in Experiment 10	337

List of tables and figures

Numbered on the basis of the chapter in which they occur.

List of tables

Caption	Page
Table 1.1. The truth tables for conjunction, disjunction, and the conditional in classical logic.	49
Table 2.1. The truth table for the material conditional.	57
Table 2.2. Examples of inferences that are invalid in the modal logical systems of Stalnaker (1968) and Lewis (1973), but valid in classical logic.	64
Table 2.3. Truth table for the example about a coin flip.	67
Table 3.1. Probability preservation properties of inferences, based on Adams (1996).	79
Table 3.2. The de Finetti table for the probability conditional.	81
Table 3.3. The Jeffrey table for the probability conditional.	82
Table 5.1. The inferences used in Experiment 1.	112
Table 5.2. The inferences used in Experiment 2.	127
Table 5.3. The four inferences of the experiment embedded in the Linda scenario.	128
Table 5.4. The inferences investigated in Experiments 3 and 4.	134
Table 5.5. The coherence intervals for the four conditional syllogisms.	135
Table 5.6. The Jeffrey table for the probability biconditional that results from adding the converse, <i>if q then p</i> , to the original conditional.	153
Table 5.7. The coherence intervals for biconditional AC (Bic AC) and biconditional DA (Bic DA).	154
Table 5.8. Variances of conclusion probability judgments in Experiment 3, separately for each group, inference, and premise probability condition.	157
Table 5.9. Variances of conclusion probability judgments in Experiment 4, separately for each group, inference, and premise probability condition.	165
Table 6.1. The inferences investigated in Experiments 5 to 7.	171
Table 6.2. The conclusion probabilities used in Experiment 5 for each inference and premise probability condition.	174
Table 7.1. The inferences used in Experiment 8.	225
Table 8.1. P-valid inferences with categorical conclusions and their p-invalid counterparts with conditional conclusions. Taken from Edgington (1995).	248
Table 8.2. The 10 inferences investigated, grouped by their probability preservation	249

properties.

Table 9.1. The inferences investigated in the 10 experiments of the thesis. 264

List of figures

Caption	Page
Figure 5.1. Observed and chance rate coherence for the eight inferences of Experiment 1, separately for each group. Error bars represent 95% CIs.	117
Figure 5.2. Above-chance coherence for the eight inferences of Experiment 1, separately for each group. Error bars show 95% CIs.	117
Figure 5.3. Above-chance coherence for the four inferences and the two task conditions of Experiment 2. Error bars show 95% CIs.	129
Figure 5.4. Distribution of the proportion of correct responses to the memory task in Group 3. Upper panel: for Experiment 3; lower panel: for Experiment 4.	137
Figure 5.5. Coherence intervals for the four conditional syllogisms: MP, MT, AC, and DA, as a function of their premise probabilities. The shaded areas in the graphs represent the coherence intervals.	138
Figure 5.6. Premise and conclusion probabilities in Experiment 3 for the inferences of type A, separately for each group and premise probability condition. Error bars show 95% CIs.	143
Figure 5.7. Premise and conclusion probabilities in Experiment 3 for the inferences of type B, separately for each group and premise probability condition. Error bars show 95% CIs.	144
Figure 5.8. Premise and conclusion probabilities in Experiment 3 for the inferences of type C, separately for each group and premise probability condition. Error bars show 95% CIs.	144
Figure 5.9. Coherence information for the IfOr inference and the OrIf inference of Experiment 3. "P(prem)" = premise probability, "P(concl)" = conclusion probability, "Observed" = observed coherence, "Chance" = chance-rate coherence, and "Above" = above-chance coherence. Error bars show 95% CIs.	145
Figure 5.10. Mean values of observed and above-chance coherence for the 12 inferences of Experiment 3, separately for each group and for three levels of measurement precision (see text for details). The black horizontal line represents a coherence rate of 0% in the panels for observed coherence, and it represents the chance rate of a coherent response in the panels for above-chance coherence. Error bars show 95% CIs.	149

Figure 5.11. Above-chance coherence for the inferences IfOr, MT, AC and DA of Experiment 3, separately for each group and premise probability condition. The horizontal line in the panels represents the chance rate of a coherent response. Error bars show 95% CIs.	151
Figure 5.12. Coherence intervals for biconditional AC and biconditional DA as a function of their premise probabilities.	155
Figure 5.13. Above-chance coherence for AC and DA of Experiment 3, in the original version of the inferences (left), in a version in which the conditional premise is substituted with a biconditional (middle), and in a version in which the conditional premise is substituted with its converse, <i>if q then p</i> (right). Error bars show 95% CIs.	155
Figure 5.14. Mean values of observed and above-chance coherence for the 12 inferences of Experiment 4, separately for each group and for three levels of measurement precision. The black horizontal line represents a coherence rate of 0% in the panels for observed coherence, and it represents the chance rate of a coherent response in the panels for above-chance coherence. Error bars show 95% CIs.	162
Figure 6.1. The conclusion probabilities used in Experiment 5 for each inference. The dots represent the conclusion probabilities, and the vertical lines represent the coherence intervals for each premise probability condition.	172
Figure 6.2. Proportion of "yes" and "no" responses to the inference <i>p, therefore not-p</i> with a premise probability of 1, observed during the practice trials.	177
Figure 6.3. Observed and above-chance coherence for the 12 inferences investigated in Experiment 5. Error bars show 95% CIs.	177
Figure 6.4. Above-chance coherence for Experiment 5 as a function of premise probability and whether the probability of the conclusion was at the edge of the coherence interval or clearly on one side of it. Error bars show 95% CIs.	181
Figure 6.5. Above-chance coherence for Experiment 5 as a function of premise probability and whether the probability of the conclusion was inside or outside of the interval. Error bars show 95% CIs.	190
Figure 6.6. Observed and above-chance coherence for the inferences in Experiment 6, excluding the data from the condition in which premise probability was 1 and the question was whether the probability of the conclusion could be higher. Error bars show 95% CIs.	200
Figure 6.7. Above-chance coherence for Experiment 6, separately for each premise probability and question condition. Higher: question of whether the probability of the conclusion can be higher than that of the premise (resp. for the two-premise inferences, whether it can be higher than .5). Lower: question of whether the probability of the	203

conclusion can be lower than that of the premise (resp. for the two-premise inferences, whether it can be lower than .5). Error bars show 95% CIs.

Figure 6.8. Observed and above-chance coherence for the 12 inferences of Experiment 7. Error bars show 95% CIs. 215

Figure 6.9. Above-chance coherence for binary instructions (Exp. 7) and probabilistic instructions (Exp. 6) when the question was whether the probability of the conclusion can be lower than the probability of the premise (resp. for the two premise inferences, whether it can be lower than 50%). The lower left corner of each panel shows the premise probability condition in Exp. 6 with which the data from Exp. 7 was compared to. Error bars show 95% CIs. 216

Figure 6.10. Above-chance coherence for binary instructions (Exp. 7) and probabilistic instructions (Exp. 6) when the question was whether the probability of the conclusion can be higher than the probability of the premise (resp. for the two premise inferences, whether it can be higher than 50%). The lower left corner of each panel shows the premise probability condition in Exp. 6 with which the data from Exp. 7 was compared to. Error bars show 95% CIs. 217

Figure 7.1. Coherence intervals for MP (upper row), DA (middle row), and and-to-if (lower row) as a function of premise probabilities. The shaded areas in the graphs represent the coherence intervals. 226

Figure 7.2. Distribution of conclusion probability judgments for MP as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals. 232

Figure 7.3. Distribution of conclusion probability judgments for DA as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals. 232

Figure 7.4. Distribution of conclusion probability judgments for &If as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals. 233

Figure 7.5. Distribution of conclusion probability judgments for inferences 4 and 5: DM and nDM, as a function of premise probabilities. The horizontal lines below the distributions show the location of the coherence interval for the condition that matches their colour. 233

Figure 7.6. Observed and above-chance coherence for the 5 inferences of the Experiment. Error bars show 95% CIs. 235

Figure 7.7. Conclusion probability judgments for MP, DA, and &If, as a function of premise probabilities. Error bars show 95% CIs. 236

Figure 7.8. Conclusion probability judgments for DM and nDM as a function of premise probabilities. The three lines in each panel display the three repetitions of each premise probability condition for these inferences. Error bars show 95% CIs.	236
Figure 7.9. Mean judgments of response confidence for MP, DA, and &If, as a function of premise probabilities. Error bars show 95% CIs.	238
Figure 7.10. Mean judgments of response confidence for DM and nDM as a function of premise probabilities. The three lines in each panel display the three repetitions of each premise probability condition for these inferences. Error bars show 95% CIs.	238
Figure 7.11. Mean probability judgments and judgments of response confidence for the conditions of Experiment 9. Error bars show 95% CIs.	243
Figure 8.1. Density curves showing the distribution of conclusion probability judgments. The shaded area represents the coherence interval for each inference and premise probability.	254
Figure 8.2. Mean values of observed and above-chance coherence for the probabilistically informative inferences. Error bars show 95% CIs.	256
Figure 8.3. Judgments of inference quality for the inferences investigated. The error bars show 95% CIs, and the grey lines in the background show the individual participant values.	258

PART 1. INTRODUCTION

CHAPTER 1. INTRODUCTION

Contents

- 1.1 Types of reasoning
- 1.2 Types of statements
- 1.3 Research questions
- 1.4 Outline of the thesis

TYPES OF REASONING

The present thesis is concerned with a specific form of thinking: reasoning. To reason is to produce one mental representation, called a conclusion, from one or more other mental representations, called premises. One says that the conclusion follows, or is inferred, from the premises, and the process of doing this is called drawing, or making, an inference. Reasoning can be distinguished from other forms of thinking, like associative thinking, in being directed: the relations between mental representations that it establishes are not always symmetrical. Reasoning can be distinguished from other forms of directed thinking, like creating a story, by the fact that it can be judged by epistemic norms: it makes sense to say that a reasoning outcome is correct or incorrect, whereas one could not say this of a story. An invented story can be of high or low artistic quality, realistic or unrealistic, interesting or boring, but not correct or incorrect. The premises and conclusion of an inference represent pieces of information, and the information in the conclusion is warranted to a higher or lower degree given the combined information in the premises.

Different forms of reasoning can be distinguished based on the type of relation established between premises and conclusion. The most common distinction made is that between deductive and inductive reasoning. An inference is deductive if and only if in each case in which the premises are true, or have a probability of 1, the conclusion is also true, or has a probability of 1. Said in another way, an inference is deductive if and only if it would be inconsistent for the premises to be true but the conclusion false. This feature of inferences has been called certainty preservation (Adams, 1996). An example of a deductive inference is *Modus Ponens* (MP): "If it is raining, the road will be muddy. It is raining. Therefore, the road will be muddy." Whenever you are certain that the two premises of this inference are true, you can also be certain that the conclusion is true, so that your certainty is preserved when going from the premises to the conclusion.

An inference is inductive if and only if it does not have the above property. That is, if it is not inconsistent for the premises to be true but the conclusion false. A conclusion can then still be likely given the truth of the premises, but it does not necessarily follow from the premises. An example of an inductive inference is: "It is raining. Therefore, the road will be muddy." The presence of rain can make it more likely that the road will be muddy, but does not necessarily imply that it will be muddy. After all, the road could be paved.

Inductive reasoning can be distinguished further into abductive reasoning and other forms of inductive reasoning. An inference is abductive if the truth of its conclusion is a good explanation for the truth of its premises. This explanation is often such that it postulates a causal link between premises and conclusion (Douven, 2011; Evans, Handley, Hadjichristidis, Thompson, Over, & Bennett, 2007; Oaksford & Chater, 2016; Oberauer, Weidenfeld, & Fischer, 2007). An example of an abductive inference is "The road is muddy. Therefore, it

rained". Other forms of inductive reasoning include category induction, e.g. "If a relatively large object moves steadily and swiftly along the road in the distance, it is probably a car ", enumerative induction, e.g. "On each occasion in which it has rained in the past, the road was muddy. Therefore, every time it rains, the road will be muddy", and analogical reasoning, e.g. "Cars are to roads like boats are to rivers".

This thesis will concern reasoning from declarative statements as premises to declarative statements as conclusions. These are statements providing *information* about events, e.g. "It is raining", as opposed to other speech acts, such as asking questions, e.g. "Is it raining?" But inferences concerning information about events can be of two kinds. They can be concerned with whether an event is the case, given that some other event is the case, or with whether an event should be the case, given that some other event is the case. Typically only the former are called declarative, whereas the latter are deontic. Deontic statements are statements about what is permissible and obligatory, e.g. in the context of a moral judgment or a legal rule. In this context, it is correct to infer that if a person should perform some action, then the person is allowed to perform the action. And if the person is not allowed to perform the action, then it is false that she should perform it. These inferences can be judged correct or incorrect without committing oneself to any specific moral or legal ideas, and it is in this sense that they can be viewed as declarative. This thesis is concerned only with declarative statements in the narrow sense: statements about what is the case or not, like "It is raining".

TYPES OF STATEMENTS

Statements like "It is raining" are often called atomic or categorical because they do not have components that are declarative sentences. But they can be combined, as components, to form compound statements by using connectives, some of which are logical. The logical connectives investigated in this thesis are *not*, *and*, *or*, and *if*, or more technically, negations, conjunctions, disjunctions, and conditionals. Formally, atomic statements are often represented through single letters like p and q (though p and q are variables that can also represent any complex statement). Negations can then be represented as *not- p* , conjunctions as $p \ \& \ q$, disjunctions as $p \ \text{or} \ q$, and conditionals as *if p then q* .

The correctness, or incorrectness, of inferences depends on the meanings we ascribe to the connectives. A typical way of characterising the meaning of the logical connectives is through a specification of the conditions under which the compound statements formed with the connectives are true. In classical logic, compound statements constructed with the above logical connectives are *truth functional*. Their truth or falsity is a function of the truth or falsity of their component statements. The simplest case is that of negation: *not- p* is true if and only if p is false. The truth conditions of the other compound statements can be conveniently

described in a truth table, which simply lists the truth or falsity of the compound statement for every possible combination of the truth or falsity of its components (Kneale & Kneale, 1962, p. 531). The truth tables for conjunction, disjunction, and conditionals in classical logic are shown in Table 1.1.

Table 1.1. The truth tables for conjunction, disjunction, and the conditional in classical logic.

	$p \ \& \ q$	$p \ or \ q$	$if \ p \ then \ q$
$p, \ q$	T	T	T
$p, \ not\text{-}q$	F	T	F
$not\text{-}p, \ q$	F	T	T
$not\text{-}p, \ not\text{-}q$	F	F	T

Note. "&" stands for *and*.

Table 1.1 shows that a conjunction is true whenever both of its elements p and q are true, and false otherwise. A disjunction is true in all cases except when p and q are both false, and a conditional is true in all cases except when p is true but q is false. This classical logical characterisation of the conditional is called the *material conditional*. The association of the truth table patterns in Table 1.1 with the natural language words *and*, *or*, and *if*, was established because, out of the 64 possible truth tables that can be built as a function of the truth or falsity of two statements p and q , those tables seemed to resemble most closely the meaning of these words. Elements of classical logic go back to Aristotle and other Greek philosophers, but it was specified more fully in the late 19th and early 20th centuries for the formalization of mathematical reasoning (Kneale & Kneale, 1962, Chs VIII-IX; Van Heijenoort, 1967). The exact relation between its formal connectives and the corresponding words of natural language has been a substantial research topic in philosophy, linguistics, and the psychology of reasoning.

There is now wide agreement in philosophy, linguistics, and psychology that at least the core meanings of natural language negations, conjunctions, and disjunctions, can be characterised through the truth tables used for them in classical logic, at least to a reasonable approximation (Douven, 2016). But this is far from true for conditionals. The natural language interpretation of conditionals has been a matter of controversy since Greek philosophers first started to study it (Kneale & Kneale, 1962, pp. 128-138). The situation is illustrated by a quote (from about 330 B.C.) attributed to the Hellenistic poet Callimachus: "Even the crows on the rooftops are cawing over the question as to which conditionals are true" (Kneale & Kneale, 1962, p. 128, see also Adams, 1998, pp. 4, 114). Different accounts of conditionals will be introduced below, but for the moment, suffice to say that none of the current main accounts of

conditionals proposes them to be represented as material conditionals (Baratgin, Douven, Evans, Oaksford, Over, & Politzer, 2015; Johnson-Laird, Khemlani, & Goodwin, 2015).

The question of how conditionals are interpreted is complicated by the fact that they can be divided into different types, and it is not clear whether the formulation of a unified account of all types of conditionals is possible (Bennett, 2003; Douven, 2016). The most important distinction made is between indicative conditionals on the one hand, and subjunctive or counterfactual conditionals on the other. Indicative conditionals are usually in the indicative mood, and are usually used in situations in which the antecedent is not known or believed to be false, e.g. "If it rained today, then the road is muddy". Subjunctives or counterfactuals are usually in the subjunctive mood, and are usually used in situations where the antecedent is either in the future, e.g. "If it were to rain tomorrow, then the road would be muddy", or the antecedent is known or believed to be false, e.g. "If it had rained yesterday, then the road would have been muddy".

A second relevant distinction is between singular and general conditionals. As the names suggest, singular conditionals refer to specific single events, e.g. "If it rains today, then this section of the road will be muddy", and general conditionals refer to classes of events, e.g. "If it rains, then roads are muddy". Generals are often described as "counterfactual supporting" (Edgington, 2011; Oaksford & Chater, 2010a). When a general conditional has high probability, there is a corresponding counterfactual that also has high probability, precisely because the relation described in the conditional holds in general, and not only on a specific occasion in which the antecedent happened to be true. There is currently little empirical research relating indicatives and counterfactuals (Cruz & Oberauer, 2014; Oaksford & Chater, 2013; Over, 2017; Pfeifer & Stöckle-Shobel, 2015; Thompson & Byrne, 2002), and more specific relations between the above two classifications still have to be established.

The present thesis will focus on singular indicative conditionals, and the use of the term "conditionals" will refer to singular indicatives unless otherwise specified.

Conditionals, singular and general, are central to the psychology of reasoning not only because the question of which inferences are valid depends on how the statements within them, including any conditionals, are interpreted, but also because every inference can be rephrased as a conditional, in which the antecedent p represents the premise (or if there is more than one premise, the conjunction of premises), and the consequent q represents the conclusion, and every conditional, *if p then q* , can be supported by an inference from p to q . Accounts of reasoning and of conditionals are therefore closely linked (Over & Baratgin, 2017).

RESEARCH QUESTIONS

A very influential and useful system of classification of research questions in experimental psychology is that by David Marr (1982). Drawing on an analogy to computer science, Marr distinguished three levels of analysis of a research topic: The computational, the algorithmic, and the implementational. The computational level of analysis asks *what* a person (or more generally a system) is aiming to do: what the person defines as the correct output to any given input. For example, if the person has the goal of finding the sum of two numbers, then the function of addition defines the correct output. If they wanted to find the difference of two numbers, a different function would apply. In reasoning research, this question translates into that of what people define as the correct conclusion in a reasoning problem: what normative principles they adhere to when deciding whether the conclusion of an inference is correct, or justified. For example, whether the normative principles people use when stating that an inference is correct or incorrect can be better modelled by classical logic or by probability logic (Oaksford & Chater, 2012).

The algorithmic level of analysis asks *how* the person carries out what they want to do: which representations and processes are used to get from the input to the output. Using the example of arithmetic, the person could sum up two numbers by adding their elements from right to left, and "carrying over" to the next position any amounts that sum to 1 or more at a given position. Or they could sum the two numbers from left to right, first adding the large components of the numbers, and then successively adding the smaller components. In reasoning research, the algorithmic level question is that of which mental representations and processes people use when reasoning, in which order these processes occur and how they interact. For example, people could arrive at the conclusion of an inference by building semantic representations of the premises in a way similar to truth tables. Then they could combine these semantic representations in a way similar to constructing a truth table for their conjunction, and then check whether the resulting representation is that of the conclusion (Johnson-Laird & Byrne, 2002). Alternatively, people could arrive at the conclusion by successively applying syntactic rules to the premises, and checking whether in this way the premises can be successively transformed until they become the conclusion (Braine & O'Brien, 1998; Rips, 1994).

At the implementational level, the question is how the representations and processes defined at the algorithmic level are implemented physically. In the arithmetic example, the sum could be implemented using a calculator or using paper and pencil. In the case of reasoning, the question is how reasoning processes are implemented in the brain (De Neys, Vartanian, & Goel, 2008; Goel, 2007, 2009; Oaksford, 2015; Prado, Chadha, & Booth, 2011).

The work in this thesis is focussed at the computational level of analysis. The general question is whether deduction plays a significant role in reasoning from uncertain premises.

When people evaluate the correctness of an inference on the basis of uncertain information, to what extent does it make a difference to them whether the inference is deductive? The two central deductive concepts in classical logic are consistency and validity. Two statements are consistent when they can both be true at the same time, and an inference is valid when as described above, it is inconsistent for the premises to be true but the conclusion false, so that it preserves certainty from premises to conclusion. These classical logical definitions are binary: they can only represent a statement as true or false, with no room for degrees of belief in between. This thesis describes a generalisation of consistency and validity to cover uncertain degrees of belief (Adams, 1998; Coletti & Scozzafava, 2002; Gilio, 2002) and investigates the role of these generalised concepts in reasoning. Consistency can be generalised to *coherence*, and validity to probabilistic validity, or *p-validity* for short. These extended definitions, described in more detail below, are studied within the framework of the *probabilistic approach* to the psychology of reasoning (Evans & Over, 2013; Oaksford & Chater, 2007; Over & Cruz, 2018; Pfeifer & Kleiter, 2009).

The investigation of the role of coherence and p-validity in reasoning from uncertain premises was accomplished using a specific probabilistic interpretation of the conditional, the probability conditional (Adams, 1998; Jeffrey, 1991), and new empirical evidence for this interpretation was obtained. However, most of the findings hold equally for the material conditional interpretation. This is because the valid inferences for the probability conditional are a subset of the valid inferences for the material conditional, so the material conditional generally poses weaker constraints on the conclusion.

OUTLINE OF THE THESIS

After this introductory section, the remainder of the thesis is divided into three further parts, covering eight chapters. Part 2, ranging through Chapters 2 to 4, provides an overview of philosophical and psychological accounts of reasoning and of the meaning of conditionals. It also describes how consistency and validity can be generalised to cover degrees of belief, why this is relevant for the interpretation of research findings and for the generation of new hypotheses within the probabilistic approach, and how the role of coherence and p-validity will be studied empirically in the subsequent chapters.

Part 3 is dedicated to the report of 10 experiments, covering Chapters 5 to 8. Chapter 5 presents four experiments that extend previous research on coherence to a range of further inferences of differing complexity, assessing whether people's responses are sensitive to the constraints of coherence more often than expected by chance. These experiments also examine the limits of people's sensitivity to coherence as a function of working memory load and as a

function of specific materials that have been found, in other settings, to lead to the incoherence of the conjunction fallacy (Tversky & Kahneman, 1983).

Chapter 6 reports three experiments that go beyond a binary assessment of whether or not the responses to an inference are coherent above chance levels, to a quantitative assessment of the extent to which responses are coherent above chance levels. It does so for the same inferences as those studied in Experiments 3 and 4 of Chapter 5. These experiments also assess whether the degree of above-chance coherence changes as a function of factors like the validity and the complexity of inferences. Further, these experiments make a direct comparison of coherence for responses under binary instructions (to assess whether the conclusion has to be true given the truth of the premises), and probabilistic instructions (to assess the probability of the conclusion, given the probabilities of the premises) using the same inferences, materials, and response format.

Chapter 7 describes two experiments that focus on people's sensitivity to the degree to which the probabilities of the premises constrain the probability of the conclusion to a coherent range of possible values.

Chapter 8 presents one experiment that focuses on the extent to which people are sensitive to the distinction between valid and invalid, i.e. between deductive and inductive inferences, over and above any sensitivity to coherence constraints for both types of inferences.

Finally, part 4 of the thesis, covering Chapter 9, provides a general discussion that summarises and relates the results of the 10 experiments to one another, outlines conclusions that can be drawn from these results, and how they can inform hypotheses for future research on different aspects of reasoning from uncertain premises.

PART 2. THEORETICAL BACKGROUND

CHAPTER 2. BINARY THEORIES OF REASONING AND THEIR ACCOUNTS OF CONDITIONALS

Contents

2.1 Classical logic

2.2 The material conditional

2.3 The truth conditions of the material conditional plus conditions of assertability: Grice and Jackson

2.3.1 Grice

2.3.2 Jackson

2.4 Possible world semantics: Stalnaker and Lewis

2.4.1 Stalnaker

2.4.2 Lewis

2.5 The triviality results

CLASSICAL LOGIC

Philosophers, and later psychologists, long thought of classical logic as the correct normative system for human reasoning. This view can be traced back at least as far as Kant (1781/1998). When outlining the principles of classical logic of the time, and providing an account of relations between them and basic laws of thought, Kant argued that the field of formal logic was completed and so not open to change.

This view of classical logic as something definite and immutable started to change drastically with the surge in developments in logic and mathematics in the late 19th and 20th centuries, leading to a range of different proposals that extended or revised classical logic in various ways, allowing the expression of a wider variety of concepts. Some of the new logical systems actually conflicted with classical binary logic (Edgington, 1995; Priest, 2008; Van Heijenoort, 1967).

In the philosophical literature on reasoning, classical logic, as rigorously developed into a truth functional system, was initially retained as the normative system for reasoning with the logical connectives "and", "or", "not", and "if". But researchers increasingly argued this logic did not fully capture their intuitions about conditionals in natural language (Bennett, 2003; Edgington, 1995).

THE MATERIAL CONDITIONAL

The conditional in classical logic is called the *material conditional*. Like all connectives in classical logic, it is *truth functional*, which means that its truth conditions can be fully described as a function of the truth conditions of its component statements. For example, consider the conditional "If it rained, then the road was muddy". If this conditional is material, then its truth or falsity can be fully determined by knowing the truth or falsity of its components "It rains" and "The road is muddy". The conditional is true when it rained and the road was muddy; false when it rained but the road was not muddy; true when it did not rain and the road was muddy; and true when it did not rain and the road was not muddy. These four logical possibilities for combining the truth or falsity of rain with the truth or falsity of the road being muddy are represented in the truth table of the material conditional, shown in Table 2.1. One can see that this conditional is true in every case except that in which its antecedent is true and its consequent false. This interpretation of the conditional is therefore equivalent to *not both p and not-q*. It is also equivalent to the disjunction *not-p or q*.

The fact that the material conditional is truth functional, and so can be fully described through its truth table, makes it a relatively simple and clear interpretation of conditionals, and one that can easily be connected with other logical connectives to form more compound

sentences (e.g.: "If it rained then the road was muddy, *and* if the road was muddy then he got delayed"). This simplicity makes the material conditional attractive theoretically, providing a reason for trying to retain it, and for attempting to explain away aspects in which it seems implausible as a meaning for everyday conditionals.

Table 2.1. The truth table for the material conditional.

	$p \supset q$
p, q	T
$p, \text{not-}q$	F
$\text{not-}p, q$	T
$\text{not-}p, \text{not-}q$	T

Note. T stands for "true" and F for "false". The horseshoe symbol \supset represents the material conditional.

One major aspect that makes the material conditional implausible as an account of natural language conditionals is that it is true whenever the consequent is true, and whenever the antecedent is false. Suppose some random road is dry. Then referring to this road, the following inferences are valid if the conditional in their conclusion is material:

- (1) "The road is dry. Therefore if it rains, then the road is dry"
- (2) "I will drive into town tomorrow morning. Therefore, if the car gets stuck in the mud, then I will drive into town tomorrow morning."

But intuitively both seem rather implausible. One could go further and use inferences that seem contradictory: "The road is dry. Therefore if the road is wet, then it is dry". The strong implausibility of these two inferences, *q, therefore if p then q*, and *not-p, therefore if p then q*, has led them to be called *paradoxes of the material conditional*.

To circumvent the counterintuitive implications of using the material conditional to characterise natural language conditionals, the material conditional has been complemented with pragmatic principles about how people use it in communication and everyday reasoning situations. In this way, a distinction was drawn between the meaning of a statement on the one hand, and the assertability, or acceptability, of the statement in a particular communicative context on the other. A prominent attempt to retain the material conditional account of conditionals, while providing an explanation for the above paradoxes in terms of pragmatic aspects of communication, was offered by Paul Grice.

THE TRUTH CONDITIONS OF THE MATERIAL CONDITIONAL PLUS CONDITIONS OF ASSERTABILITY: GRICE AND JACKSON

Grice

Grice (1989) held that the meaning people ascribe to conditionals *if p then q* is that of the material conditional $p \supset q$, and that therefore the two paradoxes above are valid inferences. He argued that the reason for why the paradoxes appear incorrect is not semantic but pragmatic. For Grice, it is pragmatically infelicitous to assert the paradoxes because they violate a general principle of being cooperative when communicating with others, as one would be expected to in a conversation. He characterised this *cooperative principle* as "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." Grice argued that this general principle can be fulfilled by observing a set of more specific principles, or maxims, that fall into four categories:

- Category of quantity: Be as informative as is required for the purposes of the exchange, but not more informative than is required.
- Category of quality: Try to be truthful (e.g. by not saying what you think is false or for which you lack evidence).
- Category of relation: Be relevant.
- Category of manner: Make your contribution clear, orderly, and brief. This relates not to what is said but to how it is said.

Grice thought that in general people tend to follow these maxims, e.g. as a habit acquired through socialisation. But he also held that, by following them, conversations and communication more generally could be conceived as a special case of purposive, *rational* behaviour. Following the maxims makes it possible for us to understand each other even when what we aim to convey in a conversation is not always made fully explicit through the semantic content of what we say. Some of the things we aim to convey may be implicit, but our conversation partners can draw on the cooperative principle and its maxims to infer what we *implicate* in addition to what we overtly say. For example, when El Padrino talked about making someone an offer that they could not refuse, he was not implicating that it was a generous offer, but was rather making a threat that was too dangerous to ignore. His *conversational implicature* worked because his conversation partners knew what he was implicating in what he said, and he knew that they knew.

However, what is implicated by an utterance in a conversation can also be used with the aim to be misleading. For example, if, asked about Mary's whereabouts, we reply "She's in the library or in town", our hearers may infer that we do not know which of these two possibilities

is the case, for if we knew that she was, for instance, in town in a pub, it would have been more informative to say so directly. If we know she is in the pub and want to cover for her, then our response is not untruthful, but it violates the maxim of informativeness.

Grice held that although the natural language conditional *if p then q* is the material conditional, and so is true whenever *p* is false or *q* is true, it would be misleading to assert *if p then q* in a conversation in situations in which we could be more informative by asserting *q* directly, or *not-p* directly. Hence for Grice the paradoxes are valid, but it is infelicitous to assert them because they violate the maxim of informativeness.

In addition to the paradoxes, Grice (1989) discussed a number of further problems for his view that the natural language conditional *if p then q* is the material conditional. For example, *if p then q* is non-commutative, i.e. it is generally not equivalent to *if q then p*. This is in contrast to *not-p or q*, which is generally considered equivalent to *q or not-p*. A related problem is that the use of *if p then q* often seems similar to that of an inference from premises *p* to a conclusion *q* (see Bennett, 2003; and Over & Baratgin, 2017, on the *inference ticket* account of conditionals), and is also often used as the major premise of a MP inference. In neither case does it seem relevant to take into account the two possibilities of the truth table for the material conditional in which the antecedent is false. A third problem discussed by Grice is that the intuitive result of negating a conditional, *if p then q*, does not seem to be that of the negation of a material conditional, *p & not-q*, but instead that of the negation of the consequent of the original conditional, *if p then not-q*. A fourth problem for the material conditional is that it renders valid the inference of *contraposition*: *if p then q, therefore if not-q then not-p*. Contraposition can also be applied in the other direction, *if not-q then not-p, therefore if p then q*, so that for the material conditional, *if p then q* and *if not-q then not-p* are logically equivalent. Logically equivalent statements have the same probability, but intuitively we can have different degrees of confidence in each. Grice (1989) remarked that the above features of the use of conditionals make them appear more similar to the statement *supposing p, (then) q*, than to the statement *not-p or q* from the material conditional.

Grice (1989) argued that his pragmatic account of the paradoxes in terms of the principle of cooperative communication was not enough to solve the additional problems for the material conditional that he outlined. To account for these additional problems he introduced a second pragmatic aspect of the use of conditionals in communication.

Grice assumed that each logical connective has, in addition to its semantic meaning, a specific recurring role in practical discourse. Negation is argued to be essential for the ability to express many matters of fact, and is not associated with a particular use beyond this essential function. Conjunction can be substituted by concatenating separate sentences, but is useful to be able to negate groups of sentences without having to specify the truth value of each of its elements (e.g. *not(p and q and r)* as opposed to having to specify separately whether each of *p*, *q*, and *r* are the case). The role of disjunction is argued to be that of

entertaining preliminary possibilities when trying to find out what is the case. These alternative possibilities can then be ruled out successively until one remains, in a way less cumbersome than would be required through the use of only conjunction and negation (e.g. one can then state *p or q or r* instead of the more unwieldy *not(not-p and not-q and not-r)*).

The role of conditionals in communication is argued to be similar to that of drawing inferences in scientific enquiry, in which one assumes *p* to be the case to see whether *q* follows, and may go on to assume that *q* is the case to see whether *r* follows, and so forth, building a chain of argument. If it then later turns out that *p* is the case, one can use this chain to infer, by MP, that *q* must be the case too. Grice called this use *interrogative subordination*, and represented it as a *bracketing device* around the antecedent, [*if p then*] *q*, indicating that the truth of *p* is common ground or has been assumed, and that one can focus one's attention on whether *q* is true under this assumption. This use of conditionals was argued to be based on *conversational implicatures*, i.e. features we can infer the speaker to have conveyed when asserting the material conditional, in addition to the explicit meaning stated through this conditional.

Grice argued that the feature of interrogative subordination was not specific to conditionals, but can be applied also to conjunctions and disjunctions. For example, in a conversation about the disjunction *p or q*, it could be common ground that *p* is highly probable, and the discussion could then focus on whether it is reasonable to believe that the disjunct *q* is highly probable too.

Grice's suggestion that conditionals are used in discourse to follow chains of questions, in a way similar to research questions, and that people do so by assuming the antecedent is common ground, and focussing on whether the antecedent holds under this assumption, arguably sounds intuitive. However, it seems less clear how this role is at the same time a role for the use of conditionals, and a procedure that can be applied equally to conjunctions and disjunctions in addition to the role Grice ascribed to them. The proposal of two elaborate pragmatic principles, the maxim of informativeness and the role of interrogative subordination, to supplement the material conditional as the semantics of natural language conditionals, also raises the question of what role each of these components plays in different situations, and whether there might be an alternative, more parsimonious way of accounting for conditionals.

Jackson

Like Grice, Jackson (1987) assumed that natural language conditionals are logically equivalent to material conditionals, but that the material conditional is supplemented with factors determining when it is justified to assert it in discourse. He consequently called his account

supplemented equivalence theory. But Jackson differed from Grice regarding the factors proposed to affect the assertability of conditionals.

Jackson argued that Grice's account of the paradoxes in terms of a violation of the maxim of quantity (being informative) does not explain why some instantiations of the paradoxes are more assertable than others. For example, if we just walked along a muddy road, then it seems assertable to say "if it rains today, the road is muddy; and if it doesn't rain today, the road is muddy", simply because we know the road to be muddy. It also does not explain why statements that by the material conditional should be logically equivalent, like *if p then q, not-p or q, and if not-q then not-p*, can differ in assertability.

To address the above problems, Jackson argued as follows. The conditions under which natural language conditionals are true are those of the material conditional, but two additional conditions must be met for a conditional to be assertable. First, the probability of the material conditional must be high enough to warrant assertion, and second, this probability must be robust with respect to the antecedent, which means we are only justified in asserting the conditional if we would still have a high degree of belief in it after learning that its antecedent is true. On this account, the paradoxes are valid but not assertable, because the conditionals in the conclusion may not be robust with respect to their antecedents. For example, if we assume the conditional "If it rains, then the road is dry" is a material conditional, then it is highly probable as long as it does not rain. But this probability is not robust with respect to its antecedent: the conditional becomes very improbable whenever it starts to rain.

Jackson proposed a very similar account for the assertability of disjunctions. He argued that the conditions under which they are true are those of classical logic, but that a disjunction is only assertable when it is highly probable, and this probability is robust with respect to both of its disjuncts. For example, the statement "Mary is in the library or out in town" is not assertable when the only reason we have for its high probability is that we think Mary is in a pub in town. If to our surprise, we were to learn that she didn't go to the pub in the end, the disjunction ceases to be probable (c.f. Gilio & Over, 2012; and Over, Evans, & Elqayam, 2010, for a similar distinction between disjunctions that are justified constructively based on one of their elements, and disjunctions that are non-constructively justified, and hence robust with respect to the probability of either of their elements taken on its own).

Why make the assertability of conditionals robust with respect to the antecedent? Jackson (1987) explains that this makes it possible to use conditionals to draw MP inferences. If a conditional is no longer probable when we learn that its antecedent is true, then that conditional is not useful for drawing MP inferences.

The proposal is that a conditional is assertable if and only if its probability, as a material conditional, is high and this probability is robust with respect to its antecedent. This claim can be described formally as the condition that, not only $p(p \supset q)$ is high, but that $P(p \supset q|p)$ is high as well. But $P(p \supset q|p) = P(q|p)$. This in turn implies that, in the supplemented

equivalence account, a conditional is assertable if and only if $P(q|p)$ is high. This brings the assertability of conditionals in Jackson's account close to that of Adams (1975), whose work Jackson tried to take into account when developing his proposal. Adams' account will be described further below.

Jackson argued that by making the assertability of a material conditional a function of $P(q|p)$, his account captures the intuition that conditionals are hypothetical rather than categorical statements. He considered this intuition to hold for both indicative and counterfactual conditionals, but incorporated it into his account in different ways for the two types of conditionals. Whereas for indicatives it is captured by making their assertability dependent on $P(q|p)$, for counterfactuals it is captured by his endorsement of the possible world accounts of conditionals of Lewis (1973) and Stalnaker (1968), described further below.

Jackson's robustness condition can also be related to Grice's description of the role of conditionals as one of interrogative subordination, and the bracketing device proposed to implement this description. However, whereas for Grice the bracketing device was a conversational implicature because it could be applied also to conjunctions and disjunctions, Jackson's robustness condition would be, in Grice's terms, a conventional implicature, because it was proposed to be specific to the use of conditionals.

Jackson's position has been criticised on a number of points (Edgington, 1995), but a telling question (raised by Appiah, 1984) is what role the material conditional is playing in the meaning of conditionals, if the intuitive understanding and use of conditionals is already captured by the conditional probability. In the interest of parsimony, it would be useful to find a specific role for the material conditional in people's actual interpretation of conditionals, beyond the fact that the material conditional is theoretically simple because it is truth functional. Jackson's attempt to reply to this criticism appears to be based on a conflation between what makes a conditional true or believable, and what makes a conditional assertable in discourse. In his own theory Jackson stated that the (degree of) assertability of *if p then q* is equal to $P(q|p)$, but also that *if p then q* is assertable if and only if $P(q|p)$ is high, which are not the same (see Over & Cruz, 2018, on this difference for probability conditionals like that of Adams).

POSSIBLE WORLD SEMANTICS: STALNAKER & LEWIS

Stalnaker

For Stalnaker (1968) the paradoxes of the material conditional are a reason for rejecting this conditional as an account of the truth conditions of the natural language conditional. For him a

conditional *if p then q* expresses a proposition that is a function of two other propositions *p* and *q*, but which is not a *truth* function of these other propositions.

He briefly discusses the alternative that a conditional expresses a logical or casual connection between *p* and *q*, so that instead of looking at the truth or falsity of *p* and of *q*, we just look at whether there is a connection between them. If there is, we say the conditional is true, and if there is not, we say it is false. But he dismisses this view by noting that a conditional can be true in cases in which antecedent and consequent are independent, for example because we are convinced that the consequent is true, whatever the value of the antecedent.

The account advocated by Stalnaker is based on a characterisation of conditionals offered by Ramsey (1929/1990). In a hugely influential footnote of this paper, Ramsey states:

"If two people are arguing 'If *p* will *q*?' and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge and arguing on that basis about *q*; [. . .] We can say they are fixing their degrees of belief in *q* given *p*. If *p* turns out false, these degrees of belief are rendered *void*."

Stalnaker points out that Ramsey's suggestion, known as the *Ramsey test*, covers only cases in which we have no fixed prior degree of belief in *p*. But he notes that it readily applies also to situations in which we believe *p* to be certain, because we can then go on to assess our degree of belief in *q* directly. Further, Stalnaker extends the Ramsey test to counterfactuals by including the provision that when *p* is added hypothetically to our stock of beliefs, we make any necessary adjustments to these beliefs to maintain consistency, before proceeding to assess our degree of belief in *q*.

Stalnaker wished to complement this epistemological account of how we decide whether or not to *believe* a conditional statement, with an ontological analogue of whether or not a conditional statement is *true*. To this end, he drew on the concept of a *possible world*, suggesting that it is the ontological analogue of a hypothetical belief. By drawing on possible worlds, Stalnaker departed from classical logic, which had been used by Jackson and Grice, and proposed an extension of modal logic (Kripke, 1963) as an alternative¹ (see also Adams, 1996, footnote 9).

Roughly, a possible world can be viewed as a hypothetical state of affairs in the world. If a statement is true, then it is not just conceivable hypothetically but holds in the actual world. If a statement is false, then it does not hold in the actual world but it may hold in a hypothetical alternative state of affairs, i.e. it may hold in a counterfactual world. A logical

¹ It is an extension of modal logic for the following reasons. Modal logic provides a way of expressing what is true in the actual world, in all possible worlds, or in at least one, unspecified world. The addition of the conditional proposed by Stalnaker makes it possible to express also what is true in particular, non-actual possible worlds. That is, it makes it possible to express counterfactual statements.

tautology is a statement that is true in all possible worlds, and a contradiction a statement that is false in all possible worlds (i.e. a contradiction refers to an isolated impossible world).

Stalnaker assumed that all worlds that are possible, or *accessible*, with respect to the actual world, can be ordered in terms of their similarity to the actual world. The actual world is most similar to itself, and the ordering of the other worlds is a *total ordering*, i.e. there is a single closest world β to the actual world α , and a single closest world γ to world β , and so on, with no ties. This is in contrast to Lewis (1973), who did allow ties, so that two or more worlds could be equally close to the actual world.

Using the above features of possible worlds, a conditional in Stalnaker's account is true if and only if the closest possible world to the actual world in which the antecedent is true, is a world in which the consequent is also true. If the consequent is false in the closest possible world to the actual world in which the antecedent is true, then the conditional is false. And if there is no possible world in which the antecedent is true, then the conditional is vacuously true.

The above implies that in Stalnaker's system, a conditional is always true or false – but the question of when it is true and when false depends not on the truth or falsity of its component statements in the actual world (as for the truth functional material conditional), but on the truth or falsity of the consequent in the closest possible world to the actual world (if there is one) in which the antecedent is true.

The extended modal logic in which Stalnaker's conditional is embedded differs from classical logic not only in the fact that it can express possibility/impossibility in addition to truth and falsity, but also with regard to which inferences containing conditionals are valid. In particular, the paradoxes of the material conditional are invalid, and so do not need to be explained away as in the systems of Grice and Jackson. For example, suppose I am sitting at my computer and writing. What is the truth value of "If I were walking, then I would be writing"? In the closest possible world, to the actual world, in which I am walking, I am not writing. This conditional is then false on Stalnaker's analysis, as it is intuitively.

Table 2.2. Examples of inferences that are invalid in the modal logical systems of Stalnaker (1968) and Lewis (1973), but valid in classical logic.

Name	Form
1 Strengthening the antecedent	$if\ p\ then\ r\ :\ if\ p\ \&\ q\ then\ r$
2 Transitivity	$if\ p\ then\ q,\ if\ q\ then\ r\ :\ if\ p\ then\ r$
3 Contraposition	$if\ p\ then\ q\ :\ if\ not-q\ then\ not-p$

Note. \therefore = "therefore".

Further inferences that are valid in classical logic but invalid in Stalnaker's system are *strengthening of the antecedent*, *transitivity*, and *contraposition*, shown in Table 2.2.

Note that the above inferences all have conditionals in the conclusion. If the antecedents of these conditionals were moved to the premises, then the resulting inferences would still be valid in Stalnaker's system (c.f. Edgington, 1995, p. 286). Thus *if p then r, p & q, therefore r* is a trivially valid inference whose second premise guarantees that the relation between *p* and *r* is robust with respect to *q*. The inference *if p then q, if q then r, p, therefore r*, is similar to transitivity and still valid; and modus tollens (MT): *if p then q, not-q, therefore not-p* remains valid.

Lewis

The possible world semantics of Lewis (1973) is very similar to that of Stalnaker (1968), and it leads to the same changes in the validity of inferences shown in Table 2.2. The main difference between the two systems is that as mentioned above, Lewis did not assume there to be an absolute ordering of worlds, with a single closest world to the actual world in every instance. Instead, he allowed there to be sets of worlds with the same degree of similarity to the actual world. He described these sets of worlds as nested spheres with the actual world at the centre, in a way similar to a planetary system. The smallest sphere around the actual world would contain the alternative possible worlds most similar to it, and successively larger spheres would contain successively more dissimilar worlds. He described the similarity relation between spheres as resembling topological altitude regions in a map, with worlds within the same sphere being equally similar to the actual world, and worlds in a larger sphere being more distant from the actual world by the same amount, creating an ordinal rather than interval scale.

A further difference between the systems of Stalnaker (1968) and Lewis (1973) is that Stalnaker used it to represent both indicative and counterfactual conditionals in a unified approach, thinking of the difference between the two types of conditionals as pragmatic and not very consequential. In contrast, Lewis applied his possible world semantics only to counterfactuals, viewing indicatives as material conditionals in a way similar to that of Grice (Lewis, 1976). A third difference is that, although formally the Stalnaker conditional is equivalent to a special case of the Lewis conditional in which there is a unique closest sphere (the so-called limit assumption in Lewis) and every sphere contains a single world, the two conditionals differ in the theoretical background in which they are placed. Stalnaker viewed his conditional as an ontological analogue of the epistemological conditional expressed by the Ramsey test, whereas Lewis viewed his conditional more in the tradition of the material conditional.

Specifically, Lewis described his counterfactual conditional as a *variably strict conditional*, $vs(not-p \text{ or } q)$ ². Lewis' application of this conditional not just to single possible worlds, as in Stalnaker, but to whole sets of possible worlds, allowed Lewis to formulate two versions of it, one version for "would" counterfactuals, as in "If it had rained today, the road would have been muddy", and another version for "might" counterfactuals, e.g. "If it had rained today, the road might have been muddy".

To establish whether a would-counterfactual is true, we search for the smallest sphere S that contains at least one world in which the antecedent p is true, i.e. that contains at least one p -world. Then the would-counterfactual is true if and only if in all the worlds of this sphere S , either p is false or $p \ \& \ q$ is true. To establish whether a might-counterfactual is true, we again search for the smallest sphere S that contains at least one p -world. Then the might-counterfactual is true if and only if there is at least one world in the sphere S in which either p is false or $p \ \& \ q$ is true. If there are no accessible spheres in which p is true, then both counterfactuals are vacuously true. Would-counterfactuals can be represented as vs -necessarily($not-p \text{ or } q$), and might-counterfactuals as vs -possibly($not-p \text{ or } q$). The distinction between would- and might-counterfactuals mirrors the distinction between cases in which we are certain that a conditional is true, and cases in which we just have an uncertain degree of belief that it is true.

One criticism of possible world accounts like those of Lewis and Stalnaker is that it is not clear how similarity between worlds can be measured, particularly if these worlds are not observable objects or events in this world but abstract possibilities. This is in spite of the fact that ordinary people do sometimes have intuitions about differences in closeness between counterfactual possibilities, for example experiencing more regret or frustration when they "just" failed to catch the train than if they were late by a larger amount of time (Byrne, 2016). Another criticism is based on a proof by Lewis (1976), which is described below.

THE TRIVIALITY RESULTS

Stalnaker (1970) sought to connect his possible worlds account of when a conditional is true or false (Stalnaker, 1968) with an account of people's degree of belief conditionals. Following the Ramsey test, he argued that the probability of the conditional if p then q in his logical system was the conditional probability of q given p , $P(\text{if } p \text{ then } q) = P(q|p)$. This equality is so important in probabilistic theories of conditionals that it has simply been called *the Equation* (Edgington, 1995). But Lewis (1976) later proved that the Equation does not hold in systems like his and Stalnaker's, in which the conditional is a proposition that is always true or false at each possible world. He proved that the Equation could only hold for conditionals like his and

² See Cariani & Rips, 2017, for a recent reasoning account using this conditional.

Stalnaker's in certain "trivial" cases, e.g. when the conditional is *if p then p*. This led his and related proofs to be known as the "triviality results".

The proof has also been termed the "bombshell" (Edgington, 1995) because of its significance for theories about the meaning of conditionals. It shows that there is no fully truth conditional³ account of *if p then q* which is such that $P(\text{if } p \text{ then } q) = P(q|p)$ for any coherent probability assignments to *p* and to *q*. Hence if our degree of belief in conditionals is in accordance with the Equation and the Ramsey test, then conditionals cannot be such full propositions.

To illustrate the argument of the proof, consider for example the truth table in Table 2.3, for a conditional about a fair coin: "If the coin is flipped (*F*), then it will land heads (*H*)". Suppose the coin is unlikely to be flipped, so that $P(F \ \& \ H) = P(F \ \& \ \text{not-}H) = .1$, and $P(\text{not-}F \ \& \ H) = P(\text{not-}F \ \& \ \text{not-}H) = .4$.

Table 2.3. Truth table for the example about a coin flip.

<i>F, H</i>	.1
<i>F, not-H</i>	.1
<i>not-F, H</i>	.4
<i>not-F, not-H</i>	.4

What is the probability of this conditional? Given that the coin is fair, the intuitive answer is that $P(\text{if } F \text{ then } H) = P(H|F) = .5$. This is the result we get when we consider only the cases of the truth table in which *F* is true: $.1/(.1+.1) = .5$. If the conditional were material, we would instead have $P(\text{if } F \text{ then } H) = P(\text{not-}F \ \text{or } H) = .1 + .4 + .4 = .9$. If the conditional is a Stalnaker or Lewis conditional, then it will be either true or false in the two false-antecedent cases, depending on whether *F & H* or *F & not-H* is the closest possible world to the *not-F* worlds. As an illustration, suppose the *F & not-H* world is certain to be the closest possible world to both *not-F* worlds. Then $P(\text{if } F \text{ then } H) = .1$. As this example illustrates, there are cases in which neither the probability of the material conditional nor the probability of the Stalnaker/Lewis conditionals matches the intuitively correct answer of .5 given by the conditional probability.

³ A truth conditional statement is one that is always either true or false for each case of its truth table, i.e. for each combination of the truth or falsity of its component statements. It differs from a truth functional statement in that the value it takes (true or false) is not uniquely determined by the truth values of its component statements. For example, for the Stalnaker and Lewis conditionals, it will depend on whether or not the consequent is true in the closest possible world to the actual world in which the antecedent is true (c.f. Adams, 1998, Ch. 8).

Generally, the conditional probability $P(q|p)$ depends only on the cases of the truth table in which p is true: if we fix the probability of those two cases, we fix $P(q|p)$. In contrast, the probability of a full proposition, be it a material conditional or a Stalnaker/Lewis conditional, is affected also by the cases of the truth table in which p is false. The two interpretations will therefore only necessarily agree in "trivial" cases, as when the conditional is a logical truth (Lewis, 1976; see also Bennett, 2003; and Edgington, 1995, for reviews of further triviality proofs by a number of authors).

The triviality results entail that, if we want to hold on to the Equation as the basis for the interpretation of natural language conditionals, we must conclude that conditionals are not propositions with full truth conditions. Lewis (1976, p. 305) said he had no conclusive objection to this conclusion, but rather an inconclusive one: "The hypothesis requires too much of a fresh start. It burdens us with too much work still to be done, and wastes too much that has been done already." But as we will see below, long before Lewis wrote those words, de Finetti had made the "fresh start", and other researchers have continued to develop his ideas, based on the Equation.

CHAPTER 3. PROBABILISTIC THEORIES OF REASONING AND PROBABILITY CONDITIONALS

Contents

- 3.1 Why represent degrees of belief with probabilities?
- 3.2 Which interpretation of probability?
 - 3.2.1 The frequentist interpretation
 - 3.2.2 The logical interpretation
 - 3.2.3 The subjectivist interpretation
- 3.3 How can we measure degrees of belief, and why would we want them to be coherent?
 - 3.3.1 Measuring beliefs by measuring actions
 - 3.3.2 Dutch book arguments
- 3.4 Coherence and p-validity: Deduction from uncertain premises
- 3.5 Probability conditionals
- 3.6 Conditionals and validity
- 3.7 Uncertain reasoning beyond deduction: Dynamic reasoning
- 3.8 Empirical evidence for the probabilistic approach
 - 3.8.1 Evidence for reasoning from uncertain premises
 - 3.8.2 Evidence for the probability conditional

The triviality results contributed to the development of probabilistic accounts of conditionals, first in philosophy and then later in psychology. The general change from classical binary logic to probability logic as the normative, computational level framework for modelling *what* people set out to accomplish when engaged in reasoning, has been described as a paradigm shift (Over, 2009). With its origins in philosophy (Adams, 1998; de Finetti, 1936/1995; Jeffrey, 1991; Ramsey, 1926/1990), the new probabilistic paradigm in psychology (Evans, 2006; Evans & Stanovich, 2013; Oaksford & Chater, 2007, 2013; Over, 2016; Pfeifer & Kleiter, 2009, 2010; Politzer & Baratgin, 2016) is based on the fundamental hypothesis that most reasoning, both in everyday life and in science, is from premises that are uncertain. The uncertainty, or degrees of belief, in the statements from which we reason cannot be modelled in classical or modal logic. But it can be in probability theory.

In the probabilistic approach advocated here, it is proposed that the probability of singular indicative conditionals is given by the Equation, $P(\text{if } p \text{ then } q) = P(q|p)$, with the consequence, by the triviality results, that conditionals are not propositions with full objective truth conditions. To have a certain degree of belief in a conditional is considered different from having a certain degree of belief about a matter of fact in the world (Edgington, 1995). The proposal is that people arrive at this conditional probability, not by first having some specific degrees of belief in $P(p \ \& \ q)$ and in $P(p)$, and then taking their ratio, $P(q|p) = P(p \ \& \ q)/P(p)$, but instead through a *Ramsey test* (Ramsey, 1929/1990; Stalnaker, 1968). That is, a mental simulation in which people hypothetically suppose the antecedent p of the conditional, *if* p *then* q , make any necessary changes to their beliefs to preserve consistency, and assess the probability of q under this supposition.

Yet another central aspect of the probabilistic approach is that it shares the computational level framework of probability theory with neighbouring areas of research, covering not just deductive but also inductive reasoning, as well as judgment and decision making, argumentation, and learning (Elqayam & Over, 2013; Hadjichristidis, Sloman, & Over, 2014; Hahn & Oaksford, 2007; Oaksford & Chater, 2013; Zhao & Osherson, 2014) leading to a stronger integration and cross-pollination of these fields. All these points about the probabilistic approach are described in more detail in what follows.

WHY REPRESENT DEGREES OF BELIEF WITH PROBABILITIES?

Some authors have argued that probability theory is too precise and fine-grained to serve as a realistic representation of degrees of belief, which are intuitively more coarse or vague than point-probabilities (Khemlani, Lotstein, & Johnson-Laird, 2015; Politzer & Baratgin, 2016; Spohn, 2013). But although degrees of belief can definitely be coarse grained, and may also change in precision with context, it is difficult to ascertain this precision a priori. If we choose

a more coarse-grained system, e.g. a Likert scale with 5 subdivisions, then we will not be able to tell whether a scale with 7 or 10 subdivisions would still have measured meaningful differences in people's degrees of belief. The use of probability theory to represent and measure degrees of belief does not imply a commitment to the idea that people will have analogues of point probabilities in their minds for every statement they may be confronted with. Probability theory can instead be treated as an approximation with which to work. As such, it has the unique advantage of making it possible to measure the coarseness of people's degrees of belief empirically for a given context, instead of having to presuppose a certain degree of coarseness from the outset. Similarly, there may well be cases in which the use of ordinal instead of interval scales to measure beliefs will reflect more closely the kind of relationships people establish between their beliefs. But we cannot find out unless we use a measurement system precise enough to capture both probabilities and ranks. For example, Tversky & Kahneman's (1979) hugely influential prospect theory about relations between degrees of belief and utilities would arguably have been much more difficult to develop if degrees of belief had been represented as ranks rather than probabilities. This argument was followed up further in some of the experiments of the thesis.

WHICH INTERPRETATION OF PROBABILITY?

In pure mathematics, the term "probability" simply refers to a non-negative, additive set function that takes a maximum value of 1. But this definition, while clear and precise, does not help us understand how the term is applied to the world, e.g. in empirical research, by insurance companies, or in everyday assertions like "Jane will probably catch her flight" (Kyburg & Smokler, 1980). There have been three main proposals for connecting the mathematical function with its use in statements about world: the empirical or frequentist, the logical, and the subjective (Kyburg & Smokler, 1980; see also Hájek, 2012).

The frequentist interpretation

In its first formulation, the empirical or frequentist conception identified probability with the limit of a relative frequency (Reichenbach, 1949; Venn, 1886; von Mises, 1951). The probability of a coin landing heads was held to be equal to the limit of the relative frequency of heads among the tosses, as the number tosses is increased towards infinity (Kyburg & Smokler, 1980). This interpretation runs into problems, for instance because it is difficult to explain the step from a sample frequency to a probability without a further conception of probability that is not itself based on frequencies (de Finetti, 1937/1980, p. 110; Hájek, 2012).

An alternative version of the frequentist interpretation is as a theoretical concept that receives its meaning through the rules or procedures through which it is applied (Braithwaite, 1953), in particular the rules for rejecting a statistical hypothesis given the evidence of a sample with specific characteristics (Fisher, 1956; Neyman, 1952).

Common to the above versions of the frequentist, empirical position is that a probability statement is an assertion about the world, like an assertion about length or weight, which can be true or false. The evidence for it is observational: in order to find out whether a probability statement is true or false, we must conduct an empirical investigation. This investigation will lead to one or more sample frequencies, e.g. a coin cannot be tossed indefinitely, and so will not usually terminate with certainty, but with a probability judgment, in some other sense, about the objective frequency (Kyburg & Smokler, 1980).

The logical interpretation

In contrast, the logical conception of probability holds that probabilities are not empirical but logical statements: statements about the logical relation between a hypothesis (one statement) and a body of knowledge or evidence (another statement or set of statements) (Kyburg & Smokler, 1980). Crucially, it argues that for a hypothesis and body of evidence, there is a single probability value that the hypothesis can take, given the evidence (Carnap, 1950/1962; Hintikka, 1965; Keynes, 1921). For example, Carnap (1950/1962) built an axiomatic system for probability theory based on this logical interpretation, and always hoped there could be a set of axioms that would rule out all but one acceptable probability function given a body of evidence. But he never found a set of intuitively acceptable axioms that led to this result (Kyburg & Smokler, 1980).

The subjectivist interpretation

The subjectivist conception probability is similar to the logical one, but differs from the latter in holding that there is no single probability function that is rationally acceptable, given some body of evidence. In the subjectivist view, probabilities represent degrees of belief. A hypothesis can be assigned any probability between 0 and 1, given a body of evidence, depending on the inclination of the person whose degrees of belief the probability represents. But this does not mean that there are no rational constraints on people's degrees of belief. The subjectivist theory of probability is a logical theory in the sense that only certain combinations of degrees of belief in related statements are admissible (Kyburg & Smokler, 1980). In line with this, the founders of subjective probability theory described it as the *logic of partial belief*

(Ramsey, 1926/1990), and as the *logic of uncertainty* (de Finetti, 1972), and contrasted it with classical logic as the logic of full belief or of certainty.

The constraints on the relations between people's degrees of belief in related statements are given by *coherence* (de Finetti, 1937/1980; Ramsey, 1926/1990), which is a generalisation of logical consistency to cover degrees of belief. The probabilities of two statements are coherent if and only if they respect the axioms of probability theory.⁴ For example, if we believe it's 80% likely to rain today, then to be coherent, we would also have to be willing to believe it's 20% likely not to rain today, otherwise the probabilities used to represent our degrees of belief would not sum to 1. The logical conception of probability also includes the constraint of coherence, but in the subjectivist interpretation coherence it is the only normative constraint (Kyburg & Smokler, 1980).

The relation between the three notions of probability can be illustrated with an urn example. Imagine there is an urn with a number of balls inside, some of which have a blue dot and others not. Imagine we draw a ball at random. What is the probability that this ball will have a blue dot? On the frequentist interpretation, we would first have to draw a series of balls before we could make an estimate based on the relative frequency of balls with blue dots in our sample. Questions about the probabilities of single events do not make sense unless they can be related to similar events (or chains of events) for which frequency information is available. On the logical interpretation, we could argue as follows. Given that the only evidence we have is that some of the balls have a blue dot and others do not, it is rational to apply the principle of indifference and say, as our best guess, that the probability of drawing a ball with a blue dot is .5. On the subjectivist interpretation, we could use the same principle and also respond .5. But our neighbour could object, pointing out that the balls with no blue dot may have a green, red, or orange dot instead. Applying the principle of indifference to this new partition of the parameter space, it would be reasonable to respond that the probability of drawing a ball with a blue dot is .25. Our neighbour could add that the original partition is not more or less correct than the new partition, as both depend on the subjective question of how the domain of the problem is interpreted. However, on a subjectivist interpretation our neighbour and ourselves would nonetheless converge in our probability judgments as we sampled more and more balls from the urn, as long as we both conformed to probability theory and so made coherent judgments (Howson & Urbach, 2006).

The present thesis follows the subjectivist interpretation of probability, studying people's degrees of belief and the extent to which these are subject to the constraints of coherence. But the subject matter and results of the thesis do not presuppose a subjectivist interpretation of

⁴ There are different axiomatic systems for probability theory even within the subjectivist view (Adams, 1998, Appendix 1; De Finetti, 1937/1980, pp. 60-61), but unless otherwise specified in this thesis, the axioms can be taken as those of Kolmogorov: (1) non-negativity: $P(p) \geq 0$; (2) normalisation: $P(\text{tautologies}) = 1$; (3) finite additivity: $P(p \text{ or } q) = P(p) + P(q)$ for all p and q that are logically incompatible, i.e. that are disjunct (Hájek, 2012).

probability in order to be meaningful. Within a frequentist or logical perspective, people can still have degrees of belief, and the relation between degrees of belief and coherence constraints may be of interest independently of the relation between degrees of belief and probabilities.

HOW CAN WE MEASURE DEGREES OF BELIEF, AND WHY WOULD WE WANT THEM TO BE COHERENT?

Ramsey (1926/1990) pointed out that to explore the subjectivist view of probability theory, we need to be able to assign probabilities to beliefs, and so we need a method with which to measure degrees of belief. One option he entertained is to imagine that there is a feeling attached to each belief, e.g. a feeling of conviction, such that the stronger our conviction, the stronger this feeling will be. But this would be inconvenient for the task of measuring beliefs because it is difficult to attach numbers to feelings, and it also seemed false to him because we usually don't feel strongly for things in which our conviction is so firm that we take them for granted (e.g. that the earth is not flat). The alternative he proposed was to examine a causal connection between beliefs and actions, assuming that the stronger we are convinced of something, the more willing we will be to act on it.

Measuring beliefs by measuring actions

Ramsey (1926/1990) argued that the established way of measuring a person's beliefs would be to propose a bet and see what the lowest odds are that the person would accept. He considered this method sound, but not sufficiently general, and necessarily inexact. His reason was that larger and larger amounts of money could have diminishing marginal utility, as could very small amounts, and that a person may be especially eager or reluctant to bet (see also Elqayam, 2016, for further problems with taking the betting analogy more literally).

To construct a more general and precise form of measuring degrees of belief, Ramsey (1926/1990) proposed a betting-like scenario in which people are not asked how much they would be prepared to bet, nor assumed to accept a *fair* bet with a zero net gain. Instead, people are simply asked which of two options they would prefer. He started off by assuming, as a useful approximation, that people's actions are determined by their desires and opinions. With desires he meant things that we want in general, which could be something for our own pleasure or for the pleasure of someone else, or anything else. He called the things people want "goods", and assumed that people act in such a way as to make most likely the realisation of these goods, given their beliefs (see Adams, 1998, Chapter 9, for a similar assumption).

Ramsey (1926/1990) proposed beginning with some statement that is itself neutral with respect to people's desires, e.g. "The coin came up tails" describing the outcome of a coin flip with no bet attached to it. The idea is that two situations differing only in whether a neutral statement is true or false will be equally desirable or undesirable.

A neutral statement like the above can be used to build an anchor on a scale for degrees of belief. A person's degree of belief in a neutral statement p is said to be .5 when the person has no preference between the options (1) A if p is true, B if p is false, and (2) B if p is true, A if p is false, and the person just has a preference between A and B. Here p could be the outcome of a coin flip, A could be e.g. a cup of coffee, and B could be no coffee. This would be equivalent to being indifferent between a bet on (1) and a bet on (2) for the same stakes.

With such an anchor for degrees of belief in hand, one can then proceed to measure differences in preferences between options. Given a neutral proposition p with degree of belief .5, the difference between options A and B on the one hand, and options C and D on the other, is said to be equal when a person has no preference between the options (1) A if p is true, D if p is false, and (2) B if p is true, C if p is false. And if the difference between A and B is equal to that between C and D, then the value of A minus B will equal the value of C minus D.

To measure a person's degree of belief in a statement p , one can then proceed as follows. If the person is indifferent between the options (1) A for certain and (2) B if p is true, and C if p is false, then the person's degree of belief in p can be said to equal the ratio of the difference between A and C to the difference between B and C. The conditional probability $P(q|p)$ is measured in a similar way. Suppose a person is indifferent between the options (1) A if p is true, B if p is false, and (2) C if p and q are both true, D if p is true but q is false, B if p is false. Then the person's conditional probability of q given p is the ratio of the difference between A and D to the difference between C and D.

In general, once an anchor for a degree of belief of .5 is established, one can develop a measure of preferences and a measure of degrees of belief at the same time. If a person is indifferent between two options that are known to have the same value, then the probabilities of the two options must differ; and if the person is indifferent between two options that are known to have the same probability, then the values of the options must differ. Using the above procedure, Ramsey (1926/1990) arrives at a ratio scale for degrees of belief, and at an interval scale for preferences.

Further, Ramsey (1926/1990) derives the axioms of probability theory from his system, proving that they must be true for any coherent set of degrees of belief. He thus states (p. 78):

"We find, therefore, that a precise account of the nature of partial belief reveals that the laws of probability are laws of consistency, an extension to partial beliefs of formal logic, the logic of consistency. They do not depend for their meaning on any

degree of belief in a proposition being uniquely determined as the rational one; they merely distinguish those sets of beliefs which obey them as consistent ones."

Dutch book arguments

Ramsey (1926/1990) observed that if people's degrees of beliefs are not consistent with probability theory, i.e. are not coherent, then their preferences would be such that a *Dutch book* could be made against them. A Dutch book is a set of bets on related statements that will necessarily lead to a net loss to one side whatever the outcome of the bets. For example, suppose we think it is 80% likely to rain today, but also 80% likely not to rain today, thereby violating probability theory. Then we should have no preference between (1) a bet on rain, and (2) a bet on not-rain, each for the price of .8 giving us 1 if we win, and 0 otherwise. Because the price of .8 corresponds to our beliefs, the two bets are fair taken individually, in the sense of being associated with a zero expected gain or loss. But if we accept them together, we will pay 1.6 and receive 1, and so lose .6 no matter what happens.

De Finetti (1937/1980) also showed that the axioms of probability theory can be derived from the assumption of coherent degrees of belief, and that if people's beliefs are incoherent, a Dutch book can be made against them. His theory of how to measure degrees of belief is perhaps less intuitive than Ramsey's because it starts out with a Dutch book scenario concerning preferences that is then related to degrees of belief in a later step, whereas Ramsey develops a way of measuring preferences and beliefs in terms of one another. But de Finetti's Dutch book argument is much more elaborated than that of Ramsey. Whereas Ramsey mentioned that if people's degrees of belief are incoherent, a Dutch book could be made against them, de Finetti proved this formally in what came to be known as the *Dutch book theorem*. The Dutch book theorem was later complemented by a *Converse Dutch book theorem* (Khemeni, 1955; Lehman, 1955), which showed that if people's degrees of belief follow the axioms of probability theory, then no Dutch book can be made against them.

Together the two Dutch book theorems make a strong case for why coherent degrees of belief are something worth trying to have, not just for the internal value of preserving a generalised notion of logical consistency (Howson & Urbach, 2006), but also for the applied purpose of being able to choose actions that increase our chances of achieving our goals. This applied aspect has sometimes been criticised e.g. for presupposing money to have linear value, or that people will find fair bets with a zero net gain to either side acceptable (Elqayam, 2016; Vineberg, 2016). But as Ramsey (1926/1990) argued, the Dutch book scenario can be taken as an illustration of a more general relation between coherent degrees of belief and goals. Ramsey said that in a sense we are making bets all the time. For example, when we go to the train station we are betting that the train will come, and if our degree of belief in this were not

high enough, we would "decline the bet" and stay at home. Stalnaker (1970, p. 67) made a similar argument: "If you find gambling games a narrow and unsuitable basis on which to build the interpretation of a belief function, consider a 'bet' as any action in the face of uncertainty, and the 'odds' as the ratio of the value of what you risk by taking the action to the value of what you hope to gain, should the uncertain event turn out in your favor."

COHERENCE AND P-VALIDITY: DEDUCTION FROM UNCERTAIN PREMISES

Coherence can be said to be a deductive concept because it puts constraints on the probability values that are logically *possible* for a statement, given the probability of a related statement or set of statements. In contrast, inductive relations rather suggest one or more probability values as the most plausible or useful among those which are possible. The constraints of coherence depend on the formal logical relations between statements, but also on their probabilities, which may change depending on the content and contexts of the statements at hand. Inductive relations, for example of conceptual similarity or causal strength, seem to depend in a more fundamental way on content and context.

In an inference, the coherence constraints that the premise (or premises) places on the probability of the conclusion can vary in strength, depending on the logical form of the argument and on the premise probabilities. The premises of some inferences pose no constraints at all on the probability of the conclusion, and so are called *probabilistically uninformative*. Typical examples – assuming a Stalnaker/Lewis or probabilistic interpretation of conditionals – are transitivity, *if p then q, if q then r, therefore if p then r*, contraposition, *if p then q, therefore if not-q then not-p*, and the paradoxes of the material conditional (see Edgington, 1995, p. 286, for a discussion). For example, the conclusion of "I am in the lab, therefore if I am not in the lab but in the cafe, then I am drinking a coffee" can take any value in a probabilistic account, depending on how likely I would be to drink a coffee if I were to go to the cafe. There are other inferences whose conclusion probability is constrained to a certain interval, given the form of the inference and the premise probabilities. Examples of these are inferences describing set-subset relations, like: I am in the cafe, therefore I am in the cafe and I am drinking a coffee. Here $P(\text{conclusion}) \leq P(\text{premise})$ by coherence. There are still other inferences where the conclusion probability is fixed to a point value, given their logical form and the premise probabilities. For example, "It will rain today, therefore it will not rain today" can only be coherent if the probability of the conclusion is the complement of the probability of the premise, so that if $P(\text{rain}) = .8$, $P(\text{not-rain})$ is fixed to $.2$.

It was mentioned above that coherence is a generalisation of logical consistency to cover uncertain degrees of belief. Next to logical consistency, a second central deductive concept is logical validity. An inference is *classically valid* if and only if there is no consistent truth value

assignment to premises and conclusion in which the premises are true but the conclusion false (Adams, 1998). Thus an inference is valid when it preserves truth, or certainty, from premises to conclusion. The definition of validity categorises inferences into two sets: those that are certainty preserving, or deductive, and those that are not certainty preserving, or inductive.

A one-premise inference is *probabilistically valid*, or *p-valid* for short, if and only if there are no coherent probability assignments to premise and conclusion in which the probability of the conclusion is lower than the probability of the premise. Hence p-validity is probability preserving just as classical validity is certainty preserving. To generalise the definition of p-validity to any number of premises, it is useful to draw on a specific concept of uncertainty. Let the uncertainty, in this sense, of a statement equal one minus its probability, $U(p) = 1 - P(p)$. Then an inference is p-valid if and only if there are no coherent probability assignments to premises and conclusion in which the uncertainty of the conclusion is larger than the sum of the uncertainties of the premises (Adams, 1998). Classical validity can be seen as a special case of p-validity in which the premises are assumed to be certain.

One can see from the above definitions that the computation of p-validity presupposes coherence, just as the computation of classical validity presupposes consistency. The general definition of p-validity draws on the concept of uncertainty, which is defined in term of probabilities, and a probability is defined as a function that respects the axioms of probability theory. This makes it impossible, by definition, for the conclusion of an inference to violate coherence but respect p-validity.

Researchers focussing on coherence have sometimes criticised p-validity on the basis that there are situations in which a conclusion violates coherence but appears to conform to p-validity (Baratgin & Politzer, 2016). One can take this view only if one leaves out the first part of the definition of p-validity, keeping only the *uncertainty sum rule*, according to which the uncertainty of the conclusion should not exceed the sum of the uncertainties of the premises. The part left out indicates that the uncertainty sum rule is applied to probabilities, which by definition must respect coherence, and so must be coherently assigned to premises and conclusion. As an example, consider the MP inference:

If it rains today, the road will be muddy
 It rains today
 Therefore, the road will be muddy.

Let us assume we think $P(\text{if rain then muddy}) = .8$, and $P(\text{rain}) = .6$. Then the minimum coherent probability we can assign to the conclusion is $P(\text{if rain then muddy})P(\text{rain}) = .8 \cdot .6 = .48$. The maximum coherent probability we can assign to the conclusion is $P(\text{if rain then muddy})P(\text{rain}) + (1 - P(\text{rain})) = .48 + .4 = .88$. Thus, for these premise probabilities, the coherence interval for the conclusion of MP is $[.48, .88]$. Now, if we view p-validity as

nothing more than the uncertainty sum rule, without taking into account that it is a rule about probability functions, which must, by definition, respect coherence, then we could argue as follows. To conform to p-validity, it would be enough to assign to the conclusion any probability whose uncertainty is not greater than the sum of the uncertainties of the premises. Since the sum of the uncertainties of the premises is $.2 + .4 = .6$, and an uncertainty of $.6$ corresponds to a probability of $.4$, this means we could assign to the conclusion any probability between $.4$ and 1 . The problem then is that probabilities from $.4$ to $.47$ are incoherent, as are probabilities from $.89$ to 1 . However, this problem does not arise when we take into account that p-validity presupposes coherence, just like classical validity presupposes consistency.

Table 3.1. Probability preservation properties of inferences, based on Adams (1996).

Name	Description
1 Probabilistically uninformative	Any conclusion probability is coherent.
2 Probabilistically informative	The conclusion is constrained to any point or interval narrower than the unit interval.
3 Certainty preserving (classical validity)	It is incoherent for the premises to be certain but the conclusion uncertain.
4 High probability preserving (p-validity)	It is incoherent for the uncertainty of the conclusion to be greater than the sum of the uncertainties of the premises.
5 Positive probability preserving	It is incoherent for the probabilities of the premises to be positive but the conclusion zero.
6 Minimum probability preserving	It is incoherent for the probability of the conclusion to be lower than the lowest premise probability.

Similarly to classical validity, p-validity classifies inferences into two sets: inferences in which coherence intervals preserve probability and that are therefore deductive, and inferences in which coherence intervals do not preserve probability and that are therefore inductive. The role of p-validity of classifying inferences on the basis of features of their coherence intervals is similar to the role of probabilistic informativeness, and both could be placed on a scale describing the degree to which inferences are bound by deductive constraints. Adams (1996) described further features of probability preservation among valid inferences, which could be placed on the same scale. These features are shown in Table 3.1, numbered in increasing degree of strictness. The stricter probability preservation properties imply the less strict ones, but not the other way around. Inferences that are minimum probability preserving are also positive and high probability preserving as well as certainty preserving, inferences that are

high probability preserving are also certainty preserving, etc. Features 1 and 2 are not in fact probability preserving, but they represent less strict ways in which the premises constrain, or in the case of 1 do not constrain, the probability of the conclusion. The six features will be returned to in Experiment 10 below.

Coherence alone does not allow us to distinguish between deductive and inductive inferences. It tells us only what the coherence interval is in each case, but not why the coherence intervals of some inferences preserve probability, while those of others do not. This means that although p-validity presupposes coherence, the two have complementary, non-overlapping roles.

Overall, the generalisation of binary consistency to coherence, and of binary validity to p-validity, makes it possible to model not only reasoning from uncertain premises in general, but also *deductive* reasoning from uncertain premises in particular (Ramsey (1926/1990, p. 82; Stevenson & Over, 1995).

PROBABILITY CONDITIONALS

As mentioned earlier, the question of which inferences are deductive and which inductive depends on the meaning of the statements in them, and most importantly on the meaning of any conditionals in them. The conditional advocated in the probabilistic approach is based on the Equation, $P(\text{if } p \text{ then } q) = P(q|p)$, and has been called the *probability conditional* (Adams, 1998), *suppositional conditional* (Edgington, 2014), and *conditional event* (de Finetti, 1937/1980). It is sometimes symbolised as $q|p$ (Pfeifer & Kleiter, 2009). The term "probability conditional" is used here simply to refer to a conditional that satisfies the Equation. In particular, a probability conditional $q|p$ does not have the same semantic meaning as *if p then the probability of q is high* (a mistake made by Goodwin, 2014; see Over & Cruz, 2018).

The valid and invalid inferences for both the probability conditional and the Stalnaker (1968) conditional exactly coincide for conditionals of the form *if p then q* where neither *p* nor *q* contain a conditional. For example, transitivity, contraposition, the paradoxes of the material conditional, and other inferences like or-to-if (*p or q, therefore if not-p then q*) are all valid in classical logic for the material conditional, but invalid in systems like that of Stalnaker (1968) and like that of Adams (1998) for Stalnaker-type conditionals and the probability conditional. These systems differ of course in that degrees of belief can be modelled only in the probabilistic approach, but they also differ in how they treat conditionals embedded in other conditionals, and conditionals building compounds with negations, conjunctions and disjunctions (Edgington, 1995, p. 273). Such more complex conditional statements are straightforward to represent in Stalnaker's (1968) system, in which conditionals are full propositions, but they are the focus of research, with different proposals being currently

developed, in the probabilistic approach (Bradley, 2012; Gilio, Over, Pfeifer, & Sanfilippo, 2016; Stalnaker & Jeffrey, 1994).

Next to the Equation and its psychological implementation in the Ramsey test, it is useful to characterise conditionals through a truth table, which in the case of the probability conditional shows the probabilities that the conditional can take as a function of each possible combination of the truth or falsity of its components. De Finetti (1936/1995) held that the indicative conditional *if p then q* is true when both *p* and *q* are true, is false when *p* is true but *q* is false, and is *void*, or not specified, when *p* is false. The de Finetti table is shown in Table 3.2, where 1 = true and 0 = false.

Table 3.2. The de Finetti table for the probability conditional.

	$q p$
p, q	1
$p, \text{not-}q$	0
$\text{not-}p, q$	Void
$\text{not-}p, \text{not-}q$	Void

This table was initially described in the psychology of reasoning as the *defective truth table*, on the grounds that it does not correspond to the predictions for the material conditional of classical logic (Evans, 1972; see Over & Baratgin, 2017, for a discussion). De Finetti (1936/1995) appears to have thought of "true" and "false" as referring to objective truth values in his table, but he escapes the triviality results because the "void" entry in the false-antecedent cases is not an objective truth value. Probability conditionals can also be regarded as "true" , or "false", in *pragmatic*, or *pleonastic*, uses of these terms. For example, a person might assert a conditional as "true" simply to express a high degree of belief in it, "beyond reasonable doubt", or because they endorse it as a convention or a matter of personal taste, as in "if you make gazpacho, you use olive oil", while someone who does not like olive oil might reject this as "false". A further use of "true" for conditionals is to refer to logical truths, like *if p then p*. This conditional is true in a logical sense, independently of whether *p* refers to a physical fact that could be observed (Over & Cruz, 2018).

De Finetti drew an analogy between indicative conditionals and conditional bets. Suppose we bet that "if it is raining, the match is cancelled". When it is raining and the match is cancelled, we win the bet, and when it is raining and the match goes ahead, we lose the bet. But if it does not rain, the bet gets called off, with no one winning or losing. Going outside and seeing that it is a clear sunny day, we do not assert the sentence "if it is raining, the match is cancelled". Ramsey (1929/1990, p. 155) agreed that in this case the indicative conditional is "void", and added that it ceases to mean anything to us, "except as a question about what

follows from certain laws or hypotheses" (see Cruz & Oberauer, 2014, on general conditionals). In this case, we might instead take an interest in the counterfactual "if it had rained, the match would have been cancelled".

Jeffrey (1991) asked which probability could be assigned to the false-antecedent cases of a probability conditional, so that the expected value of the conditional as a whole follows the Equation. He proved that this had to be the conditional probability itself. The resulting *Jeffrey table* is shown in Table 3.3. This table, like de Finetti's, sidesteps the triviality results because the conditional is not objectively true or false in false-antecedent cases. The 1 and 0 in the table can be interpreted as objective or subjective, but it makes no difference to the overall probability of the conditional.

Table 3.3. The Jeffrey table
for the probability conditional.

	$q p$
$p \ \& \ q$	1
$p \ \& \ \text{not-}q$	0
$\text{not-}p \ \& \ q$	$P(q p)$
$\text{not-}p \ \& \ \text{not-}q$	$P(q p)$

Specifying $P(q|p)$ as the value of a probability conditional in false-antecedent cases makes it possible to model counterfactuals. Counterfactuals were impossible to represent in the de Finetti table, which was limited to the description of indicatives as void, and it was also difficult in Adams' (1998) specification of the probability conditional, in which the false-antecedent cases of the truth table were either left empty, or were given the value 1 per convention. Adams needed such a convention because (following Kolmogorov strictly) he defined conditional probabilities as derived from the ratio formula, $P(q|p) = P(p \ \& \ q)|P(p)$, with the consequence that conditionals are undefined when their antecedents are known to be false. Such a convention is not needed when conditional probabilities are instead considered primitive in the formal system, as in de Finetti's approach. For de Finetti all probabilities are in fact conditional probabilities, either conditional on a specific statement, or more generally conditional on background knowledge k . The Equation would then more precisely be written $P(\text{if } p \text{ then } q) = P(q|p, k)$. One can then turn the ratio formula around and define conjunctions in terms of conditional probabilities, $P(p \ \& \ q) = p(p)P(q|p)$. And outside the formal system, one can use the psychological procedure of the Ramsey test to fix the conditional probabilities.

It can be argued that when people are asked to evaluate the probability of a conditional for false-antecedent cases, they may attempt to solve the task by switching from an evaluation of the indicative to an evaluation of the corresponding counterfactual. Drawing an analogy once more between conditionals and bets, in the false-antecedent cases in which the bet is

called off, we receive our money back, which for a fair bet corresponds to our degree of belief in the conditional (Coletti & Scozzafava, 2002).

The Jeffrey table prevents misunderstandings that come from focusing on the de Finetti table alone, without considering it together with the Equation and the Ramsey test (Byrne & Johnson-Laird, 2009; Douven, 2015a; Gilio et al., 2016). For example, the conditional *if p then p* has the intuitive probability of 1 in the false-antecedent cases of the Jeffrey table, and not the undifferentiated void value it would have in the false-antecedent cases of the de Finetti table (Baratgin, Over, & Politzer, 2013). Nonetheless, the Jeffrey table can be considered a specification and extension the de Finetti table, and several contemporary researchers in the de Finetti tradition adopt the Jeffrey table as a basis of their approach (Coletti & Scozzafava, 2002; Over & Baratgin, 2017; Pfeifer & Kleiter, 2009).

CONDITIONALS AND VALIDITY

It was mentioned above that inferences like contraposition and the paradoxes are valid for the material conditional but invalid for the probability conditional. This has led to the suggestion that the inferences that are valid in probability logic are a subset of the inferences that are valid in classical logic (Pfeifer & Kleiter, 2009, Figure 1). However, this characterisation can be misleading because it raises the question of whether the dividing line between deductive and inductive inferences in classical logic is different from that in probability logic. When $P(q|p)$ is specified through the Ramsey test and so is considered primitive within the logical system, and is thus defined also in cases in which the antecedent has probability zero, the dividing line between deductive and inductive inferences is in fact the same. Among the inferences that can be represented in classical logic (i.e. those that do not contain probability conditionals), all inferences that are certainty preserving in classical logic are also certainty and high probability preserving in probability logic, and all inferences that are certainty preserving in probability logic are also certainty preserving in classical logic. But with the probability conditional, probability logic can represent a set of further inferences that could not even be expressed in classical logic. Among these further inferences, some are certainty and high probability preserving, and others not. The inferences that are certainty and high probability preserving in probability logic are a proper subset of the corresponding (but not identical) inferences that would be certainty preserving in classical logic, if the probability conditionals they contain were replaced with material conditionals. For example, the paradox of the material conditional: "not-rain, therefore not-rain or muddy", where "not-rain or muddy" represents the material conditional, is certainty preserving in classical logic, and it is also certainty and high probability preserving in probability logic, as a case of the inference of *or-introduction*. In contrast, the corresponding inference "not-rain, therefore muddy|rain" where "muddy|rain"

represents the probability conditional, is neither certainty nor high probability preserving – but it cannot even be formulated in classical logic.

The misleading suggestion that the dividing line between deductive and inductive inferences may be different in classical and probability logic comes from applying the same term "conditional", to two concepts that are not equivalent, the material conditional *not-p or q* and the probability conditional $q|p$. Referring to both as conditionals is of course difficult to avoid, especially given that there is no universal agreement among researchers on the meaning(s) of natural language conditionals. But it is nonetheless necessary to treat the two concepts as formally distinct, in which case classical and probability logic coincide in their dividing line, and so in their definition of deduction.

A further point on the relation between conditionals and validity is that when the conditional probability in the Equation is specified through the Ramsey test and so is considered primitive in the formal system, certainty preservation (classical validity) and high probability preservation (p-validity) coincide (Adams, 1996, 1998; Gilio, 2002). They only diverge in Adams' (1998) system because of the default assumption made in it that conditionals with zero antecedents have probability 1. This assumption has as a consequence that, although inferences like the paradox of the material conditional, *not-p, therefore if p then q*, are not high probability preserving, they are nonetheless certainty preserving because in the special case in which the premise has probability 1, $P(\text{not-}p) = 1$, the antecedent p of the conditional in the conclusion has probability zero, rendering the probability of the conditional as a whole equal to 1 by default. When the conditional probability is instead primitive, the conclusion in this inference can have any probability when the probability of the premise is 1. The value that the conclusion probability will take in any particular instance will be determined by a Ramsey test, and correspond to the conditional probability itself, as given in the Jeffrey table. Because of the equivalence in the set of inferences that are classically valid and that are p-valid for conditionals based on the Jeffrey table, the terms "validity" and "p-validity" will be used interchangeably in the thesis unless otherwise specified.

UNCERTAIN REASONING BEYOND DEDUCTION: DYNAMIC REASONING

A further aspect of the probabilistic approach is its application to reasoning beyond deduction, even on a probabilistic interpretation of the term. This is the field of dynamic reasoning, reasoning over time, in contrast to the focus on static reasoning that had characterised the earlier, binary approach (Ali, Chater, & Oaksford, 2011; Baratgin & Politzer, 2010; Douven, 2012; Hadjichristidis et al., 2014; Hartmann & Rafiee-Rad, 2014; Oaksford & Chater, 2013). An important question in the probabilistic approach is how people change their degrees of belief over time after learning new information (Oaksford & Chater, 2013; Over, 2016).

Imagine you want to assess whether a certain substance is acid. At time 1, you have some suspicion that it might be acid, and you buy a piece of litmus paper to test this. At time 1 you may also have an idea of the likelihood that the litmus paper turns red, given that it is an acid, and the likelihood that it turns red, given that it is not an acid. You can then infer, still at time 1, a degree of belief that the substance is an acid, given that the litmus paper turns red, using *Bayes theorem*:

$$P_1(Acid|Red) = \frac{P_1(Red|Acid)P_1(Acid)}{P_1(Red|Acid)P_1(Acid) + P_1(Red|notAcid)P_1(notAcid)}.$$

(you would come to the same result by using a Ramsey test, in which you hypothetically assume that the litmus paper turned red and assessed how likely the substance is to be an acid under this assumption). Bayes theorem describes relationships between static beliefs: beliefs that are held at a single point in time.

When you then get home after buying the litmus paper and actually perform the litmus test at time 2, you can update your degree of belief that the substance is an acid using *Bayesian conditionalisation*, by making your degree of belief that the substance is acid at time 2 equal your degree of belief at time 1 that it is acid, given that the litmus paper turned red: $P_2(Acid) = P_1(Acid|Red)$ (Hadjichristidis et al., 2014; Oaksford & Chater, 2013; Over, 2016).

Bayes theorem only depends on the axioms of probability theory (Chater & Oaksford, 2012), but because Bayesian conditionalisation describes relationships between beliefs over time, it makes two important further assumptions, which are to some extent idealisations (Adams, 1998). The first is that the information you learn at time 2 becomes certain for you. Often the information we learn will be less than certain. For example, it could be that you are making the test under dim light. If the information you learn is not certain, then you can update your belief that the substance is acid using Jeffrey conditionalisation: $P_2(Acid) = P_1(Acid|Red)P_2(Red) + P_1(Acid|not-red)P_2(not-red)$. One can see that this equality reduces to the total probability theorem if its components are all represented at the same time point, i.e. if $P_1(Acid|Red) = P_2(Acid|Red)$, and $P_1(Acid|not-red) = P_2(Acid|not-red)$.

This points to the second additional assumption, made both in Bayesian and in Jeffrey conditionalisation. It is that when you learn the new information at time 2, you learn only it, and your other beliefs remain *invariant*. However, it could be that when you see whether the litmus paper turned red or not, you also see that it is beyond its "use by" date. In such a case not only your probability that the substance is an acid, given that the paper turns red will change from time 1 to time 2, but also your probability that the paper turns red, given that the substance is an acid, and your probability that it turns red given that the substance is not an acid. When invariance does not hold, we are in a situation of non-monotonic reasoning. Alternative criteria for updating our beliefs in this case may be given by imaging (Baratgin &

Politzer, 2010; Lewis, 1976; Over, 2017; Zhao & Osherson, 2014), or by minimising the Kullback-Leibler divergence (Hartmann & Rafie-Raad, 2014). But there are no hard and fast criteria in this case, and it is an interesting question in itself when it is justified to assume that invariance holds, and when not (Oaksford & Chater, 2013; Zhao & Osherson, 2014).

EMPIRICAL EVIDENCE FOR THE PROBABILISTIC APPROACH

There is vast evidence that people tend to reason from uncertain premises, even when asked to assume the premises to be certain for the sake of argument. The role of uncertainty, or degrees of belief, goes beyond the field of reasoning and has gained prominence in accounts of a range of cognitive processes from perception and language processing (Chater & Manning, 2006; Kersten & Yuille, 2003) to learning and argumentation (Bramley, Dayan, Griffiths, & Lagnado, 2017; Griffiths & Tenenbaum, 2009; Hahn & Oaksford, 2007), in what has been called the *probabilistic turn* in cognitive science (Chater, Oaksford, Hahn, & Heit, 2010; Oaksford & Chater, 2007, Ch. 4).

Evidence for reasoning from uncertain premises

In the area of deductive reasoning, most studies have focussed on three specific groups of inferences or tasks: conditional syllogisms, categorical syllogisms, and the Wason selection task (Evans & Over, 2004; Oaksford & Chater, 2001, 2007). Conditional syllogisms are four inferences with a conditional or *major* premise, a categorical or *minor* premise that asserts or denies one of the components of the conditional premise, and a conclusion that asserts or denies the other component of the conditional. They are: *Modus ponens* (MP, *if p then q, p, therefore q*), *Modus tollens* (MT, *if p then q, not-q, therefore not-p*), *Affirmation of the consequent* (AC, *if p then q, q, therefore p*) and *Denial of the antecedent* (DA, *if p then q, not-p, therefore not-q*). MP and MT are deductively valid, whereas AC and DA are not.

Conditional syllogisms are represented in sentential logic, so that the *p* and *q* referred to stand for whole sentences. In contrast, categorical syllogisms are represented in quantified predicate logic. In predicate logic, the sentences are broken down into smaller units, similar to objects and descriptions of those objects. For example, instead of representing "The road is muddy" as *p*, one could represent it as Mr, where M is a predicate, or description, of the object r. Such more fine grained constructions are quantified if one says that the property M applies not just to the individual object r, but to one or more members of a set of objects, which we could call x. For example, x could be the set of roads in our district. One can then say things like "for all x, Mx", which would mean "all roads in our district are muddy". Categorical

sylogisms are inferences in predicate logic that have two premises and a conclusion, each containing two terms, with the premises sharing one term and the conclusion referring to the other two terms, i.e. to the terms that occurred only in one premise. Both premises and conclusion contain quantifiers like "all", "some", "none" and "some not". For example: "All roads that are wet are muddy. All roads that are muddy are impassable. Therefore, all roads that are wet are impassable."

In the original version of the Wason selection task (Wason, 1968), participants are shown four cards, each with a number on one side and a letter on the other, e.g. D, 3, B, 7. Participants are then asked which cards need to be turned over to determine whether the conditional rule "if there is a D on one side, then there is a 3 on the other" is true or false of the four cards.

The task posed to participants in the Wason selection task is conceptually ambiguous, and can be interpreted in a deductive or inductive way (Johnson-Laird & Byrne, 2002; Manktelow & Over, 1991; Oaksford & Chater, 1996, 2003; c.f. Crupi, Tentori, & Lombardi, 2009; Klayman & Ha, 1987). It is also not studied in this thesis. Further, all the inferences investigated in the thesis can be expressed in sentential logic. The following exposition therefore focusses on reasoning with conditional syllogisms.

Early findings on conditional syllogisms, using binary paradigm instructions to assume the premises are true, to then judge whether the conclusion also has to be true, found that people do not simply accept the valid MP and MT, and reject the invalid AC and DA. The acceptance rates for MP tend to be at ceiling, and people tend to accept valid inferences more often than invalid ones. But people also accept the two invalid inferences with non-negligible frequency, and tend to accept the *forward* inferences (MP and DA) more often than the *backward* inferences (MT and AC) (Evans & Over, 2004; Klauer, Beller, & Hütter, 2010; Oberauer, 2006). However, the latter difference can be reduced or reversed when accounting for negation effects (Evans, Clibbens, & Rood, 1995), i.e. for findings suggesting that it is more difficult for people to process inferences that contain negations (Evans & Handley, 1999; Oaksford, Chater, & Larkin, 2000).

A further series of findings that deviate from the expectations of classical logic more directly suggests people treat the premises of conditional syllogisms as uncertain, even when instructed to assume them to be true with certainty. It was found that people accept the valid inferences of MP and MT less often when given an additional premise that casts doubt on the conditional premise, e.g. by pointing to a precondition that must be met for the conditional premise to hold (Byrne, 1989) or suggesting that the conditional is more or less likely to hold (Stevenson & Over, 1995; see also Over, 1993), and also if the conditional premise itself is varied, explicitly or through the choice of its content, in a way that varies $P(q|p)$ (George, 1997; Liu, Lo, & Wu, 1996). Further studies showed that while the endorsement rates of MP and MT were most affected by $P(q|p)$, those of AC and DA were most affected by $P(p|q)$

(Cummins, Lubart, Alksnis, & Rist, 1991; Thompson, 1994) which makes sense because AC and DA would be equivalent to MP and MT, respectively, if the conditional premise *if p then q* were swapped with its converse, *if q then p*.

The above findings were called *suppression effects*, suggesting that people suppress the inferences when their degree of belief in the premises decreases. In line with the above results, both the number of different counterexamples to the conclusion (instances of *p & not-q* for MP and MT, and instances of *not-p & q* for AC and DA), and the overall frequency of counterexamples, was found to play a role in creating suppression effects. This makes sense from a probabilistic point of view, because both types of information are correlated and both undermine the probability of the conclusion. From a binary paradigm point of view, a single counterexample, or any probability lower than 1 in the premises, would be expected to render the conclusion false with certainty. The question of whether one type of undermining evidence comes first and is an indicator of the other, seems to depend on whether participants receive classical binary or probabilistic instructions (De Neys, Schaeken, & d'Ydewalle, 2003; Geiger & Oberauer, 2007; Markovits, Lortie-Forgues, & Brunet, 2010).

From a probabilistic point of view, it is rash to call these findings *suppression effects*. This is because the inference itself is only suppressed if people's conclusion probability is lower than the lower bound of the coherence interval for the inference – or in a generalisation of the term *suppression*, then also if it is higher than the upper bound of the interval (Over & Cruz, 2018). Consider a MP inference in which $P(\text{if } p \text{ then } q) = P(p) = .1$. Then by coherence, the probability of the conclusion must lie in the interval [.01, .11], and any conclusion probability above or below this interval would *suppress* the logical constraints of MP.

A similar finding to suppression effects is that of *belief bias*, the tendency to accept a conclusion more often when it is believable given the premises, regardless of whether it follows logically from the truth of the premises (Evans, Handley, & Bacon, 2009). In the probabilistic approach this could again be a reasonable response in cases in which the unbelievable premises yield a low coherence interval for the conclusion. Such a response would of course be incorrect when binary paradigm instructions are given to assume the premises, however implausible or arbitrary. But participants might not follow such instructions and yet they might still be engaged in a deductive reasoning process, following the logic of partial belief or of uncertainty. Further research will be necessary to establish whether people suppress, not only the conclusion of an inference, but the inference itself, and respond on the basis of belief bias in a way that violates the logical constraints of coherence.

Evidence for the probability conditional

The psychological hypothesis that people's degrees of belief in conditionals will follow the Equation is called the *conditional probability hypothesis* (Evans, Handley, & Over, 2003). There is strong converging evidence for this hypothesis. When in *classical truth table tasks* people are given the four cases of the truth table, and are asked for each case whether it renders the conditional true, false, or whether the case is irrelevant for the truth or falsity of the conditional, people's modal response pattern corresponds to the de Finetti table (Baratgin et al., 2013; Evans, 1972; Johnson-Laird & Tagart, 1969; see Evans & Over, 2004, for a review). And when in *probabilistic truth table tasks* people are asked to judge the probability of the conditional given information about the frequency of each truth table case, people's modal response closely matches the conditional probability (Evans, Handley, Neilens, & Over, 2007; Evans et al., 2003; Oberauer, Geiger, Fischer, & Weidenfeld, 2007; Oberauer & Wilhelm, 2003).

A minority response found in these studies corresponded to the conjunction of p & q , whereas the response corresponding to the material conditional was virtually absent. The minority conjunctive response was not reliable, but decreased with practice on the task (Fugard, Pfeifer, Mayerhofer, & Kleiter, 2011), with general cognitive ability (Evans et al., 2007), with children's age (Barrouillet & Gauffroy, 2015) and when real world materials were used (Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Singmann, Klauer, & Over, 2014) – in each case being replaced not with the material conditional, but with the probability conditional response.

A further minority response pattern sometimes found is that of the probability biconditional, *if p then q & if q then p* , which has $P((p \ \& \ q)|(p \ \text{or} \ q))$ as its probability (Fugard, Pfeifer, et al., 2011; Oaksford & Chater, 2013, 2017; Over, 2017). This pattern might be relevant to the study of some general conditionals, e.g. "if an animal is a bird, then it can fly". Another possibility for general conditionals is that, not only must $P(q|p)$ be high, but $P(q|not-p)$ must be low. $P(q|p) - P(q|not-p)$ is termed delta- p (Allan, 1980), and it is logically related to another probability biconditional, *if p then q & if not- p then not- q* , since $P(q|not-p) = 1 - P(not-q|not-p)$. Appendix A gives the Jeffrey tables for these two biconditionals. These interpretations appear to occur more often when conditionals describe causal relations, but the evidence for their presence is not reliable (Dieussaert, Schaeken, & d'Ydewalle, 2002; Oaksford & Chater, 2013; Oberauer, Geiger, et al., 2007; Oberauer, Weidenfeld, et al., 2007; Singmann et al., 2014). It may be that when they occur in causal conditionals this is because the conditionals are interpreted as general conditionals about law-like relations, and that in these cases people's degree of belief in them is similar to a judgment of causal power (Cheng, 1997; c.f. Fernbach, Darlow, & Sloman, 2010). But further research on the distinction between

singular and general conditionals is necessary to address this question (Cruz & Oberauer, 2014).

Further evidence for the Equation draws on the betting scenario introduced by de Finetti (1937/1980) and Ramsey (1926/1990) as a way of measuring subjective probability. Studies comparing people's probability judgments with people's judgments about bets on conditionals found a close correspondence between the two (Baratgin et al., 2013; Oberauer & Wilhelm, 2003; Politzer, Over, & Baratgin, 2010).

Current work on the meaning of conditionals within the probabilistic approach is investigating to what extent the scope of the conditional probability hypothesis extends to general conditionals (Cruz & Oberauer, 2014; Oaksford & Chater, 2017; Over, 2017); to counterfactuals (Over et al., 2007; Pfeifer & Stöckle-Schobel, 2015, September; Sloman & Lagnado, 2005); to compounds of conditionals (Bradley, 2012; Gilio et al., 2016; Gilio & Sanfilippo, 2014; Stalnaker & Jeffrey, 1994; Van Wijnbergen-Huitink, Elqayam, & Over, 2015); and to missing-link conditionals, in which the antecedent and consequent are independent (Cruz, Over, Oaksford, & Baratgin, 2016; Douven, 2015b; Oberauer, Weidenfeld, et al., 2007; Skovgaard-Olsen, Singmann, & Klauer, 2016). However, this thesis will focus on singular indicative conditionals without missing links.

The following section provides a brief overview of current psychological theories that draw on alternative interpretations of conditionals, or that integrate probabilistic elements into a broader dual-process framework, and discuss how they relate to the probabilistic approach.

CHAPTER 4. ALTERNATIVES TO THE PROBABILISTIC APPROACH IN PSYCHOLOGY

Contents

4.1 Mental model theory (MMT)

4.1.1 Conditionals in MMT

4.1.2 Reasoning with conditional syllogisms in MMT

4.1.3 Mental models and probabilities

4.1.4 New MMT

4.2 Dual-component theories

4.2.1 "Logic" vs. "belief" in dual-component theories

4.2.2 Breaking the association of "logic" to type 2 and "belief" to type 1 processes

4.2.3 Breaking the "logic" vs. "belief" dichotomy itself

4.3 Research question

This section begins with a description of mental model theory (MMT) (Johnson-Laird & Byrne, 1991, 2002; Johnson-Laird et al., 2015), one of the most influential psychological theories of reasoning for the past decades. It is mostly considered an alternative to the probabilistic approach, though there are features of the two approaches that have been integrated with one another (Evans, 2006; Geiger & Oberauer, 2010; Girotto & Johnson-Laird, 2004; Manktelow & Over, 1991; Oaksford & Chater, 2010b). MMT is referred to at points in the interpretation of results in the thesis, but the thesis as a whole arguably has a stronger bearing on dual-process theories (De Neys, 2012; Evans & Stanovich, 2013; Markovits, Brunet, Thompson, & Brisson, 2013; Oaksford & Chater, 2011; Sloman, 1996; Trippas, Thompson, & Handley, 2017; Verschueren, Schaeken, & d'Ydewalle, 2005), or dual-component theories more generally (Klauer et al., 2010; Singmann et al., 2014). This section goes on to describe some of the proposals of dual-component theories, before it concludes with an outline of the hypotheses.

MENTAL MODEL THEORY

MMT has been applied to reasoning in a range of deductive (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird & Byrne, 2002; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Johnson-Laird & Savary, 1999) and inductive (Johnson-Laird, 1994) tasks, including ones about spatial relationships (Ragni & Knauff, 2013) and about causality (Khemlani, Barbey, & Johnson-Laird, 2014). It has sometimes been criticised because the MMT specifications for these tasks can appear independent, task specific accounts (Baratgin et al., 2015). But the general assumptions of the theory may be summarised as follows.

People reason by making and manipulating isomorphic representations of aspects of the reasoning domain, called models. In the theory's earlier account of deductive reasoning with sentence particles like *if*, *and*, *or*, and *not* (Byrne & Johnson-Laird, 2009; Johnson-Laird & Byrne, 2002), each model refers to a logical possibility for a statement, corresponding to the truth table cases that make the statement true in classical logic. For example, a conjunction is true in one possibility, that in which both *p* and *q* are true, so it is represented with one model:

p *q*

The inclusive disjunction is true in three possibilities, so it is represented with three models:

p

q

p *q*

Note that the above *mental models* do not explicitly represent negations. But they can be *fleshed out* into *fully explicit models*, which also include information about which statements

are false at each possibility in which the statement as a whole is true. For the disjunction this yields:

p	not-q
not-p	q
p	q

The logical possibilities in which the statement as a whole is false (in the case of the disjunction the *not-p, not-q* possibility) are generally not included in any models. The assumption that people tend to represent what is true in a model, rather than what is false, and that they represent the models that render a statement true but not those that render a statement false, is called *the principle of truth* in MMT. Logically, the fully explicit models are equivalent to the *disjunctive normal form* for the statement. For example, this form for the disjunction *p or q* is $(p \ \& \ \text{not-}q)$ or $(\text{not-}p \ \& \ q)$ or $(p \ \& \ q)$.

Conditionals in MMT

The initial *mental model* for conditionals is represented as a conjunction, followed by an *implicit model* represented as an ellipsis:

p q
...

The implicit model symbolises that there are further possibilities in which the statement could be true, that are not yet fleshed out. The *fully explicit model* for the conditional is equivalent to the cases of the truth table in which the material conditional is true:

p	q
not-p	q
not-p	not-q

Consequently these models are the same as the models for *not-p or q*.

The above models represent what the theory calls the *basic* meanings of conjunctions, disjunctions and conditionals. These are the meanings people are thought to give to the statements when their contents are "abstract", or arbitrary. For example, a basic conditional about figures drawn on a black board could be "If there is a circle then there is a square". The theory posits that these meanings can change through *semantic and pragmatic modulation*, i.e. through features of the content or context (Byrne & Johnson-Laird, 2010; Johnson-Laird & Byrne, 2002). For example, "If it rained then it poured" would be represented only by two models:

it rained	it poured
it did not-rain	it did not pour

with the model "it did not rain, it poured" ruled out by the content, which indicates that pouring is a form of raining. Modulation is assumed to rule out models and also annotate models with extra information, e.g. about temporal or causal order, but it is assumed not to lead to the addition of models (Byrne & Johnson-Laird, 2010).

It is argued that although the meanings of basic conditionals, conjunctions and disjunctions are truth functional, and have the same truth conditions as in classical logic, the meanings of these statement types after semantic and pragmatic modulation is not truth functional, so that overall, conditionals and other statement types are not truth functional in the theory (Johnson-Laird & Byrne, 2002). There are two difficulties with this assumption. The first is that MMT categorically endorsed the paradoxes of the material conditional as logically valid inferences (Johnson-Laird & Byrne, 1991, 2002), but the validity of these inferences (plus the inference from the truth of p and falsity of q to the falsity of *if p then q*) entails truth functionality (Evans, Over, & Handley, 2005). The second is that when and how semantic and pragmatic modulation affects the meaning of conditionals, and other statement types, is not clearly specified. Byrne and Johnson-Laird (2010, p. 66; see also Johnson-Laird & Byrne, 2002, p. 674) are explicit on this point: "One consequence of modulation is that conditionals have an indefinite number of meanings – ten sets of possibilities and a variety of relations between the antecedent if-clause and the consequent then-clause." The result is that when modulation is taken into account, any experimental finding can be explained, and none can be predicted. There has consequently been little research testing the theory's account of non-basic conditionals other than by proponents of the theory (for a recent example see Orenes & Johnson-Laird, 2012).

Proponents of MMT (Barrouillet, Gauffroy, & Lecas, 2008; Girotto & Johnson-Laird, 2010) have argued that one of the interpretations of conditionals posited by the theory corresponds to the de Finetti table, and so explains people's responses in truth table tasks. They held that the de Finetti response pattern for conditionals, *if p then q* , may be based on the initial mental model in which only the p & q possibility is represented explicitly, and the other two possibilities remain implicit (denoted by the ellipsis). Being implicit, these two possibilities are considered irrelevant, while at the same time the p & *not- q* possibility is somehow acknowledged as the only case that makes the conditional false.

However, this position seems inconsistent with other parts of MMT. If the implicit possibilities for the conditional are treated as irrelevant, then implicit possibilities should also be treated as irrelevant in conjunctions and disjunctions, which would change their meanings (for example, the *not- p , q* model should then be judged irrelevant, and not false, for conjunctions). And if the p , *not- q* model for the conditional is fleshed out alongside the p , q model, then this would violate the principle of truth. There is no mechanism within MMT for distinguishing between false and irrelevant cases (Oberauer & Oaksford, 2008), and it is

generally difficult to find a representation for a statement that is not a full proposition using the propositional tools for representing mental models.

Reasoning with conditional syllogisms in MMT

The earlier version of MMT (Johnson-Laird & Byrne, 2002) assumed that when people draw inferences with conditionals in conditional syllogisms, they first build mental models for the two premises, and then integrate these models. The integration is done by adding the models of one premise to the models of the other, and removing any models that are inconsistent. Next people take a few (typically one or two) of the logical possibilities in which the combined models are true, and assess whether the conclusion is true in these possibilities. If it is, the process stops, and they take the inference to be deductively valid. If it is not, the process also stops, and they take the inference to be invalid. But if the conclusion does not feature in the premises, and people have enough time and resources, then they look at further models of the premises, possibly fleshing them out into fully explicit models, to see whether they can find a counterexample to their initial conclusion. A counterexample is a logical possibility in which the premises are true but the conclusion false. Markovits & Barrouillet (2002) proposed that counterexamples are based not just on abstract logical possibilities, but are retrieved from long term memory, with ease of retrieval varying with the specific content of the conditional (c.f. Cummins, 1995).

The above procedure implies that, if people flesh out all the models of the premises, and modulation is not an issue, then they will respond exactly in the way predicted by the rules of classical logic. But the theory posits that people make errors because they fail to take into account relevant logical possibilities. For instance, they might not flesh out mental models into fully explicit models, or not flesh out the negations within a model. This is assumed to occur mainly because of working memory limitations, and it implies that the more models are required to fully flesh out the premises of an inference, the more difficult the inference will be.

Work in developmental psychology (e.g. Barrouillet & Gauffroy, 2015; Barrouillet & Lecas, 1999; Markovits & Barrouillet, 2002) suggested a developmental trend in the interpretation of conditionals, going from a conjunctive interpretation in young children, over a biconditional interpretation, to a conditional interpretation in adults – with the conditional interpretation being in accordance with the material conditional under binary paradigm instructions, and in accordance with the probability conditional under probabilistic instructions. MMT proposed that this trend comes about because, when people flesh out the initial conjunctive model of the conditional, they first add the *not-p, not-q* model that yields a biconditional interpretation, and only afterwards add the *not-p, q* model that results in the conditional interpretation. There is no rationale within the theory for this specific ordering

(e.g. why is the *not-p, not-q* model added before and not after the *not-p, q* model?), but the interpretation of the developmental trend, in the successive addition of models, corresponds to a situation in which increasing *numbers* of models are considered, in line with the predictions of the theory.

The above ordering can also be used to explain the pattern of responses in the conditional inference task. A conjunctive interpretation of the conditional in which the ellipsis is forgotten would lead to the acceptance of MP and AC. The addition of the *not-p, not-q* model would lead to the acceptance of all four inferences, and the subsequent addition of the *not-p, q* model would result in participants accepting MP and MT and rejecting AC and DA. This account of conditional syllogisms has found some support in the literature, but mostly when combined with an additional assumption about conditionals that is not part of MMT, called *directionality* (Barrouillet, Grosset, & Lecas, 2000; Evans et al., 2005; Oberauer, 2006). This is the assumption that people tend to process a conditional in direction from antecedent to consequent, and that the outcome of processing the conditional can differ if it is instead carried out from consequent to antecedent. Directionality is hard to explain within MMT when reasoning with basic conditionals, but it is a feature that follows from the Ramsey test (see also Verschueren et al., 2005).

Mental models and probabilities

MMT was extended early on to cover situations in which the premises are not certain but only probable to a higher or lower degree. It proposed that by default models are equiprobable, so that the probability of a statement corresponds to the proportion of models that make the statement true. But it was also held that people are able to assign distinct probabilities to models. In simple cases, they could do this by creating repetitions of models of the same kind, and letting the proportion of models of each kind stand for their probability. In more complex cases, tags with numerical probabilities could be added to the models (Geiger & Oberauer, 2010; Girotto & Johnson-Laird, 2004; Johnson-Laird et al., 1999). Conditional probabilities can be represented by using the *subset principle*, in which a person focuses on the subset of models in which the denominator is true, and assesses the proportion of models in this subset in which the numerator is also true (Johnson-Laird et al., 1999). This procedure seems functionally equivalent to a Ramsey test. Later on Khemlani et al. (2015) also described a way of representing subjective probabilities, using a scale set a priori to have eight subdivisions, on which positive and negative evidence for a statement could be averaged (see also Juslin, Nilsson, & Winman, 2009, on the averaging hypothesis). But MMT never proposed a rational procedure for transferring premise probabilities to conclusion probabilities for relations more complex than that of a set to a subset.

New MMT

Johnson-Laird and colleagues (Johnson-Laird et al., 2015; see also Johnson-Laird & Ragni, 2017, July; Khemlani, Hinterecker, & Johnson-Laird, 2017) have recently proposed a radical revision of MMT, in which the paradoxes of the material conditional, but also other classically valid inferences, notably *or-introduction*, p , *therefore* p or q , are declared invalid in the theory. The revision is based on a change in the meanings of *and*, *or*, and *if*. They are still represented by the same models, but whereas before a statement was true when one of the possibilities used to represent it was actually the case, now a statement is held to be equivalent to the conjunction of those possibilities. For example, as pointed out above, the disjunction p or q was at first fully represented, in effect, as its disjunctive normal form, $(p \ \& \ \text{not-}q)$ or $(\text{not-}p \ \& \ q)$ or $(p \ \& \ q)$. But it is now taken to be equivalent to the conjunction of possibilities: *possibly* $(p \ \& \ \text{not-}q)$ & *possibly* $(\text{not-}p \ \& \ q)$ & *possibly* $(p \ \& \ q)$. One consequence of this change is that the paradox of the material conditional *not- p , therefore if p then q* , is argued to be invalid, because the model used to represent *not- p* does not guarantee that all the models used to represent the material conditional are possible. *Or-introduction* is declared invalid for the same reason: the model for p does not establish that the three models for the disjunction are possible.

The revision arguably has the advantage that the highly counterintuitive paradoxes are no longer considered valid. But it apparently entails inconsistent and counterintuitive consequences that make it not worth pursuing (for overviews see Baratgin et al., 2015; Cruz, Over, & Oaksford, 2017; Oaksford, Over, & Cruz, 2018). For example, if *or-introduction* is invalid, then so is *and-elimination*, $p \ \& \ q$, *therefore* p , because the two inferences can be derived from one another by De Morgan's rules. But *and-elimination* remains valid in the revised MMT, as the models of the premise do establish that the model for the conclusion is possible. The claim that *or-introduction* is invalid may be based on experiments with binary paradigm instructions, showing that it is then endorsed less often than other valid inferences (e.g. Orenes & Johnson-Laird, 2012). But under probabilistic instructions, asking people directly about their degrees of belief, *or-introduction* is accepted to the same degree as *and-elimination*, and to a much higher degree than the paradoxes of the material conditional (Cruz et al., 2017). Further, if p , *therefore* p or q is invalid, then the premise p should be consistent with the negation of the conclusion, *not- $(p$ or $q)$* , which is equivalent by De Morgan's rules to *not- p & not- q* . But MMT continues to consider two statements consistent when they have at least one model in common, and p does not share any model with *not- p & not- q* (Baratgin et al., 2015).

It is also not clear what Johnson-Laird et al. (2015) mean when they state that *or-introduction* is invalid "in MMT". An inference is valid or invalid within a specific logical system. This might be classical logic, probability logic, or a new logic defined for MMT. But

the theory would have to specify either a known logical system on which it bases itself when making assertions about validity, or specify its own provably consistent system, or substitute for "validity" a non-logical term to characterise inferences, e.g. "endorsed often by most people". In the current formulation of the theory, it is unclear whether the proposals made about validity are at the computational or the algorithmic level (c.f. Oaksford et al., 2018).

MMT is referred to at points in the interpretation of the results of the thesis, but there are perhaps two main reasons why the theory has no strong bearing on the work of the thesis as a whole. The first is that the mental model account of reasoning with probabilities has not yet been specified for relations more complex than that between a set and a subset. This means that it is not clear, for example, how the probabilities of two-premise inferences, like the conditional syllogisms, could be transferred in a principled way to the conclusion. The second is that the current revision of MMT raises many questions that make it difficult to derive predictions from it. However, the revision is still under development, and removing the apparent inconsistencies may be a part of that process, with the implications for reasoning from uncertain premises becoming clearer over time.

DUAL-COMPONENT THEORIES

This thesis uses the term "dual-component theories" to refer summarily to dual-process theories of thinking (De Neys, 2012; Evans & Stanovich, 2013; Oaksford & Chater, 2011; Trippas et al., 2017; Verschueren et al., 2005), and to theories postulating two factors in reasoning that do not map clearly onto the dual-process distinction, for example two sources of information used for reasoning (Klauer et al., 2010; Singmann & Klauer, 2011; Singmann et al., 2014), or two strategies in reasoning (Markovits, Brisson, & Chantal, 2015).

Arguably the most central assumption of dual-process theories is that reasoning, or cognition more generally, takes place at two levels of awareness, reflexivity, or overt attention. In line with this, type 1 processes have often been described as intuitive and type 2 processes as reflective (Evans, 2010).

The distinction between intuitive and reflective processes is drawn at the algorithmic level, i.e. it is concerned with people's awareness of the representations and processes that they use in reasoning, rather than with the nature of the task they set out to solve when engaged in reasoning. As such, it is orthogonal to the probabilistic approach, whose proposals have until now been focussed at the computational level (Oaksford & Chater, 2012). One can see this in the example of Evans' dual process theory, which, as he points out, started off in the binary paradigm and then evolved into the probabilistic one (Evans, 2012, p. 21). However, once in the probabilistic approach, dual-process theories can of course contribute greatly to it, and provide much needed algorithmic level proposals to complement the computational level ones

(Bonnefon, 2013; Elqayam & Over, 2013). There seems to be ample evidence for a distinction between intuitive and reflective processes, not least in the form of the evidence that cognitive processes differ in their dependence on working memory and attention resources (Evans, 2008; Evans & Stanovich, 2013).

A further distinction often made in dual-process theories (Evans, 2003; 2012, p. 16; Trippas et al., 2017; Verschueren et al., 2005), though also studied its own right (Klauer et al., 2010; Markovits et al., 2015; Singmann & Klauer, 2011), is between "deductive", "logical", or "form-based" reasoning on the one hand, and "inductive", "probabilistic", or "belief based" reasoning on the other. This is a distinction at the computational level, as it concerns the norms defining correct inferences, or definitions of the task, that people use in reasoning. In what follows, the distinction between intuitive and reflective reasoning will be called *levels of reasoning*, and the distinction between "logical" and "probabilistic" reasoning will be called *forms of reasoning*. A question raised in this thesis concerns the consequences for the above forms of reasoning of generalising deduction to degrees of belief, so that logical reasoning can itself be probabilistic.

"Logic" vs. "belief" in dual-component theories

Evans (2012, p. 22) argued that when dual process theories were first developed in the binary paradigm, they "originally attempted to find a theoretical basis for the observed competition between logical and non-logical processing in paradigms such as syllogistic belief bias" (see also Evans et al., 2009, p. 78). As mentioned above, belief bias refers to the finding that, under binary paradigm instructions, people's evaluations of the conclusions of inferences seem to be influenced not only by whether the inferences are deductively valid or not, but also by whether the conclusions, given the premises, or the conclusions taken on their own, are believable or unbelievable. By allocating the "logical" responses, in the sense of classical binary logic, to type 2 processing, and the "belief based" responses to type 1 processing, the perceived discrepancy between the two could arguably be accounted for.

Evans' earlier dual-process account (Evans, 2003, 2006) proposed that type 1 processes are heuristic and draw on evolved behavioural tendencies, mechanisms of associative learning and world knowledge (c.f. Sloman, 1996; 2014). Type 1 processes were said to be responsible for building a semantic representation of the premises, and if possible cue an initial conclusion. This initial conclusion was supposed to be based on a *single* mental representation *relevant* for the task. Type 2 processes were proposed to be rule-based and have the exclusive ability to compute deductive relations. Also, only type 2 processes were said to draw on working memory resources, allowing people to reason about abstract and novel materials. The role of type 2 processes was to potentially intervene to revise the initial response, depending

on whether or not the initial response *satisfied*, or worked well enough, for the purposes of the task. This form of interaction between type 1 and type 2 processes is called *default-interventionism*. In later versions of the theory, the only defining difference between type 1 and type 2 processes was proposed to be the involvement of working memory, with the concomitant ability to solve novel reasoning tasks (Evans, 2017; Evans & Stanovich, 2013).

The dual-source model of Klauer, Singmann, and colleagues (Klauer et al., 2010; Singmann & Klauer, 2011; Singmann et al., 2014) also draws a contrast between a binary "logic based" and a probabilistic "belief based" component in reasoning, but it is concerned with forms and not levels of reasoning, and so is not a dual-process theory. Drawing on earlier work by Liu et al. (1996), it proposed that people use two sources of evidence, or information, when evaluating conditional syllogisms: *logical form* and *prior knowledge* (Klauer et al., 2010). The form component was measured as the observed pattern of acceptance of the four conditional syllogisms under binary paradigm instructions. The knowledge component was measured by drawing on accounts of conditional syllogisms within the probabilistic approach (Oaksford et al., 2000; Pfeifer & Kleiter, 2009).

The conceptual distinction between two sources of information in reasoning, one based on logical form and the other on premise content, seems perfectly consistent with the probabilistic approach, where logical form can be related to coherence constraints, and premise content to premise probabilities. Such a mapping between accounts is more difficult in later formulations of the theory, in which the form component is identified with deductive reasoning, and the knowledge component with inductive reasoning (Singmann & Klauer, 2011). The mapping is also more difficult when considering how the components are measured. The probabilistic theories used to model the knowledge component are defined at the computational level, and are logical theories in a wide sense. These theories set coherence constraints on deductive reasoning. The context and content of the premises can be used to refine these constraints further up to a point value for the conclusion (Oaksford et al., 2000). From this perspective, it is not clear what is being measured by the form component that is not already entailed by the knowledge component (c.f. Singmann et al., 2014).

The dual-strategy account of Markovits et al. (2015), originally formulated within a dual-process framework (Verschuere et al., 2005) but then generalised, again contrasts a "logic based" component with a "belief based" component. The "logic based" component is characterised as a binary deductive reasoning strategy based on the search for specific counterexamples to a conclusion, as proposed in MMT. The "belief based" component is characterised as an inductive, statistical strategy, based on estimating the overall frequency of counterexamples to a conclusion. The theory proposes that people use the statistical strategy by default, but that the choice between strategies also depends on factors such as task instructions, cognitive capacity, and metacognitive control (Markovits et al., 2013).

An argument emphasised in this thesis is that the dichotomy between "logic" and belief" in dual-component theories is unnecessary once logic is generalised to cover degrees of belief. Given that this generalisation exists, it is worth exploring its implications for the components proposed to be involved in reasoning.

Breaking the association of "logic" to type 2 and "belief" to type 1 processes

The allocation of binary "logic" to type 2 processes, and probabilistic "belief" to type 1 processes, was broken up by findings from Handley, Trippas and colleagues (Handley, Newstead, & Trippas, 2011; Trippas et al., 2017). Building on earlier work by Rips (2001), they showed that in belief bias tasks, not only does the believability of an inference interfere with judgments of validity, but the validity of an inference also interferes with judgments of believability. For simple logical inferences, validity was even found to interfere more with judgments of believability than believability interfered with judgments of validity (Handley et al., 2011; Trippas et al., 2017). This is contrary to the default-interventionist assumption that the computation of logical structure is limited to reflective type 2 processes. If the response cued by type 1 processes is never logical, then it is hard to explain why when the instructions ask for a believability and not a validity judgment, validity still has an influence on people's responses. The authors proposed that instead, intuitive and reflective processes operate in parallel. Both intuitive and reflective processes are said to cue responses based on both logical structure and beliefs (Handley & Trippas, 2015). The resulting redundancy is considered advantageous because it is said to lead to more robust responses in non-conflict cases. In cases of conflict between responses based on logical structure and responses based on beliefs, the prevailing response type is said to be a function of the complexity of the computations required for each. For example, it is held that simple logical relations like MP interfere more with belief based responses than do more complex relations like MT (Trippas et al., 2017).

In the above proposal it does not seem clear what role is being fulfilled by dividing "logic based" and "belief based" computations into heuristic and reflective ones. But more generally, the theory shares with the dual-component theories described earlier the assumption of a dichotomy between binary logic on the one side, and probabilistic beliefs on the other. This seems entirely unnecessary from the perspective of the probabilistic approach given that logic can also be probabilistic.

Breaking the "logic" vs. "belief" dichotomy itself

The binary paradigm dichotomy between "logic" and "belief" is left aside in the dual-process theory of De Neys (De Neys, 2012, 2014). De Neys takes into account that both classical logic and probability theory are general, computational level frameworks for reasoning, and refers to them summarily as "logic". This use of the term was also advocated by Ramsey (1926/1990) and de Finetti (1972), who referred to probability theory as the "logic of partial belief" and the "logic of uncertainty", respectively. Logical computations in this generalised sense are contrasted with heuristics based on semantic and stereotypical associations (De Neys, 2012), making it possible to extend the study of dual-processes to tasks in which what is considered a bias refers to a departure from the principles of probability theory, as in studies of the conjunction fallacy, base rate neglect, or the bat and ball problem (De Neys, 2014).

In a series of experiments, De Neys and colleagues found evidence that people are sensitive to the principles of both binary logic and probability theory in an apparently intuitive, implicit way, even when this is not reflected in their overt responses. For example, in experimental conditions in which a logic based response conflicts with a heuristic based response and people's response sides with the heuristic, people nonetheless show longer response times (De Neys & Glumicic, 2008), lower confidence in their responses (De Neys, Cromheeke, & Osman, 2011), a higher skin conductance response (De Neys, Moyens, & Vansteenwegen, 2010), and higher activation in the anterior cingulate cortex, involved in error monitoring (De Neys et al., 2008). These effects did not depend on reasoning accuracy (De Neys et al., 2010), and were not affected by working memory load (Franssens & De Neys, 2009), suggesting that the detection of a conflict between logical and heuristic responses is implicit and effortless. The effortlessness of conflict detection in turn suggests that interindividual differences in response choice appear to be based, not on the ability to detect a conflict between a logical and a heuristic response, but on the ability to inhibit the heuristic response (De Neys et al., 2008; see also Dube, Rotello, & Heit, 2010, on interpreting belief bias as a response bias). These findings provide further evidence against a default-interventionist account of the interaction between type 1 and type 2 processes, at least when associating "beliefs" with the former and "logic" with the later.

From a theoretical perspective, De Neys (2012) argued that, in default-interventionist accounts, it is hard to explain how reflective type 2 processes can determine whether the initial, intuitive response is good enough for the task at hand, or whether there should be intervention to revise it, without already being engaged, at least to some extent, from the beginning. But for parallel processing accounts arguing that reflective processes are indeed engaged from the beginning, it may be hard to explain why this parallel engagement is not a waste of resources. He proposed a third option, called a logical intuition model, in which intuitive type 1 processes can cue both logical responses – in a wider sense that includes

responses conforming to probability theory – and heuristic responses based on semantic and stereotypical associations. Reflective type 2 processes come into play when there is a conflict between an intuitive logical and an intuitive heuristic response. The role of type 2 processes then seems to be to try to resolve this conflict to keep the individual functioning, even if the result will not invariably favour the logic based response (De Neys, 2012; c.f. Oaksford & Chater, 2011; Thompson & Johnson, 2014; Trippas et al., 2017).

The above proposal is in line with the dual-process theory proposed by Oaksford, Chater and colleagues (Oaksford & Chater, 2011, 2012; Oaksford & Hall, 2016). Their account is fully probabilistic, and posits that the main difference between intuitive and reflective processing is that the first relies on long-term memory and the second on working memory. The role of working memory in this *single function, dual-process* account is similar to that in the present account of Evans (Evans & Stanovitch, 2013): it makes it possible to selectively restrict the information considered when making an inference to fulfil specific task goals, and engage in hypothetical and counterfactual thinking (Oaksford & Chater, 2012).

Deviations of people's responses from the conclusions sanctioned by probability theory are said to arise for instance when people draw on information from long term memory to determine their degree of belief in the premises. They might thereby include information that is not part of the premises, or that misses something that is part of the premises, with corresponding differences in their degree of belief in the conclusion (Oaksford & Chater, 2012). Divergences from probabilistic principles are also said to occur when, due to working memory limitations, people forget relevant information in the process of drawing inferences, or distort their degrees of belief in premises and conclusion when attempting to verbalise them using the discrete tools of language (Oaksford & Hall, 2016). But the verbalisation of people's inferential processes brings with it the advantage that they can be recorded and shared with others, who can then help us consider alternative interpretations and solutions in rational discourse (Oaksford & Hall, 2016).

Leaving aside the binary paradigm contrast between "logic" and "belief" makes it possible for dual-component theories to be integrated with the probabilistic approach, offering new ways of interpreting findings and formulating research questions. For example, belief bias can be reinterpreted as an effect of deductive reasoning from uncertain premises, in spite of instructions to artificially assume the premises to be certain. It is then an open question whether people show belief bias in a wider sense, in which they move outside of probabilistic deduction by violating coherence constraints. Similarly, it could be assessed whether differences in responses under binary paradigm vs. probabilistic instructions can be usefully understood as concerning the difference between certainty preservation and probability preservation, rather than between deduction and induction.

RESEARCH QUESTION

Generalising deduction to inferences from degrees of belief has possible implications for dual-component theories of thinking, but before investigating these, it is necessary to examine whether this generalisation is descriptively adequate. That deduction can be generalised to probabilities in formal logic (Ramsey, 1926/1990, p. 82) does not of course imply that people will actually use deduction in a probabilistic way. The role of coherence and p-validity in people's reasoning has only started to be investigated empirically (Evans, Thompson, & Over, 2015; Pfeifer & Kleiter, 2005, 2009; Singmann et al., 2014), but this investigation is the focus of the following ten experiments.

PART 3. EXPERIMENTS

CHAPTER 5. EXPERIMENTS 1 TO 4: COHERENCE ABOVE CHANCE LEVELS

Contents

- 5.1 Methodological points relevant across experiments
 - 5.1.1 Above-chance coherence
 - 5.1.2 Linear mixed models
- 5.2 Experiment 1: Ifs and ors
 - 5.2.1 Method
 - 5.2.2 Results and discussion
 - 5.2.3 General discussion
- 5.3 Experiment 2: Ifs, ands, and the conjunction fallacy
 - 5.3.1 Overview of the conjunction fallacy
 - 5.3.2 Ifs and ands
 - 5.3.3 Method
 - 5.3.4 Results and discussion
 - 5.3.5 General discussion
- 5.4 Experiments 3 and 4: Intuition, reflection, and working memory
 - 5.4.1 Experiment 3
 - 5.4.2 Experiment 4
 - 5.4.3 General discussion

METHODOLOGICAL POINTS RELEVANT ACROSS EXPERIMENTS

Above-chance coherence

The main dependent variable in most of the experiments of the thesis was above-chance coherence, i. e. the extent to which the observed rate of coherent responses was higher than expected by chance. It was computed using a method introduced by Evans et al. (2015; see also Pfeifer & Kleiter, 2009; Politzer & Baratgin, 2016). The observed rate of coherence responses was computed by creating a binary variable taking the value 1 when a response was coherent, and the value 0 when it was incoherent. For the one-premise inferences investigated here, the response to a valid inference is coherent when the probability assigned to the conclusion is equal to or higher than the probability assigned to the premise. Conversely, for the one-premise inferences investigated, the response to an invalid inference is coherent when the probability assigned to the conclusion is equal to or lower than the probability assigned to the premise.

This binary variable for *observed coherence* was then compared to a variable representing the chance rate of a coherent response, *chance coherence*. On the assumption that a random response can fall equally likely on any point of the probability scale, the probability of complying to coherence by chance corresponds to the width of the coherence interval. This is a simplifying assumption, for probability estimates could be based on sampling procedures that lead to higher chance rates for extreme cases (Stewart, Chater, & Brown, 2006), and otherwise random responses could still show response tendencies towards the middle or the extreme of the scale. However, we considered a uniform distribution of chance rates a sufficiently accurate approximation to allow an assessment of the hypotheses at hand. On this assumption, if a person assigns for instance a probability of .6 to the premise of the valid inference of *or-introduction*, then the probability they assign to the conclusion is coherent if it falls within the interval between .6 and 1. Because the width of this interval is .4, the chance rate of conforming to coherence in this case is also .4. The values for *above-chance coherence* were obtained by subtracting the values for chance coherence from those for observed coherence for each participant and condition.

Linear mixed models

Above-chance coherence was used as the dependent variable in a series of linear mixed models. Linear mixed models (LMMs) can be considered a special case of multilevel linear models (MLMs), which are themselves generalizations of regression models (and hence also

of ANOVA's). In a MLM, the variance in the data can be partitioned into different levels of aggregation. For example, at level 1 could be the variance in individual responses of a student, at level 2 the variance in mean responses of the students in a classroom, and at level 3 the variance in mean responses of the classrooms in a school (Hox, Moerbeek, & van de Schoot, 2018; Snijders & Bosker, 2012; Tabachnick & Fidell, 2007).

Being able to distinguish variance in the data at different levels of aggregation is important when the variance within a unit (e. g. within individuals, or within classrooms) is smaller than the variance between units. For example, when different responses of the same person are more similar to each other than they are to the responses of a different person, or when test results are more similar for students within the same classroom than for students of different classrooms. When this is the case, the independence of errors assumption for ordinary linear regression is violated. As a result, analysing the data at the individual level without taking its hierarchical structure into account can lead to dramatical alpha error inflation rates, because the analysis is based on too many degrees of freedom that are not truly independent (Tabachnick & Fidell, 2007).

Moreover, possible differences in the pattern of the data between levels of aggregation can make it impossible to extrapolate conclusions from one level to another. An example is the ecological fallacy of drawing conclusions at the level of individuals on the basis of data patterns measured at the level of groups (Robinson, 1950; see also Te Grotenhuis, Eisinga, & Subramanian, 2011). Consider the case of the relationship between local language performance and number of years living in a location, for students in different cities of that location. Assume we have contingency tables summarising this relationship for each city. The correlation between language performance and language exposure at the individual level will be based on the data within each contingency table. In contrast, the same correlation at the city level will be based on the marginals of each table. But the marginals of a contingency table do not uniquely determine the values within it (Robinson, 1950). Therefore the correlation at the individual level can be positive, while the correlation at the city level may be negative (e. g. if people arriving more recently to the location tend to live in larger cities, where average values of language proficiency are higher) (Te Grotenhuis et al., 2011). The fact that a statistical correlation can be positive at one level of aggregation and negative at another has been referred to as Simpson's paradox, or Simpson's reversals (Malinas & Bigelow, 2016).

Discussions of the ecological fallacy, or of Simpson's paradox, sometimes revolve around the question of which level of aggregation is the most adequate to draw on for data interpretation. But when the pattern in the data is different at different levels of aggregation, this can be interesting in itself, and the different levels can be analysed and related, instead of having to choose between them. This can be done with MLM.

In a MLM, the assumption of independence of errors is not required. The degree to which the variance within units of analysis is smaller than that between units is instead included

explicitly in the model. This is done by allowing the intercepts (means) and slopes (IV-DV relationships) for the regression equation at one level (e. g. students) to vary at the next highest level of aggregation (e. g. classrooms). This variability is then modelled by treating the intercept and slope for the regression equation at one level as the DVs in a regression equation at the next highest level. For example, if one is measuring student achievement (individual level DV) as a function of student motivation (individual level IV), then one can allow mean student achievement (intercept), as well as the relation between student achievement and student motivation (slope), to vary between classrooms.

A MLM allows predictors to be included at each level of analysis. For example, student achievement could be predicted by student motivation at the individual level, by teacher emphasis on homework at the class level, and by school poverty at the school level. It is possible to assess cross-level interactions, for example, the interaction between teacher emphasis on homework and student motivation in predicting student achievement. And it is possible to assess interactions between an intercept and a slope, as opposed to only interactions between slopes. For example, the relationship between student achievement and student motivation may vary as a function of school poverty level (Tabachnick & Fidell, 2007). Overall, this means that MLMs not only allow one to avoid errors in interpretation resulting from not taking the nested structure of the data into account, but they also make it possible to test hypotheses that could not be formulated in ordinary regression or ANOVA analyses.

The present work draws on a special case of multilevel linear models (MLM): linear mixed models (LMM). These are models in which the variance in the data is partitioned into different levels of aggregation, but not all higher level units are nested within each other. For example, assume a school takes pupils from different neighbourhoods. Then the students of this school will be nested within classrooms, and within neighbourhoods, but the classrooms and the neighbourhoods will not themselves be nested – they are instead said to be crossed (Hox et al., 2018, Ch. 9; Snijders & Bosker, 2012, Ch. 13).

The modelling of crossed random effects is useful in situations like the above example, in which students are nested in classrooms but also in neighbourhoods, or when students attend more than one school during the time period assessed by a study. But an important further field of application is to experimental settings in which individual responses are nested within participants and within items, but the participants and the items are not themselves nested. Using a LMM in this case makes it possible to treat both participants and items as random variables, and therefore to increase the generalizability of the results without having to compute separate by participant and by item analyses (Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012). This is the context in which the method is used in the present thesis.

Because MLMs are more complicated to compute, and sometimes to interpret, than ordinary linear regression or ANOVA, it makes sense to ask when it is worth the effort. In earlier work on the topic, it was often recommended that researchers explicitly check whether the hierarchical structure of the data actually makes a difference in the case of the sample at hand. If for each level of aggregation, the variance within groups is not smaller than that between groups (and hence the so called intra-class correlation is low), then it was considered sufficient to analyse the data using ordinary linear regression or ANOVA. Similarly, if the inclusion of a random intercept or slope in the model did not significantly improve model fit, or its effect was not large enough to be statistically significant, then it was considered unnecessary to include it (Tabachnick & Fidell, 2007).

More recently, it has been instead recommended that researchers include in the model the maximum random effect structure justified by the design (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). This is because there are cases in which a random effect is too small to be statistically significant, but its exclusion nonetheless leads to significant biases in the pattern of results and their interpretation. The present thesis follows these more recent recommendations.

For the experimental designs used, this means that random intercepts for participants and for item scenarios were included in the model when possible. It was not possible to do so in each analysis because there were cases in which the ratio of participants to model parameters would have been too small – particularly in follow-up tests to interaction effects. When it was possible to include one random intercept but not both, the intercept for participants was given priority. It was generally not possible to include random slopes, because there were generally no, or not sufficient, repeated observations within each cell of the design. In most cases, each participant provided a single response for each design cell, and similarly, each item scenario was associated with a single design cell per participant. This allowed the inclusion of further experimental variables in the design without making an experiment prohibitively long. However, it had the drawback that any variance in slopes could not be distinguished from the error variance, and so could not be modelled separately. In the words of Barr: "If observations are not replicated (i.e., there is only a single observation per unit per cell), random slope variance cannot be distinguished from random error variance and thus random slopes need not be included" (Barr, 2013, p. 1; see also Baayen et al., 2008).

Perhaps partly because of the ability of MLMs to differentiate more precisely between sources of variation, it is not yet straightforward to compute measures of explained variance, or of effect size, in MLMs. An immediate difficulty arises when it comes to determining the sample size to use for such a computation, because sample sizes mostly differ between each level of aggregation. Further, MLMs assume that the data has been obtained by successive random sampling: sampling a number of groups from a population of potential groups, and then sampling a number of individuals from each group. When under these circumstances a

variable added to the model explains variance at level 1, it can also be expected to explain some variance at level 2. If the amount of explained variance at each level differs from that expected by the assumption of successive random sampling, this can lead to the occurrence of negative values for explained variance, which do not make sense (Snijders & Bosker, 1999). Various correction formulas have been developed for this, but it appears that all still sometimes lead to counterintuitive results (Nakagawa & Schielzeth, 2013). For this reason the present work will restrict itself to providing confidence intervals for statistical results and their graphical representation.

Two further technical decision points in the use of MLM are the estimation method and the type of covariance matrix. The present analyses used Maximum Likelihood estimation (ML), as opposed to restricted Maximum Likelihood estimation (reML). ML is considered to be more robust and to yield more precise estimates than reML, although reML is considered to yield more unbiased estimates. But the principal reason for choosing ML over reML was that only ML allows the statistical comparison of nested models using likelihood ratio tests. This is useful when an effect is only marginally significant. The effect can then not only be tested individually within a given model, but also by comparing the fit of two nested models, one with and one without the effect included. In general, marginal effects on both sides of the significance threshold were complemented with likelihood ratio tests and an inspection of confidence intervals, and were interpreted with caution.

The covariance structure used in the analyses was always unstructured, reflecting the fact that no prior assumptions were made about its form. In all analyses, predictors were centred around their grand mean to avoid issues of collinearity when computing interactions.

EXPERIMENT 1: IFS AND ORS⁵

The first studies investigating whether people's probability judgments are coherent above chance levels focussed on conditional syllogisms (Evans et al., 2015; Singmann et al., 2014). They found people to be coherent above chance levels for MP and to a lesser degree for DA, but not for MT and AC. On the other hand, studies on whether people's assessment of the probability of a (simple or complex) statement and of its negation sum up to 1 have found strong evidence for coherence (e. g. Costello & Watts, 2014, 2016; Klauer et al., 2010). The difference in these two groups of findings may be explained as an effect of task complexity: That a probability and its complement sum to 1 is a basic principle of probability theory, which can be applied to the more complex case of deriving a conclusion probability for a two-

⁵ The majority of the data from this experiment was published using an ANOVA analysis in a collaborative paper with Jean Baratgin, Mike Oaksford, and David Over (Cruz, Baratgin, Oaksford, & Over, 2015).

premise inference. If people are sensitive to the constraint of coherence but fail to comply with it when task complexity increases, then one would expect response coherence to be more reliably above chance level for simpler one-premise inferences. Unlike two-premise inferences, one-premise inferences do not require the integration of premise probabilities. This simplifies their coherence intervals, which, for the inferences studied in this experiment, go either from the probability of the premise to 1 (in the case of p-valid inferences) or from the probability of the premise to 0 (in the case of p-invalid inferences).

This experiment assessed the coherence of responses to the eight one-premise inferences in Table 5.1.

Table 5.1. The inferences used in Experiment 1.

	Name	Form
1	OrIf1	<i>p or q, therefore if not-p then q</i>
2	OfIf2	<i>not-p or q, therefore if p then q</i>
3	IfOr1	<i>if p then q, therefore not-p or q</i>
4	IfOr2	<i>if not-p then q, therefore p or q</i>
5	OrI1	<i>p, therefore p or q</i>
6	OrI2	<i>not-p, therefore not-p or q</i>
7	OrI3	<i>q, therefore p or q</i>
8	OrI4	<i>q, therefore not-p or q</i>

Inferences 1 and 2 are logically equivalent, as are inferences 3 and 4, as well as inferences 5 to 8, respectively. However, the equivalence of inferences 5 and 6 on the one hand, with inferences 7 and 8 on the other, presupposes the fact that the disjunction *p or q* is commutative, so that *p or q = q or p*. Apart from this, the sets of equivalent inferences differ only in the position of the negation they contain. The positions of the negation used are those for which the largest negation effects have been reported in the literature (Espino & Byrne, 2013; Oberauer et al., 2011). The above variants of equivalent inferences were introduced in order to control for negation and order effects.

Inferences 1 and 2 are logically equivalent *or-to-if* inferences, going from a disjunction to a conditional. When the natural language conditional in these inferences is interpreted as the material conditional, the premise and the conclusion are equivalent, and then judgments are only coherent when the premise and the conclusion are assigned the same probability, $P(\text{if } p \text{ then } q) = P(\text{not-}p \text{ or } q)$. When the conditional is interpreted as the probability conditional, $P(\text{if } p \text{ then } q) = P(q|p)$, judgments are coherent when they conform to the relation $P(q|p) \leq P(\text{not-}p \text{ or } q)$; that is, when the probability of the conclusion is equal to or lower than the probability of the premise.

Inferences 3 and 4 are logically equivalent *if-to-or* inferences, going from a conditional to a disjunction. Under the assumption that *if p then q* is equivalent to the material conditional, premise and conclusion again have the same probability $P(\text{if } p \text{ then } q) = P(\text{not-}p \text{ or } q)$, and any other probability judgment is incoherent. Under the assumption that *if p then q* is the probability conditional, $P(\text{if } p \text{ then } q) = P(q|p)$, it follows from the axioms of probability theory that $P(q|p) \leq P(\text{not-}p \text{ or } q)$, and probability judgments must conform to this relation to be coherent. Hence the relation that must hold for the inferences to be coherent is the same for inferences 3 & 4 as for inferences 1 & 2. The difference is that in the first two the conditional is the conclusion, and in the second two the conditional is the premise.

Overall, this implies that if one interprets the conditional as the probability conditional, the *if-to-or* inference is coherent when the probability of the conclusion is equal to or higher than the probability of the premise, and the *or-to-if* inference is coherent when the probability of the conclusion is equal to or lower than the probability of the premise.

The difference in the conditions for coherence of the two inferences is reflected in the fact that under a probability conditional interpretation, the *if-to-or* inference is p-valid, whereas the *or-to-if* inference is p-invalid and can even be a quite weak inference. Consider an instance of inference 5. We might have a high degree of confidence that our bicycle is outside our apartment in Paris where we left it. That should, if we are coherent in the *or-introduction* inference, give us a high degree of confidence that our bicycle is outside our apartment in Paris or in Antarctica. But we do not have any confidence that, if our bicycle is not outside our apartment in Paris, then it is in Antarctica. It is much more reasonable to infer that, if our bike is not there, it is somewhere else in Paris after being stolen. Gilio and Over (2012) have an analysis of when inferences 1 & 2 are, and are not, reasonable inferences to make, and Over, Evans, & Elqayam (2010) have supporting findings.

Inferences 5 and 6 are equivalent forms of the p-valid inference of *or-introduction*. As such, any probability assignment to the conclusion is coherent when it is equal to or higher than the probability of the premise. This is analogous to the treatment of the inference in classical logic, where the conclusion validly follows from the premise if and only if it is true in every instance in which the premise is true. Inferences 7 and 8 are also equivalent forms of *or-introduction*, and differ from inferences 5 and 6 only with regard to whether the statement in the premise refers to the first or the second element of the disjunctive conclusion. The results for inferences 7 and 8 were not reported in Cruz et al. (2015) due to space constraints.

Studies within the binary paradigm found that people endorse *or-introduction* less often than other logically valid inferences (Braine, Reiser, & Rumin, 1984; Orenes & Johnson-Laird, 2012; Rips, 1983). This finding has generally been explained as a pragmatic effect: people are unwilling to draw the inference because it would be misleading in a conversation to say *p or q* when one can make the more informative statement *p* (Grice, 1989; see also Bar-Hillel & Neter, 1993; Fugard, Pfeifer, & Mayerhofer, 2011; Orenes & Johnson-Laird, 2012;

Tversky & Köhler, 1994;). If we map the response "true" to "P(conclusion) = 1" and the response "false" to "P(conclusion) = 0", then the failure to endorse *or-introduction* can be said to reveal incoherence in reasoning about it. However, as argued by Cruz et al. (2015), the probabilistic approach predicts that this incoherence is likely to be reduced when people are asked directly for their degree of belief in the conclusion, given their degree of belief in the premise, as opposed to stating whether the conclusion necessarily has to be true assuming the truth of the premise. This is because asking directly about a person's degrees of belief may lessen pragmatic effects based on what is, or is not, misleading in an open conversation with another person (including an experimenter). One prediction for this experiment is thus that people's responses for *or-introduction* will be coherent above chance levels.

Because the question of whether people's responses to the *or-to-if* inference are coherent depends on how the conditional is interpreted, this experiment not only investigates response coherence in general. It also tells us something about the modal interpretation of the conditional. If people's judgments are highly incoherent for one interpretation, yet highly coherent for another, there is an argument in favour of the interpretation that renders their judgments coherent.

Thus, this experiment made an assessment of whether people's judgments for the inference forms of Table 5.1 are coherent above chance levels, and a comparison between the material conditional and the probability conditional interpretation of natural language conditionals. A third question investigated was whether people's sensitivity to coherence is something that requires an explicit focus of attention on the statements in an inference (Oberauer, 2013), with explicit *reasoning* about what the probability of a premise statement implies for the probability of a conclusion statement; or whether the establishment of coherence is something that occurs spontaneously and implicitly, even in the absence of an explicit reasoning task (Evans et al., 2015).

To assess the third question, participants were divided into an *inferences group* and a *statements group*. Participants in the inferences group were shown all the statements composing an inference (i. e. the premises and the conclusion of the inference) together on the same screen. They received the explicit instruction to reason about the implications of the probability of the premise for the probability of the conclusion. In the statements group, the statements that had served as premise and conclusion for the inferences group were presented one at a time on the screen, in random order. Participants in this group were simply asked to rate their degree of belief in each statement.

Evans et al. (2015) used a similar manipulation in a lab experiment with conditional syllogisms, and found that responses were more often coherent above chance levels in the inferences group than in the statements group. A small difference between the experimental design in Evans et al. and the present experiment was that in Evans et al., participants in the statements group were not shown all statements in random order. Instead, the statements were

first sorted into two groups: a group of conditional statements on the one side, and a group containing the antecedents and consequents of these conditionals (which in the inferences group served as the minor premises and conclusions of inferences) on the other. Participants were asked to rate the probabilities of the statements within each statement group separately.

This task grouping variable assessed whether the effect of task found in Evans et al. (2015) could be replicated in an internet experiment using simpler one-premise inferences.

Method

Participants

A total of 1140 participants from English speaking countries completed the online experiment in exchange for € 1. Among them, 566 completed one of the eight booklets of the inferences group, and 305 completed one of the two booklets of the statements group. Participants were removed from the sample if they indicated being younger than 18 or having only "basic" English language skills, if they provided the same responses across all trials, or if they had one or more trial response times of 2 seconds or less. The final sample had the following characteristics for each group. The inferences group consisted of 456 participants. Their median age was 33 (range (18-78), and they reported a variety of levels of formal educational training. The final sample for the statements group consisted of 204 participants. Their median age was 36 (range 18-72), and they also reported a variety of levels of formal educational training.

Material and design

Participants were shown a short scenario describing a person, and then presented with a series of statements about the person. In the statements group, these statements appeared one at a time on the screen, in random order for each participant. In the inferences group, the statements were presented in pairs as premises and conclusions of explicit inferences. Participants in the statements group were asked to judge how confident they were in each statement, by typing in a percentage between 0% ("no confidence at all") and 100% ("complete confidence"). Participants in the inferences group were asked to judge how confident they were in the premise of the argument, and then how confident they were in the conclusion, given the premise. Participants in the inferences group used the same percentage scale as those in the statements group to provide their answers.

Two scenarios were varied between participants: The Linda scenario (Tversky & Kahneman, 1983) with the standard description of Linda, and a scenario describing a person conforming to a stereotype quite unlike that of Linda. The frame below shows a sample trial in

the statements group and in the inferences group, using the Linda scenario. The complete list of materials can be found in Appendix B.

In the inferences group, participants judged each inference twice with different contents. One of the contents was typical for the scenario, and the other atypical. The allocation of scenario contents to inferences was counterbalanced across participants, leading to eight different booklets, four for each scenario. In the statements group, each participant rated the entire set of contents created for the relevant scenario, leading to two booklets, one for each scenario. In order to compensate for the difference in sample size between groups resulting from the different number of booklets in each group, a weight was placed on the otherwise random procedure for assigning participants to booklets, such that participants were twice as likely to receive any one of the booklets of the statements group than any one of the booklets of the inferences group. This resulted in sample sizes of $n = 305$ and $n = 566$ for the statements and inferences group, respectively.

Statements group:

Now consider the following statement about Linda:

Please indicate how much confidence you would have in this statement. Please give a percentage rating from 0% (no confidence at all) to 100% (complete confidence).

"Linda votes for the Labour Party or the Green Party"

Inferences group:

Now consider the following argument about Linda:

Next to A please indicate how much confidence you would have in the premise of the argument. Next to B please indicate how much confidence you would have in the conclusion, given the premise. Please give a percentage rating from 0% (no confidence at all) to 100% (complete confidence).

A. "Linda votes for the Labour Party or the Green Party."

B. "Therefore, if Linda does not vote for the Labour Party, then she votes for the Green Party."

Procedure

The experiment took place online using the platform CrowdFlower (c. f. Peer, Brandimarte, Samat, & Acquisti, 2017). On the first screen participants viewed the instructions and a sample trial. The next screen showed the scenario within which the statements, or respectively the inferences, were to be assessed. These then followed, presented one at a time on the screen. A further screen asked for demographical information, and a final screen provided debriefing information. The median duration of the experiment was 5.63 minutes in the inferences group, and 4.75 minutes in the statements group.

Results and discussion

The values of observed and chance rate coherence are shown in Figure 5.1 for each inference and group. The corresponding values of above-chance coherence are shown in Figure 5.2. Above-chance coherence was entered as the dependent variable in a linear mixed model with group (statements, inferences) and inference (the 8 inferences of Table 5.1) as predictors. The model included random intercepts for participants and for scenarios.

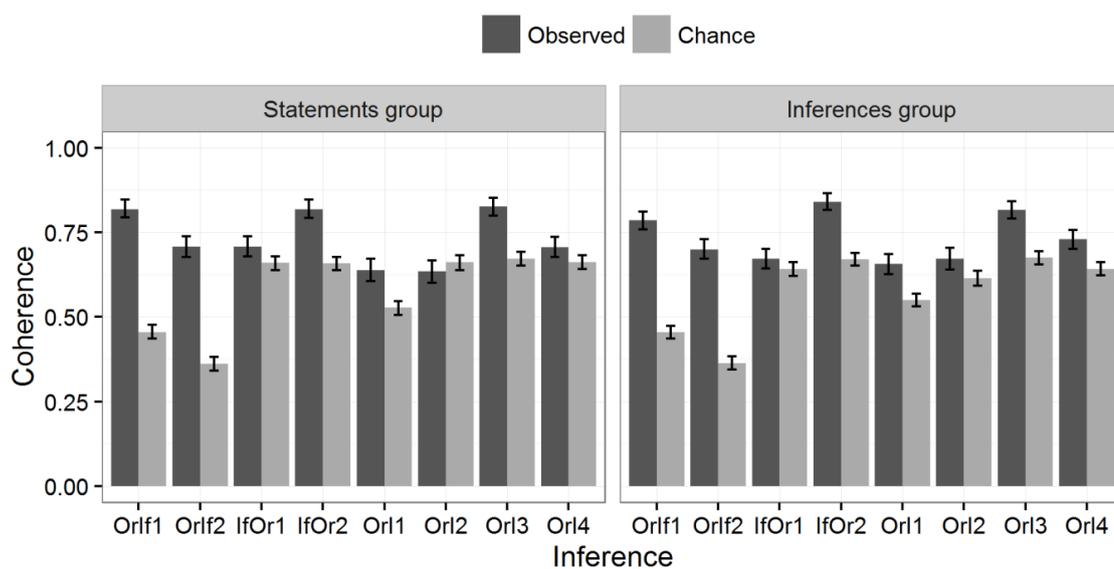


Figure 5.1. Observed and chance rate coherence for the eight inferences of Experiment 1, separately for each group. Error bars represent 95% CIs.

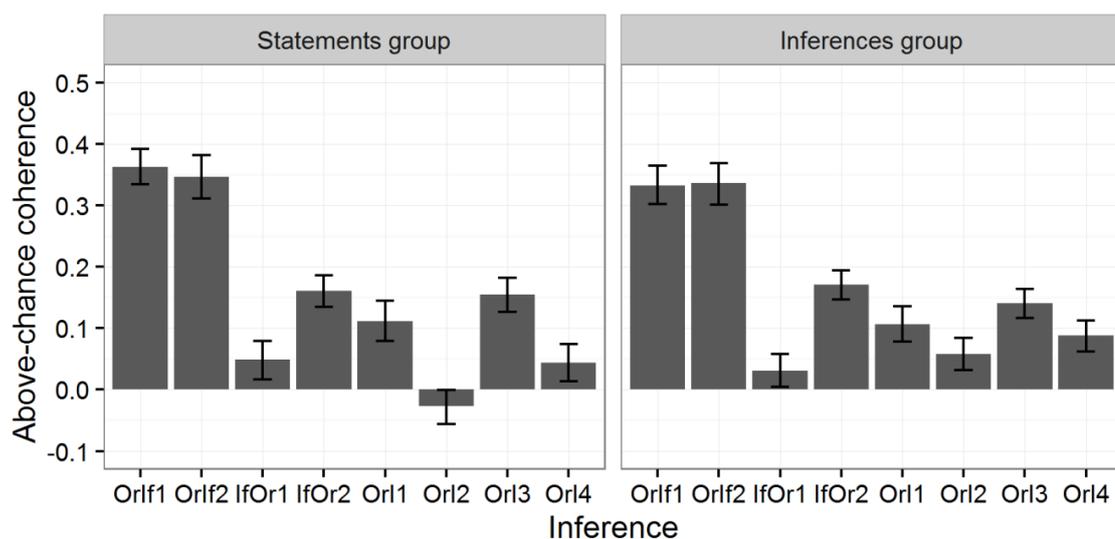


Figure 5.2. Above-chance coherence for the eight inferences of Experiment 1, separately for each group. Error bars show 95% CIs.

Overall responses were coherent 15% more often than expected by chance ($F(1, 423.048, p = .002)$). Above-chance coherence did not differ between the two experimental groups ($EMM_{stat} = .150, EMM_{inf} = .158, F(1, 561.790) = .422, p = .516$). However, above-chance coherence differed between inferences ($F(7, 13172.182) = 161.275, p < .001$), and the size of the effect of inference differed between groups ($F(7, 13172.182) = 3.618, p = .001$).

To examine further the interaction between inference and group, a series of inference specific analyses were performed. The comparisons included a random intercept for participants, but not for scenarios, because the more complex models that included a random intercept for scenarios failed to converge for the smaller sample sizes involved.

For inference 1, coherence was above chance levels ($EMM = .348, F(1, 492.682) = 614.773, p < .001$), and the extent of this effect did not differ between the statements and the inferences group ($EMM_{stat} = .363, EMM_{inf} = .332, F(1, 492.682) = 1.230, p = .268$). For inference 2, coherence was also above chance levels ($EMM = .342, F(1, 501.100) = 615.381, p < .001$), and above-chance coherence did not differ between groups ($EMM_{stat} = .347, EMM_{inf} = .337, F(1, 501.100) = .135, p = .714$).

For inference 3 coherence was again above-chance levels ($EMM = .040, F(1, 530.644) = 10.373, p = .001$), and above-chance coherence did not differ between groups ($EMM_{stat} = .049, EMM_{inf} = .031, F(1, 530.644) = .530, p = .467$). The same pattern was observed for inference 4: coherence was above chance levels ($EMM = .166, F(1, 540.080) = 228.343, p < .001$), and did not differ between groups ($EMM_{stat} = .161, EMM_{inf} = .170, F(1, 540.080) = .200, p = .655$).

Inference 5 followed the same pattern: coherence was above chance levels ($EMM = .109, F(1, 524.932) = 71.740, p < .001$), and was similar for both groups ($EMM_{stat} = .111, EMM_{inf} = .107, F(1, 524.932) = .029, p = .866$). Inference 6 was an exception to the above. Overall coherence was not significantly higher than expected by chance ($EMM = .015, F(1, 520.720) = 1.815, p = .178$), but it differed between the two groups ($EMM_{stat} = -.027, EMM_{inf} = .058, F(1, 520.720) = 13.739, p < .001$). Follow-up analyses showed that responses were at chance level in the statements group ($F(1, 204) = 2.554, p = .112$), but were above chance in the inferences group ($F(1, 456) = 14.827, p < .001$).

Inferences 7 and 8 followed a similar pattern as inferences 1 to 5. For inference 7, coherence was above chance levels ($EMM = .147, F(1, 539.761) = 195.851, p < .001$), and did not differ between groups (For inference 7: $EMM_{stat} = .155, EMM_{inf} = .140, F(1, 539.761) = .455, p = .500$). For inference 8, coherence was also above chance levels ($EMM = .066, F(1, 570.824) = 32.081, p < .001$), and was marginally, but not significantly, higher in the inferences group than in the statements group ($EMM_{stat} = .044, EMM_{inf} = .088, F(1, 570.824) = .059, p = .059$).

Overall, for 7 of the 8 inferences responses were coherent above chance levels in both the statements and the inferences group, and the presence of an explicit inference task did not increase coherence further. For inference 6, coherence was at chance level in the statements

group but above chance level in the inferences group. Hence, in the one case out of eight in which coherence was not already above chance levels when judging the probability of statements in isolation, the presence of an explicit inference task did increase response coherence. This pattern of results is in accordance with that of Evans et al. (2015). The main difference seems to be that for the simpler one-premise inferences studied here, coherence was already above chance levels in most cases in the statements group, leaving less room for observing an effect of an explicit inference task.

The confidence intervals in Figure 5.2 suggest that above-chance coherence was similar for the two equivalent inferences 1 and 2, but differed more strongly e. g. between the equivalent inferences 3 and 4. However, it would be difficult to interpret such differences in the current experiment. This is because the chance rate for a coherent response depended on the probability participants assigned to the premise, and so differed between inferences. The higher the chance rate, the more difficult it becomes to detect above-chance coherence when it is there. For example, Figure 5.2 shows that above-chance coherence was highest for inferences 1 and 2; but Figure 5.1 reveals that this was mainly because inferences 1 and 2 had a lower chance rate coherence, even though their observed coherence rate was similar to that of other inferences. Similarly, Figure 5.1 shows that the highest rate of observed coherence was found for inferences 1, 4 and 7; but differences in their chance rates meant that the rate of above-chance coherence, shown in Figure 5.2, differed strongly between the three inferences. This comparability issue will be discussed in more detail in Experiments 3 and 4. Ways to render coherence comparisons between inferences possible are explored in experiments 5 to 7.

A further analysis was conducted for inferences 1 and 2: the *or-to-if* inferences *p or q*, *therefore if not-p then q*, and *not-p or q, therefore if p then q*. This analysis compared the material and the probability conditionals as accounts of people's modal understanding of conditionals. Above-chance coherence was computed again, but this time excluding from the probability interval for a coherent conclusion the cases in which premise and conclusion were assigned the same probability.

It had been mentioned above that the coherence interval for the material conditional is respected only when the conditional and the disjunction have the same probability, whereas the coherence interval for the probability conditional is respected only when the probability of the conditional is equal to or lower than the probability of the disjunction. In this new computation of coherence, the overlapping area between the two intervals is excluded, so that a response is categorised as coherent only when it is coherent from the perspective of the probability conditional and incoherent from the perspective of the material conditional.

If people follow a material conditional interpretation for inferences 1 and 2, then the mean difference between the premise and conclusion probability is expected to be zero. There may be some scattering of probabilities above and below zero, but no systematic drift in any

direction. In an analysis using the above new measure of coherence, one would therefore predict no effect of coherence, i. e. one would expect coherence to be at chance levels.

If people follow a probability conditional interpretation for inferences 1 and 2, then one would predict probability assignments to the conclusion that are equal to or lower than the probabilities assigned to the premise. In an analysis using the above new measure of coherence, one would expect the effect of coherence to be weaker because it would exclude the subset of coherent responses for which premise and conclusion were assigned the same probability. It would even be compatible with a probability conditional interpretation that the effect of coherence ceases to be significant. In the latter case the analysis would be uninformative to the question at hand. However, a remaining effect of coherence in the expected direction would constitute specific evidence for a probability conditional interpretation and against a material conditional interpretation of the conditional.

A linear mixed model on the new measure of above-chance coherence, with inference (1, 2) and group (statements, inferences) as predictors, and participants as random intercepts, showed that responses were coherent 35% more often than expected by chance ($F(1, 530.676) = 771.528, p < .001$). No other effects were significant (all F s < 1 . For the effect of inference: $EMM_1 = .348, EMM_2 = .342, F(1, 2761.313) = .147, p = .702$. For the effect of group: $EMM_{stat} = .355, EMM_{inf} = .334$. For the interaction between inference and group, $F(1, 2761.313) = .524, p = .469$).

Overall, coherence was above chance levels when considering as coherent only those responses that are coherent for the probability conditional and incoherent for the material conditional interpretation of the conditional.

General discussion

The present experiment extended the method introduced by Evans et al. (2015) for computing the extent to which people's conclusion probability judgments are coherent above the level that would be expected by chance (see also Politzer et al., 2016, for similar considerations developed independently for qualitative probability judgements). Evans et al. as well as Singmann et al. (2014) applied this method to conditional syllogisms, and found people to be coherent above chance levels mainly for MP (see Pfeifer & Kleiter, 2009, for earlier convergent findings). Evans et al. argued that the special standing of MP in assessments of response coherence makes sense within the probabilistic approach. It is the inference at the centre of Bayes rule, and by extension, at the centre of dynamic reasoning by conditionalization. However, the ability to establish coherence between assertions is central to reasoning and decision making as a whole, as illustrated by the fact that it guarantees the avoidance of Dutch books (Vineberg, 2016). As such, it is far more general than the ability to

reason in accordance with Bayes rule. There is therefore a case for exploring people's sensitivity to coherence further before concluding that it may be restricted to the use of specific inferences like MP.

The present experiment assessed above-chance coherence for a series of simpler one-premise inferences, rather than for the conditional syllogisms. Replicating Evans et al. (2015), it had two task conditions. In the inferences task participants were asked to judge, for each inference, the probability of the premise, and the probability of the conclusion given the premise. In the statements task the statements that constituted the premises and conclusions in the inferences task were presented in isolation, one at a time in random order. Coherence was found to be above chance levels in 7 out of 8 cases in the statements task. In the one case in which coherence was at chance levels in the statements task, coherence was above chance levels in the inferences task. It is interesting that in this one case, *not-p or q* from *not-p*, the inference has a negation as its premise, which is also present as a component of its conclusion.

This pattern of results suggests a much more optimistic picture of people's sensitivity to coherence constraints than had been obtained in the study of the more complex conditional syllogisms. The results suggest that people have a spontaneous tendency to give coherent responses; and that in cases in which coherence is not established automatically – for instance due to the presence of negations or of more complex inferential structures – it is increased through the context of an explicit inference task (see De Neys, 2012, and Trippas, Handley, Verde, & Morsanyi, 2016, for similar arguments about the role of inference complexity for detecting sensitivity to deductive constraints).

As mentioned above, it was not feasible in this experiment to compare the degree of above-chance coherence between individual inferences. However, it is notable that in the statements group, responses were coherent above chance levels in three of the four versions of the or-introduction inference, and that in the inferences group responses were coherent above chance levels for all four versions. This is in accordance with the hypothesis that under probabilistic instructions, the pragmatic awkwardness of drawing the inference is reduced because people are asked directly about their beliefs. As a result, people seem to treat the inference as valid just as they do other less controversially valid inferences. This finding also goes counter to the proposal in a recent revision of mental model theory (Johnson-Laird et al., 2015) that the lower endorsement rates of the inference under binary instructions reflect the fact that the inference is actually invalid. If the inference were considered invalid, one would expect it to be rejected across a variety of experimental conditions, and not only in contexts in which there are strong alternative pragmatic reasons for the rejection (see Politzer et al., 2016, and Cruz et al., 2017, for further evidence that, under probabilistic instructions, or-introduction is endorsed to a similarly high degree as other, less controversially valid inferences).

An additional analysis showed that responses were coherent above chance levels under the assumption that the natural language conditional is interpreted as the probability

conditional, but incoherent above chance levels under the assumption that the conditional is interpreted as the material conditional. This constitutes a strong and new form of evidence for the descriptive adequacy of the probability conditional and against that of the material conditional.

The experiment that follows below explores whether an effect of increased coherence under an inferences task relative to a statements task, found in cases in which responses were not already coherent to begin with, can be extended to contexts in which people have been found to be pervasively incoherent, specifically that of the conjunction fallacy.

EXPERIMENT 2: IFS, ANDS, AND THE CONJUNCTION FALLACY⁶

Overview of the conjunction fallacy

The conjunction fallacy refers to the judgment that $P(p \ \& \ q) > P(p)$, which is not possible in classical probability theory because $p \ \& \ q$ is a subset of p . So for instance, the set of people who are bank tellers and feminists is equal to or smaller than the set of people who are bank tellers, because some bank tellers may be feminists and others not.

This scenario about bank tellers and feminists, also used in Experiment 1, was the initial context in which the fallacy was found by Tversky & Kahneman (1983), who coined the term. The original format of the task was one of a list of around eight statements: the two critical ones about being a bank teller (B), and being a bank teller and active in the feminist movement ($B \ \& \ F$), and six filler items. The task was to rank the probability of the statements from the most to the least likely. But the basic finding was replicated when the task was to judge the probability of each statement, or to judge which of two arguments was better, one stating that (B) is more likely due to set-inclusion relations, and the other stating that ($B \ \& \ F$) is more likely due to its higher representativeness (Tversky & Kahneman, 1983). The latter makes the set-inclusion relation between the two statements entirely explicit. The finding was also replicated when the statement "she is a bank teller" was replaced with "she is a bank teller whether or not she is active in the feminist movement" (Tversky & Kahneman, 1983) or when participants were given the three options $B \ \& \ F$, $B \ \& \ \text{not-}F$, and B (Bonini, Tentori, & Osherson, 2004), thus avoiding the possibility that people might interpret B as $B \ \& \ \text{not-}F$. Further, the finding was replicated when the terms "probability" and "conjunction" were not mentioned, and participants were simply asked to choose the most likely of a sequences of throws of a die, given a particular constellation of the die sides (Tversky & Kahneman, 1983; see also Politzer & Noveck, 1991, and Tentori et al., 2004, on the roles of terminology and pragmatic implicatures). The incidence of the fallacy was sometimes reduced, but still present, in betting contexts and when participants had sophisticated knowledge of statistics (Bonini et al., 2004; Sides, Osherson, Bonini, & Viale, 2002; Tversky & Kahneman, 1983).

The finding of the conjunction fallacy generated an explosion of further studies and attempts to provide an explanation for it. Tversky and Kahneman (1983) themselves attributed it to the use of the heuristics of representativeness and availability. They distinguished two situations in which the fallacy occurs: when there is a model or situation (e. g. the description of Linda's personality) that is highly representative of one element of the conjunction (A) and

⁶ The data from this experiment was published using an ANOVA analysis in a collaborative paper with Jean Baratgin, Mike Oaksford, and David Over (Cruz et al., 2015).

highly unrepresentative of the other (B) (the M-A paradigm); and when A and B are not necessarily related to a specific model or situation, but there is a causal, correlational or otherwise explanatory relation between the two elements of the conjunction A and B (the A-B paradigm). In line with this, the incidence of the fallacy was greatly reduced in M-A situations when the scenario description was omitted (e. g. it was only stated that Linda is a 31 year old person), and it was greatly reduced in A-B situations when the causal link between A and B was severed, while leaving the probability of each constant. For example, in one scenario participants were told that two unrelated individuals were randomly selected among respondents of a representative health survey. Participants were asked about the probability that the first person has had one or more heart attacks, and that the second person is over 55 years old. Even though being over 55 years old is associated with a higher probability of having had one or more heart attacks, this association plays no role when the features of age and of medical history for heart attacks refer to different individuals.

The authors argued that natural, spontaneous assessments e. g. of similarity, representativeness, or causal relations can bias probability judgments in three ways. They can be used explicitly as heuristics for solving the task, they can exert an influence by providing an anchor or association that diverts the probability judgment, or people may find it difficult to distinguish between the natural assessment triggered by the materials and the assessment required by the task.

Two accounts of the fallacy that may at least partly be considered specifications or extensions of the explanation brought forward by Tversky & Kahneman (1983), are that of Crupi, Fitelson, & Tentori (2008), and that of von Sydow (2017). Crupi et al. argued that the fallacy may arise because the answer to the question about the probability of the statements is biased by the availability of an answer to another question: that of inductive Bayesian confirmation. The evidence about the personality of Linda provides better inductive confirmation of the hypothesis $B \& F$ than it does of the hypothesis B (where inductive confirmation is a measure assessing the extent to which a person increases their belief in a hypothesis after learning the evidence, compared to before obtaining the evidence). Von Sydow (2017) argues that the fallacy can be accounted for by the idea that people sometimes compute probabilities in an intensional as opposed to extensional way. With an intensional reading, there is a specific data pattern that is approximated best by the hypothesis of a conjunction, and another data pattern that is approximated best by the hypothesis of one of the elements of the conjunction. But the two hypotheses refer to separate concepts and are not nested. Along these lines, the data describing the personality of Linda fits, or confirms (Crupi et al., 2008) more the $B \& F$ hypothesis than the B hypothesis, regardless of the fact that from an extensional perspective, $B \& F$ is a subset of the B .

An account that looks at the incoherence of the conjunction error from a different perspective, offering a defence of the in principle rational processes of probability estimation

that lead to it, is that of Costello & Watts (2014). They propose that the findings on the conjunction fallacy, as well as on other biases in probability judgment like the disjunction fallacy, subadditivity and conservatism, can be explained as resulting from probability estimates that follow probability theory but are subject to random noise. This noise is larger for composite than for simple events, and it can be cancelled out through specific algebraic computations. The authors provide impressive empirical evidence for the precise predictions made by their theory. They have also extended the account to reasoning with conditional probabilities (Costello & Watts, 2016a) and to findings previously argued to support the approach of quantum probability theory (Costello & Watts, 2017; see also Sanborn & Chater, 2016, for a related perspective). Although the probability theory plus noise account may not explain all instances of the fallacy, the relevance of its findings goes beyond it and builds a counterweight to the findings of the fallacy, which accounts suggesting people have no sensitivity to the principles of probability theory are in need to explain.

Two accounts that depart more strongly from the original interpretation of the findings are the frequentist (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) and the quantum probability account (Pothos & Busemeyer, 2013; Pothos, Busemeyer, Shiffrin, & Yearsley, 2017). In brief, the frequentist argument is that people are naturally equipped to reason in accordance with the principles of probability theory when probabilities are presented as natural frequencies, but not when they are presented as single event probabilities. Contrary to this proposal, several studies have shown that the frequentist format does not always lead to lower fallacy rates, and that it can promote biases of its own. The effect of the frequency format has been explained as resulting from a more transparent description of the nested set relations involved in the problem, and it has been shown that such transparency in the problem structure, with corresponding reductions in fallacy rates, can be achieved in other ways than through frequency formats (Evans, Handley, Perham, Over, & Thompson, 2000; Sloman, Over, Slovak, & Stibel, 2003; Tentori et al., 2004).

The quantum probability perspective is that the conjunction effect is not in fact a fallacy, but a consequence of the contextual representation of events. It argues that the probability of B in the context of its conjunction with F may not be the same as the probability of B presented on its own. If we hear the statement that Linda is a bank teller, this may initially seem very improbable, but if we first hear that she is a feminist and then that she is a bank teller, we may be more amenable to the fact that bank tellers come in many shapes and sizes. This is an order effect, which could also be described as a failure of invariance in dynamic reasoning (Oaksford, 2013; Oaksford & Chater, 2013). It is argued to occur when people reduce the dimensionality of the problem space, e. g. because of processing limitations. The predictions of the theory for unreduced problem spaces are equivalent to those of classical probability theory. One strength of the quantum approach is that it allows the formulation of precise predictions about the size of contextual effects when they occur. However, there currently

seem to be no clear criteria for predicting when contextual effects will occur and when not. Further, the normative, computational level foundation for the quantum approach seems less clear than for classical probability theory, particularly with respect to coherence and the avoidance of Dutch books (Hahn, 2014). Given this uncertain foundation and the fact that the findings used to support it can also be explained within classical probability theory (Costello & Watts, 2016b, 2017; Kellen, Singmann, & Batchelder, 2017), it seems as though more arguments are needed to appreciate the benefits of the arguably less parsimonious quantum approach.

Tversky & Kahneman (1983), and Crupi et al. (2008) further argue that a finding hard to explain within accounts positing that the conjunction effect is not a fallacy, is that participants in experiments usually do not defend their conjunction effect answers when debriefed about the normative solution in classical probability theory. They instead usually concede having made an error, and seem to experience some regret for it. This was one of the reasons for why Tversky and Kahneman described it as a fallacy rather than as a misunderstanding.

Ifs and ands

The present experiment addresses this last point, concerning the conditions under which the normative solution in classical probability theory is transparent. It is not directly concerned with a comparison of explanations for the conjunction fallacy, but rather with factors that might increase the coherence of reasoning in the special case of this fallacy, in which the modal answer is an incoherent one.

In addition, this experiment extends the study of conditionals and disjunctions reported in Experiment 1 to conditionals and conjunctions. It looks at whether it makes a difference to people whether an inference from *and* to *if* has one premise: *p & q, therefore if p then q*, or two premises: *p, q, therefore if p then q*. Normatively the coherence intervals differ for the two cases. For the one-premise case, the interval for the conclusion lies between the probability of the premise and 1. For the two-premise case, the interval is more complex. Assuming the conditional is defined through the Equation, it is given by:

$$P(q|p) \in [\max\{0, (P(p) + P(q) - 1)/P(p)\}, \min\{P(q)/P(p), 1\}]$$

Depending on the premise probabilities, the minimum value of this interval may lie above 0, and the maximum value may lie below 1. This means that conclusion probability judgments for this inference can be incoherent in two directions, resulting in either underconfidence or overconfidence. In contrast, in the one-premise version it is only possible to be incoherent by being underconfident in the conclusion.

The above interval presupposes an interpretation of conditionals based on the Equation. But the coherence bounds for this interpretation are stronger than those for the material conditional. Therefore, any response that is coherent for the above interpretation is also coherent for the material conditional interpretation.

Method

Participants

Forty-eight students from the University of Orsay, France, took part in the experiment on a voluntary basis. Their mean age was 20 years (range 18-24). They had different majors, although the majority studied biology or medicine. All participants were French native speakers.

Material and design

The material and design were similar to those of Experiment 1. However, only the Linda scenario was used, and because the original inferences contained no negations, no variation of the inferences on the basis of the position of negated terms was introduced. The four inferences investigated are shown in Table 5.2.

Table 5.2. The inferences used in Experiment 2.

Name	Form
1 and-to-if (one premise)	$p \ \& \ q \ \therefore \ \text{if } p \ \text{then } q$
2 and-to-if (two premises)	$p, \ q \ \therefore \ \text{if } p \ \text{then } q$
3 and-elimination (prototypical)	$p \ \& \ q \ \therefore \ p$
4 and-elimination (counter-prototypical)	$p \ \& \ q \ \therefore \ q$

The and-to-if inferences (also called *centering*, Cruz et al., 2016; Over & Cruz, 2018), served to assess whether response coherence differed as a function of whether the inference from a conjunction to a conditional had one or two premises; in particular, whether coherence was lower in the two-premise version than in the one-premise version. This may occur because the more complex interval for the two-premise version requires the integration of two premise probabilities p , and q , taking into account the minimum and maximum degree to which p and q may overlap, or co-occur. In contrast, in the simpler interval of the one-premise version, a fixed co-occurrence of p and q is already given as the probability of the premise, and so no integration is required. These two inference versions are shown again in Table 5.3 in the context of the Linda scenario. The inferences had contents that seemed likely given the

scenario description, to encourage higher premise probability judgments and so lower probabilities of conforming to coherence just by chance.

The and-elimination inferences of Table 5.2 served to assess whether the incoherence of the conjunction fallacy is reduced when people are asked to make an explicit inference from one of the statements involved in the fallacy to the other. Inferences 3 and 4 are shown again in Table 5.3 embedded in the content of the Linda scenario. One can see that inference 3 has prototypical content, and inference 4 has counter-prototypical content for the scenario. The experiment included six further inferences that are not relevant for the present study, and are not reported further.

Table 5.3. The four inferences of the experiment embedded in the Linda scenario.

-
- 1 Linda votes in the municipal elections and she votes for the Socialist Party. Therefore, if Linda votes in the municipal elections, then she votes for the Socialist Party.
 - 2 Linda votes in the municipal elections. Linda votes for the Socialist Party. Therefore, if Linda votes in the municipal elections, then she votes for the Socialist Party.
 - 3 Linda is a feminist and she works in a bank. Therefore, Linda is a feminist.
 - 4 Linda is a feminist and she works in a bank. Therefore, Linda works in a bank.
-

As in Experiment 1, the inferences were presented under two task conditions: a statements task and an inferences task. In the statements task the statements appeared as single items on a list, and participants were asked to judge how much confidence they would have in each statement. In the inferences task, the statements were arranged into the premises and conclusion of inferences, and each inference was presented on a separate page. Participants were asked to judge how much confidence they would have in the premise of the inference, and how much confidence they would have in the conclusion, given the premise. Participants in both groups provided their answers by writing a percentage number into a box next to each relevant statement, ranging between 0 ("no confidence at all"), to 100 ("complete confidence"). Four booklets were created for each task condition, which varied only in the order in which the statements resp. the inferences were presented.

Procedure

Participants were tested in the university library in small groups of up to four participants. They worked at their own pace, and took 10 to 15 minutes to complete.

Results and discussion

The data was analysed using linear mixed models with above-chance coherence as the dependent variable, and with inference and task as predictors. Given the relatively low ratio of participants to data points, the only random effects included in the models were random intercepts for participants.

And-to-if inferences

The first analysis looked at the and-to-if inference, and assessed whether the more complex coherence interval for the two-premise version than for the one-premise version was associated with a higher frequency of coherent responses for the latter than for the former. The results are shown in the left panel of Figure 5.3. A linear mixed model for the effects of inference and task on above-chance coherence showed that overall, responses were 42% more likely than expected by chance ($EMM = .417$, $F(1, 80.686) = 17.931$, $p < .001$). No other effects were significant (for inference: $F(1, 46) = 1.028$, $p = .316$; for task: $F(1, 80.686) = .004$, $p = .941$; for the interaction: $F(1, 46) = .049$, $p = .826$). There was thus no indication that the computation of a coherent conclusion probability was more difficult for the two-premise version than for the one-premise version of the inference.

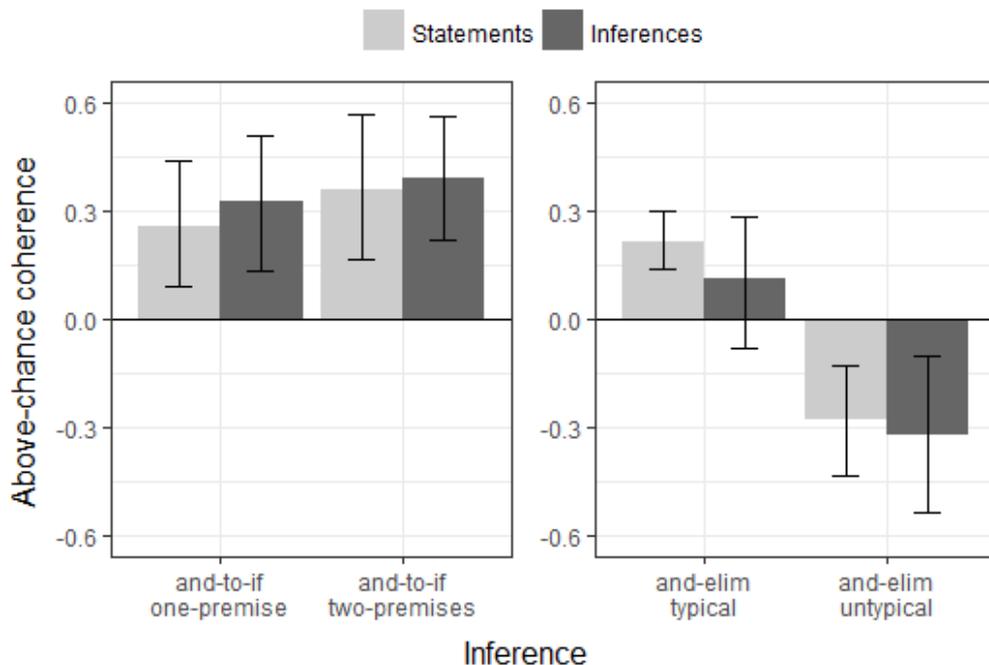


Figure 5.3. Above-chance coherence for the four inferences and the two task conditions of Experiment 2. Error bars show 95% CIs.

And-elimination inferences

To assess whether the frequency of occurrence of the conjunction fallacy is reduced in the context of an explicit inference task, a linear mixed model was computed for the effects of inference content and task on above-chance coherence, again with random intercepts for participants. The results are shown in the right panel of Figure 5.3. Overall coherence was again above chance levels ($EMM = .216$, $F(1, 90.199) = 7.194$, $p = .009$). The extent to which coherence was above-chance levels was larger for inference 3 with prototypical content than for inference 4 with counter-prototypical content ($EMM_{typical} = .166$, $EMM_{untypical} = -.301$, $F(1, 46) = 33.314$, $p < .001$). No other effects were significant (for task: $F(1, 90.199) = .661$, $p = .418$; for the interaction: $F(1, 46) = .126$, $p = .724$).

Follow-up analyses for each inference confirmed that coherence was above chance for inference 3 ($EMM = .166$, $t(47) = 3.201$, $p = .002$); but below chance for inference 4 ($EMM = -.301$, $t(47) = -4.328$, $p < .001$).

Overall, for the same logical structure responses were coherent above chance levels when the content of the conclusion was in accordance with the scenario, but below chance level when the content was untypical for the scenario, the latter being the content originally found to lead to the conjunction fallacy (Tversky & Kahneman, 1983). The frequency of the fallacy was thereby not reduced by the presence of an explicit inference task.

General discussion

The finding that coherence for the and-to-if inferences was above chance level extends the findings of Experiment 1 to inferences relating conjunctions and conditionals. The observation that responses were coherent above chance level in the statements task, so that there was little opportunity for coherence to increase further in the inferences task, is in line with the findings of Experiment 1 as well as with Evans et al. (2015). It suggests that an explicit inference task can be most helpful in cases in which coherence is not already above chance levels to begin with. That coherence was above chance levels in the statements task suggests people may tend to comply with it in a spontaneous, intuitive way, even in contexts in which there is no explicit suggestion that it should be taken into account. This was not originally expected, but is in line with findings on good conformity to probabilistic principles in domains outside of reasoning, where tasks are carried out in a more implicit way, like perception and language comprehension (Fiser, Berkes, Orbán, & Lengyel, 2010; Hsu, Chater, & Vitányi, 2011).

It also made no difference to coherence whether the inference from the conjunction to the conditional had one or two premises. Before drawing further conclusions from this finding, it would be worth assessing how far it can be generalised to other inferences, as well as to

situations in which people have differing knowledge of the overlap between p and q that would allow them to narrow down their conclusion probability estimates.

People's responses to the and-elimination inferences were coherent at levels above chance when the materials did not involve a conflict between prototypicality and probability (Kahneman & Tversky, 1983), or between informativeness about an inductive hypothesis and probability (Tentori, Crupi, & Russo, 2013; von Sydow, 2017). This is in line with further findings showing responses to be coherent for this inference when using neutral contents (Pfeifer & Kleiter, 2005; Politzer & Baratgin, 2016; Tversky & Kahneman, 1983). However, when the material did have the content known to cause the fallacy, responses were incoherent above chance levels, and it made no difference whether the task was to rate the probability of each statement individually, or to draw an explicit inference from the probability of one statement to the probability of the other. This finding does not directly help arbitrate between different accounts of the fallacy, but it extends its scope to a further arguably transparent context (Tversky & Kahneman, 1983), underlining the view that it is a deep fallacy that is hard to overcome.

The following two experiments extend the study of people's intuitions about coherence constraints to further inferences, and to different kinds of materials. They also provide further discussion of the role of premise probabilities, and with them the chance rate of a coherent response, for the interpretation of above-chance coherence, setting the scene for an implementation of those discussion points in the next chapter.

EXPERIMENTS 3 AND 4: INTUITION, REFLECTION, AND WORKING MEMORY

Experiment 1 found people's judgments of conclusion probability to be coherent at above chance levels for a range of simple one-premise inferences (Cruz et al., 2015). This finding has been replicated and extended to further one-premise inferences, as well as to an ordinal response format (Politzer & Baratgin, 2016). These consistent findings suggesting that people are sensitive to coherence constraints appear to be at odds with the findings for conditional syllogisms (Evans et al., 2015; Singmann et al., 2014; c. f. Pfeifer & Kleiter, 2009), for which coherence at above chance levels was observed only for MP and to a lesser extent for DA.

Evans et al. (2015) further observed that more responses were coherent in an explicit inference task than in a task asking participants to judge the probabilities of the statements in the inferences presented one at a time in random order. Specifically, they found that in the statements task, probability judgments were coherent at above chance levels for MP and DA, but were *incoherent* at above chance levels for MT and AC. In the inferences task, coherence was above-chance levels for MP, DA and also AC, whereas it remained below chance levels for MT. However, no significant difference in above-chance coherence between the statements and the inferences task was found in Experiment 1 of this thesis.

What led to the differences between these studies? Were responses reliably coherent in Cruz et al. (2015) and in Politzer & Baratgin (2016) because they involved simpler inferences? Or was it because the chance rate of conforming to coherence in Evans et al. (2015) and in Singmann et al. (2014) was too large to allow a reliable detection of above-chance performance? Did Cruz et al. fail to find a difference between the statements and the inferences task because the simpler inferences they studied were already reliably coherent across both tasks, or was it because the study of Cruz et al. was performed on the internet, whereas that of Evans et al. was conducted in the lab, and the data from the internet experiment might have had a higher error variance? Or does the failure to replicate the effect of task indicate that the effect itself is not reliably present?

These questions were addressed in two follow-up experiments. Experiment 3 was conducted in the lab, and Experiment 4 on the Internet. Both experiments included, in addition to the statements and an inferences task, a third condition in which participants worked through an inferences task with working memory load. The aim of this third condition was to assess whether the difference between the statements and the inferences task can at least in part be explained through lower working memory demands involved in the inferences task.

Experiment 3

Method

Participants. A total of 142 participants from the recruitment pool of the Department of Psychological Sciences of Birkbeck, University of London, completed the experiment. They were divided into three groups: a statements group (Group 1), an inferences group (Group 2), and an inferences group with working memory load (Group 3) (see the design and materials section for the characterisation of each group). All participants indicated at the end of the experiment that they had taken part seriously, as opposed to just "clicking through". Three participants (one from each group) were excluded because they had problems understanding the task; eight were excluded because they failed a catch trial asking them not to respond but instead to click "next" to continue with the experiment; three participants were excluded from the statements group because they had two or more trial reaction times of 2 seconds or less, and two from the inferences groups for having two or more trial reaction times of 3 seconds or less. Finally, one further participant was excluded for not reporting at least "good" English language skills. The final sample consisted of 131 participants (42 in Group 1, 50 in Group 2, and 39 in Group 3). Participants' median age was 24 (range 18-56). Most had some university education, with 64% reporting an undergraduate degree, 22% a postgraduate degree and 3% a doctoral degree. 8% reported having finished 12th grade, and 2% having a technical/applied degree. Participants' median ratings of task difficulty were 24% in Group 1, 51% in Group 2, and 62% in Group 3.

Materials and design.

Inferences. The experiment investigated the 12 inferences shown in Table 5.4. These inferences were classified into three groups based on the complexity of the coherence intervals for their conclusion. The inferences of type A (inferences 1 and 2) are one-premise equivalences or contradictions. The coherence interval for these inferences is a point value equal to the probability of the premise (in the case of equivalences) or to the complement of the probability of the premise (in the case of contradictions). Inference 1 is called *De Morgan* (DM) and shows how a conjunction can be transformed into an equivalent disjunction. Inference 2, *not De Morgan* (nDM), is the result of deleting the negation in the premise of the DM inference.

The inferences of Type B (inferences 3 to 8) are one-premise inferences describing set-subset relations, making them p-valid in one direction and p-invalid in the other. Inferences 3, 5 and 7 (*and-elimination*: &I, *and-to-or*: &Or, and *if-to-or*: IfOr) are p-valid, and the coherence interval for their conclusion goes from the probability of the premise (inclusive) to 1. Inferences 4, 6 and 8 (*and-introduction*: &I, *or-to-and*: Or&, and *or-to-if*: OrIf) are p-invalid, and the coherence interval for their conclusion goes from the probability of the

premise (inclusive) to 0. Inferences 7 and 8 (IfOr and OrIf) were also tested in Experiment 1 (Cruz et al., 2015).

Table 5.4. The inferences investigated in Experiments 3 and 4.

Type	Name	Validity	Form
A. One-premise, equivalence and contradiction	1. De Morgan (DM)	1	$\text{not}(p \ \& \ q) \ \therefore \text{not-}p \ \text{or} \ \text{not-}q$
	2. not De Morgan (nDM)	0	$p \ \& \ q \ \therefore \text{not-}p \ \text{or} \ \text{not-}q$
	3. and-elimination (&E)	1	$p \ \& \ q \ \therefore p$
	4. and-introduction (&I)	0	$p \ \therefore p \ \& \ q$
B. One-premise, valid in only one direction, left to right, or right to left	5. and-to-or (&Or)	1	$p \ \& \ q \ \therefore p \ \text{or} \ q$
	6. or-to-and (Or&)	0	$p \ \text{or} \ q \ \therefore p \ \& \ q$
	7. if-to-or (IfOr)	1	$\text{if not-}p \ \text{then} \ q \ \therefore p \ \text{or} \ q$
	8. or-to-if (OrIf)	0	$p \ \text{or} \ q \ \therefore \text{if not-}p \ \text{then} \ q$
C. Two-premise, conditional syllogisms	9. Modus ponens (MP)	1	$\text{if } p \ \text{then } q, p \ \therefore q$
	10. Modus tollens (MT)	1	$\text{if } p \ \text{then } q, \text{not-}q, \ \therefore \text{not-}p$
	11. Affirmation of the consequent (AC)	0	$\text{if } p \ \text{then } q, q, \ \therefore p$
	12. Denial of the antecedent (DA)	0	$\text{if } p \ \text{then } q, \text{not-}p, \ \therefore \text{not-}q$

Note. "1" = "valid", "0" = "invalid", "∴" = "therefore".

The inferences of Type C (inferences 9 to 12) are two-premise inferences. They have more complex coherence intervals that do not stand in a simple relation to the probabilities of the premises. The four inferences of this type included here are the conditional syllogisms, which were also investigated by Evans et al. (2015). Their intervals are displayed in Table 5.5.

Experimental groups. Participants were divided into three groups: (1) a statements group, (2) an inferences group, and (3) an inferences group with a secondary working memory load task. Groups (1) and (2) were similar to the groups in Evans et al. (2015) and Cruz et al. (2015). Group (3) was introduced to assess the extent to which any benefit of having an explicit inference task would decrease with increasing working memory load. Groups (1) and (3) thus constitute alternative comparison conditions to group (2), to help ascertain more precisely where the difference between the inferences group and the statements group in Evans et al. (2015) comes from.

Table 5.5. The coherence intervals for the four conditional syllogisms.

Inference	Coherence interval for the conclusion	
	lower bound	upper bound
MP	xy	$xy+(1-y)$
MT	$\max[(1-x-y)/(1-x), (x+y-1)/x]$	1
AC	0	$\min[y/x, (1-y)/(1-x)]$
DA	$(1-x)(1-y)$	$1-x(1-y)$

Note. x = the probability of the major premise, y = the probability of the minor premise, min = minimum value, max = maximum value. For example, both MP and MT have *if p then q* as their major premise. But the minor premise for MP is p , while the minor premise for MT is *not-q*. Taken from Evans et al. (2015).

Inferences group. Participants in Group 2 were shown one inference at a time on the screen, with the premises and conclusion presented at the same time. Participants were asked to judge how likely it was that each premise was true, and how likely it was that the conclusion was true, given the likelihood of the premise(s). Participants indicated their probability judgments for the premise(s) and conclusion by clicking on a slider, with the anchors "0% likely/certainly false" and "100% likely/certainly true". The slider for the conclusion was shown in a different colour to set it apart from the sliders for the premises. This task was similar to the one used in Experiment 1.

Statements group. Participants in Group 1 were shown the same statements that they would have seen in the inferences task, but one statement at a time on the screen in random order. For example, what would have been the first premise of MP in the inferences task could have been shown in trial 1, followed by unrelated statements involved in other inferences, followed by the conclusion of MP on trial 5, followed by more unrelated statements, followed by the second premise of MP on trial 8. For each statement participants were asked to judge how likely it was that the statement was true, using the same type of slider as for the inferences task.

The randomisation procedure for the inferences group was the same as that of Experiment 1: All participants were shown all inferences and scenarios, but only a particular pairing of inference with scenario, and this pairing was randomly generated for each participant. However, the randomisation procedure for the statements group differed between the present experiment and Experiment 1 above. In the present experiment, the procedure used to allocate scenarios randomly to inferences for each participant was the same for the statements and for the inferences group. It was only in a second step that the inferences for the statements group were separated into their component statements so that they could be displayed one at a time

on the screen in random order. In contrast, in Experiment 1 participants in the statements group received all pairings of statement type with scenario.

The randomisation procedure in Experiment 1 was chosen with the aim of making the statements and the inferences group parallel in the time it took to complete the task. The faster processing time for each trial in the statements task was balanced by adding further trials to that experimental condition. The procedure used in Experiment 3 instead aimed at a parallel between conditions in the amount and diversity of material worked through by each participant, so that the only difference in the materials between the two experimental conditions was whether the statements were shown grouped into inferences, or one at a time in random order. Hence in Experiment 3, a parallel between conditions in the amount and variety of the materials was given priority over a parallel between conditions in task duration.

Inferences group with working memory load. In Group 3 the inferences task was interleaved with a visuospatial memory task similar to the one used in studies by De Neys and colleagues (De Neys, 2006; Franssens & De Neys, 2009). Before each trial of the inferences task (which was itself identical in the conditions with and without working memory load) participants were shown a dot pattern for a short period of time. Participants were instructed to try to remember this pattern because they would be asked to reproduce it after working through the trial of the inferences task.

As in Franssens & De Neys (2009), the dot pattern consisted of four dots in a 3x3 grid, displayed in the centre of the screen for 900 ms. At the end of the inferences task trial, an empty grid appeared and participants were asked to click with the mouse on the locations of the grid in which the dots had been.

The dot patterns were selected randomly for each trial and participant out of a pool of 48 patterns created by varying the possibilities in which a pattern could feature three dots on a single line (horizontal, vertical or diagonal). The presence of three aligned dots simplified the patterns, reducing the risk of a floor effect in either the memory or the reasoning task.

A further difference between this and the other two groups concerned the presence of time pressure. In Groups 1 and 2 participants were instructed to take their time to think through the questions and answer as carefully as they could. In contrast, participants in Group 3 were instructed to respond to the inference and the memory task as quickly but also as accurately as they could. The inclusion of time pressure in the instructions aimed to prevent participants from developing mnemonic or other strategies that could reduce the load of the parallel task on working memory.

In Group 3, the percentage of correct responses to the 26 trials of the memory task ranged between 15.4% and 92.3% (median: 69.2%). The chance rate for a correct response in the memory task, that is, for identifying correctly the four locations on the 3x3 grid, is $9 \times 8 \times 7 \times 6 = 1/3024$ per trial. This suggests that the manipulation was successful in creating an additional

task for participants. The upper panel of Figure 5.4 shows the distribution of correct responses in the memory task.

Premise probabilities. Previous studies investigating whether people's responses to conditional syllogisms are coherent above chance levels (Evans et al., 2015; Singmann et al., 2014) used real world materials. The researchers in those studies selected materials to elicit a variety of premise probabilities in participants, with the aim of generalising the results across premise probabilities. But a possible problem with this procedure is that it did not systematically consider the effect of premise probabilities on the chance-rate of assigning a coherent probability to the conclusion. As a consequence, the chance rate might have been too large in some conditions to allow the reliable detection of above-chance performance. See for example Figure 3 of Singmann et al. (2014), where the chance rate was the unit interval for some participants and conditions, rendering above-chance coherence = 0 by definition.

The way that premise probability determines the chance rate depends on the form of the inference, and in particular on whether the inference is p-valid or p-invalid. Consider the classification of inferences in Table 5.4 as of type (A), (B), or (C). Because the coherence interval for inferences of type A is always a point value, p-validity has no effect on the width of the interval for these inferences. But p-validity affects the width of the interval, and with it the chance rate, for inferences of type (B) and type (C).

For inferences of type (B) the coherence interval for p-valid inferences goes from the probability of the premise to 1, and the coherence interval for p-invalid inferences goes from the probability of the premise to 0. This means that for the p-valid inferences, the lower the probability of the premise, the higher the chance-rate of conforming to coherence, and hence the harder it is to detect above-chance coherence when it is there. So a sensitive test of coherence for p-valid inferences of type (B) requires high premise probabilities. For the p-

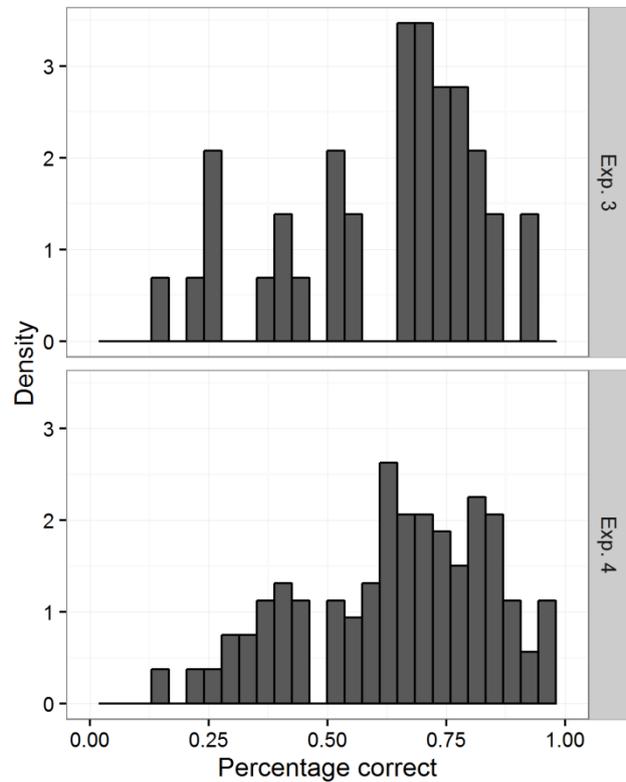


Figure 5.4. Distribution of the proportion of correct responses to the memory task in Group 3. Upper panel: for Experiment 3; lower panel: for Experiment 4.

invalid inferences the opposite relation holds: the higher the probability of the premise, the higher the chance-rate of conforming to coherence, and so the harder it is to detect above-chance coherence when it is there. Therefore, a sensitive test of coherence for p-invalid inferences of type (B) requires low premise probabilities.

The materials for the one-premise inferences (Types A and B) were varied with the aim of creating two conditions, one with high premise probability (*high condition*) and the other with low premise probability (*low condition*). The idea was to ensure that there is a condition for both p-valid and p-invalid inferences in which the test for above-chance coherence is highly sensitive, making it more likely that it will be sensitive enough across conditions.

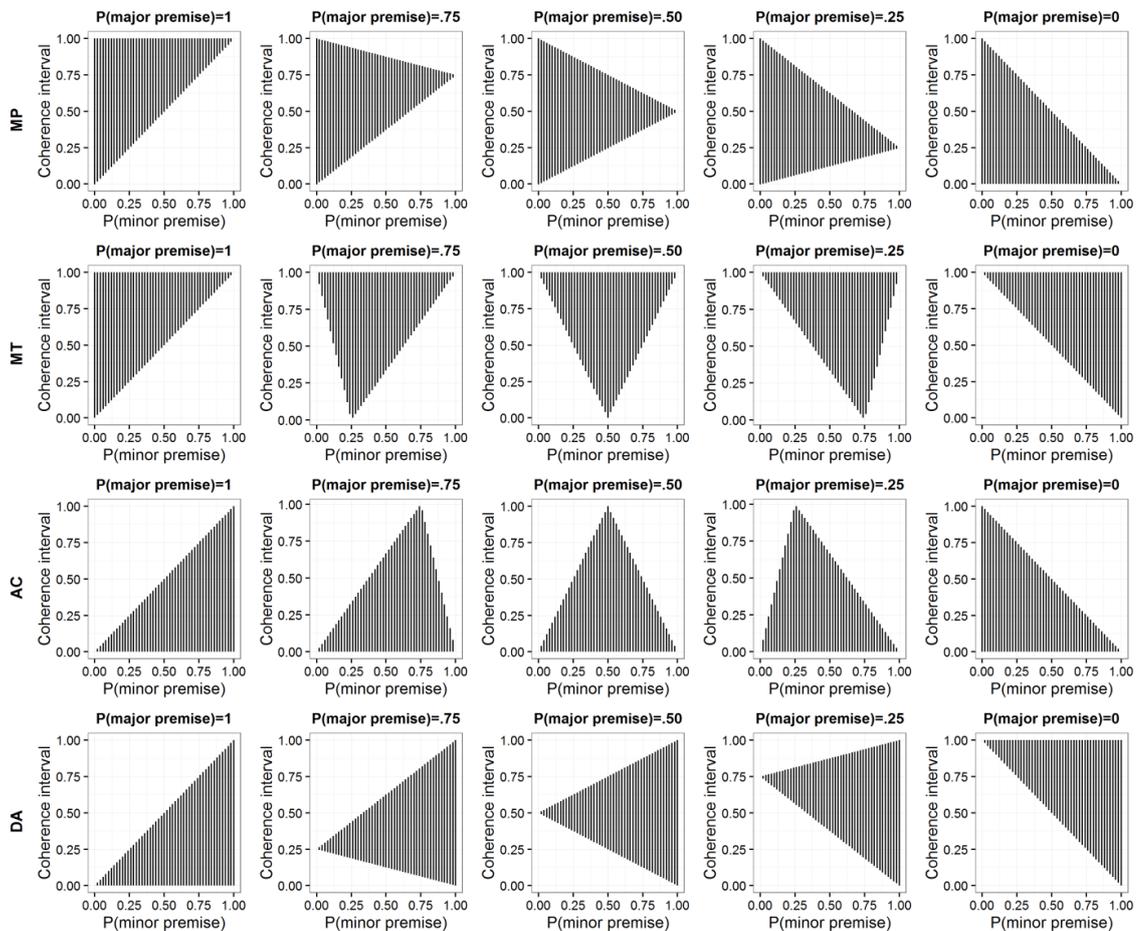


Figure 5.5. Coherence intervals for the four conditional syllogisms: MP, MT, AC, and DA, as a function of their premise probabilities. The shaded areas in the graphs represent the coherence intervals.

The more complex coherence intervals for the inferences of type (C) are displayed in Figure 5.5 as a function of their premise probabilities. One can see that for the p-valid MP and MT inferences, the coherence intervals are narrow – and so the sensitivity of a test for above-chance coherence higher – when both premises have a high probability. In contrast, the coherence intervals for MP and MT are wide – and so the sensitivity of a test for above-chance

coherence lower – when the major premise has a high probability and the minor premise a low probability. The opposite relation holds for the p-invalid inferences AC and DA: The coherence intervals for the latter two inferences are wide when both premise probabilities are high, whereas they are narrow when the major premise has a high probability and the minor premise a low probability.

Accordingly, premise probability for the two-premise inferences was varied over two conditions, *high-high*, and *high-low* (with the first term in each pair referring to the probability of the major premise, and the second to the probability of the minor premise). The purpose of this variation in premise probability was again to assess the degree to which responses were coherent above chance levels in a way that increases the sensitivity of the test, i.e. the chances of finding above-chance coherence when it is there.

Above-chance coherence was calculated following the procedure used in previous studies (Cruz et al., 2015; Evans et al., 2015, Singmann et al., 2014): A conclusion probability judgment was classified as coherent if it fell within the coherence interval, and as incoherent if it fell outside of it. Coherent judgments were coded with 1 and incoherent judgments with 0. These values for *observed coherence*, 1 or 0, were compared to the *chance rate* of a coherent response, which was defined as being equal to the width of the coherence interval. *Above-chance* coherence was computed by subtracting, for each response, the chance rate from the observed value for coherence. For example, if the chance rate of being coherent was 0.25 in a particular case, and the response was coherent, $1 - 0.25 = 0.75$ would be the measure of above-chance coherence.

The variation in premise probabilities made it possible to assess not only whether people were sensitive (at some level) to the location of the coherence interval, and therefore coherent above chance levels, but also whether they were sensitive to the width of the interval. This is because when the interval is wide, one can expect the variability of responses between conditions to be higher. For inferences of type A (whose coherence interval was a point value), response variance was predicted to be similar for high and low premise probabilities. For inferences of type B, response variance was predicted to be higher for valid inferences when premise probability is low, and higher for invalid inferences when premise probability is high. Similarly, for inferences of type C a higher response variance for valid inferences was predicted in the high-low condition, and a higher response variance for invalid inferences in the high-high condition.

The above variables were entered into a mixed design with group (statements (1), inferences (2), inferences with working memory load (3)) as between participants variable, and inference (the 12 inferences of Table 5.4) and probability (*high-high*, *high-low* for the two-premise inferences; *high*, *low* for one-premise inferences) as within participant variables. The dependent variables were the mean and variance of participants' conclusion probability judgments, and above-chance coherence computed from these judgments.

Materials. The premise probabilities were varied using real-world materials, as in Evans et al. (2015) and Singmann et al. (2014). This means that it was not possible to have strict control over the value of the premise probabilities, because real world experiences will be different for each person, and people's probability judgments may be calibrated differently. Nonetheless, an attempt was made to create trends for higher or lower probability ratings strong enough to show an effect.

The high and low probability contents were constructed with a view to avoiding the use of stereotypes about groups of people. They instead referred to objects, animals and services considered typical or untypical for a random, unnamed city. Below is a sample trial for the inferences group (Group 2), showing a MP inference in the *high-high* and the *high-low* condition. The full set of materials is provided in Appendix C.

Berta walks through a random city and goes into a bar.		
How likely are the following statements?		
(a) high-high	0% likely	100% likely
Premise 1:	certainly false	certainly true
If the bar has beer then it has wine.	(slider for Premise 1)	
Premise 2:		
The bar has beer.	(slider for Premise 2)	
Conclusion:		
Therefore, the bar has wine.	(slider for Conclusion)	

(b) high-low		
Premise 1:		
If the bar has a book store, then it has a place to read.	(slider for Premise 1)	
Premise 2:		
The bar has a book store.	(slider for Premise 2)	
Conclusion:		
Therefore, the bar has a place to read.	(slider for Conclusion)	

Suppose that in the above example we indeed assign a high probability to both premises of inference (a), and assign a high probability to the first premise but a low probability to the second premise of inference (b). Then the coherence constraints of the inferences would demand that we assign a high conclusion probability to inference (a), but would allow us to assign a wide range of conclusion probabilities to inference (b).

Note that it is not possible to cross inference with premise probability while holding the type of material constant. For example, a *high-high* condition for MP, as shown above, will be a *high-low* condition for DA; and a *high-high* condition for MT will be a *high-low* condition for AC. However, if the same material is used for the *high-high* condition of MP as for the *high-low* condition of DA, and analogously for the other cases (pairing the materials of the *high-high* condition for MT with those of the *high-low* condition for AC, those of the *high-low* condition of MP with those of the *high-high* condition for DA, and those of the *high-low* condition for MT with those of the *high-high* condition for AC), then overall the materials for the p-valid inferences will be the same as those for the p-invalid inferences. This was the case in the present experiment. It meant that the same materials expected to lead to higher and less heterogeneous responses for MP were expected to lead to lower and less heterogeneous responses for DA; and the same materials that were expected to lead to more heterogeneous responses for MP were expected to lead to more heterogeneous responses for DA (and analogously for MT and AC). There is a risk that a confirmation of this prediction could be an artefact of the materials, but this risk can be reduced by using a wide range of materials. The present experiment used 12 different materials for each inference, referring to: a dog, restaurant, family house, tree, dove, bar, neighbourhood, bus from the local public transport system, cat, park, bicycle, and train station.

As can be seen in the above example of the experimental materials, the probability conditions were varied within the topics. As a result, each inference was randomly allocated to a single scenario for each participant, and this mapping remained constant within participants for the different probability conditions.

For the two-premise inferences, the variation in the topics was the same for the *high-high* condition of MP as for the *high-low* condition of DA; it was the same for the *high-high* condition of MT as for the *high-low* condition of AC, and analogously for the other conditions. The topic variations used for the one-premise inferences were the same for each one-premise inference. But the topic variations differed between the two-premise and the one-premise inferences. The reason was that some materials that seemed natural in a given condition of the two-premise inferences seemed pragmatically odd in the same condition for the one-premise inferences, and vice versa. In particular, and interestingly, the conditionals used for the conditional syllogisms featured a "connection" between the antecedent p and the consequent q . It was mostly a causal connection, but sometimes also a conceptual or similarity based connection, as in the example about the bar. The reason for this was that it seemed impossible to create materials for the *high-low* condition without such a connection: if the probability of the antecedent and/or of the consequent is low, it seems implausible that the probability of the conditional itself will be high unless the conditional describes a general relation between antecedent and consequent that comes into play whenever the antecedent and/or consequent hold, even if they may not hold at that particular moment. In contrast, for

the conditionals in the one-premise inferences, which stood in relation to disjunctions, the focus was on assuring that the antecedent p was not a subset or superset of q , but was defined at the same level of generality – a criterion that often failed to hold when a connection was present. These considerations were implemented with the aim of avoiding possible pragmatic infelicities in the materials (for more on the role of a "connection" in the understanding and use of conditionals and disjunctions, see Cruz et al., 2016; Douven, 2015b; Oberauer, Weidenfeld, et al, 2007; Oberauer & Wilhelm, 2000; Skovgaard-Olsen et al., 2016; Vidal & Baratgin, 2017).

With 8 one-premise inferences, 4 two-premise inferences, and 2 premise probability conditions, the experiment for the inferences groups had $12*2 = 24$ trials, and that for the statements group had $(8*2 + 4*3) *2 = 56$ trials. In addition, for all the groups, the experiment included two catch trials to make sure participants were paying attention. The catch trials were similar in format to the experimental trials, but the text displaying the statements to be evaluated was replaced with the information that it was a control trial to make sure participants were paying attention, and asking them not to respond with a probability judgment, but to instead click *next* to continue with the experiment.

Procedure. The experiment took place in a quiet testing room of the Department of Psychological Sciences of Birkbeck, University of London. In case there were any questions, the experimenter was present while the participants went through the instructions and three practice trials with inferences different from those tested in the experiment. At the end of the experiment participants provided demographical information and indicated whether they had taken part seriously, or just "clicked through". The final page provided debriefing information. The entire session took approximately 20 minutes to complete.

Results and Discussion

Preliminaries: The problem of quantitative comparability. The data analysis began with an assessment of some preliminary matters to determine which questions can and cannot be addressed in the present experiment.

To check whether the manipulation of premise probability using real world materials was successful, the mean values assigned to the premises were computed for each inference and group. These values, together with those of the conclusion, are shown in Figures 5.6 to 5.8.

One can see that overall, the manipulation worked in that participants tended to give premise probability ratings at or above .75 in the *high* (resp. *high-high*) condition and at or below .25 in the *low* (resp. *high-low*) condition. An exception was the ifOr inference, *if not-p then q, therefore p or q*, for which mean premise probability judgments in the *high* condition were below 50%. Note that this difference was observed across the three groups, and across 12 scenarios shared with the other one-premise inferences. A possible explanation for it is that the negation in the antecedent of the conditional premise led to a failure of invariance. An

example of this conditional in the *high* condition was "If the bar does not have beer, then it has wine". This conditional was expected to have a high probability because the consequent has a high probability. But this assumption might have been wrong: It did not take into account that a situation in which a bar in some random city in the world has no beer is so rare, that it might be exceptional in ways that could also have affected the supply of wine. For example, there might be no beer because of a severe flood or other disaster, cutting the city off from its supply chain. In such a situation the supply of wine is also likely to be affected. This problem would have been avoided if the high condition had been one in which not only the consequent, but also the antecedent had a high probability, for example: "If the bar has no book store, then it has wine". However, a consequence of this would be that the disjunction in the conclusion, "The bar has a book store or it has wine" would be justified only by one of its disjuncts, i.e. it would be constructively justified (Gilio & Over, 2012). Such a disjunction may still have a high probability, but its constructive justification could bring with it complications of its own (e. g., it is pragmatically infelicitous to assert a disjunction only on the basis of a belief in one of the disjuncts, because it is then more informative to assert the disjunct directly).

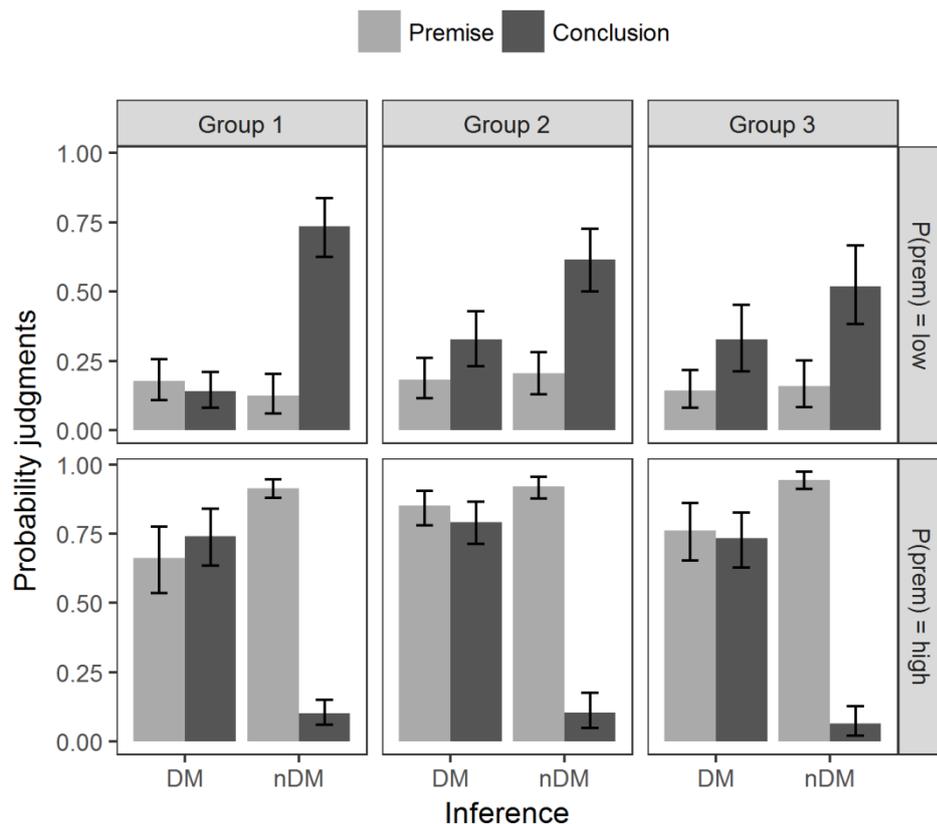


Figure 5.6. Premise and conclusion probabilities in Experiment 3 for the inferences of type A, separately for each group and premise probability condition. Error bars show 95% CIs.

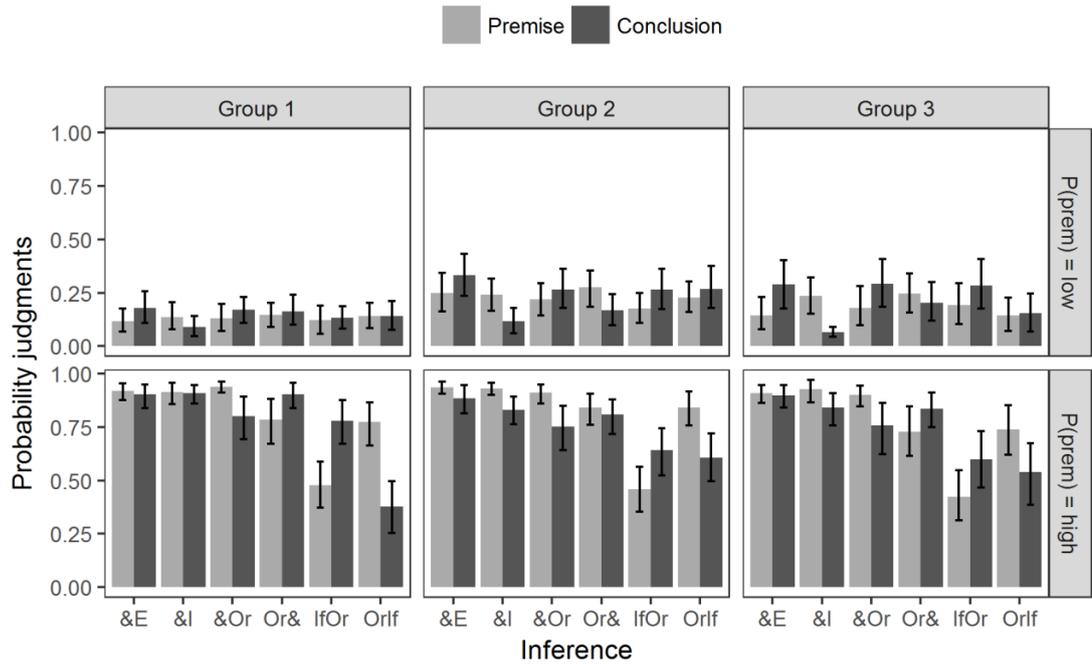


Figure 5.7. Premise and conclusion probabilities in Experiment 3 for the inferences of type B, separately for each group and premise probability condition. Error bars show 95% CIs.

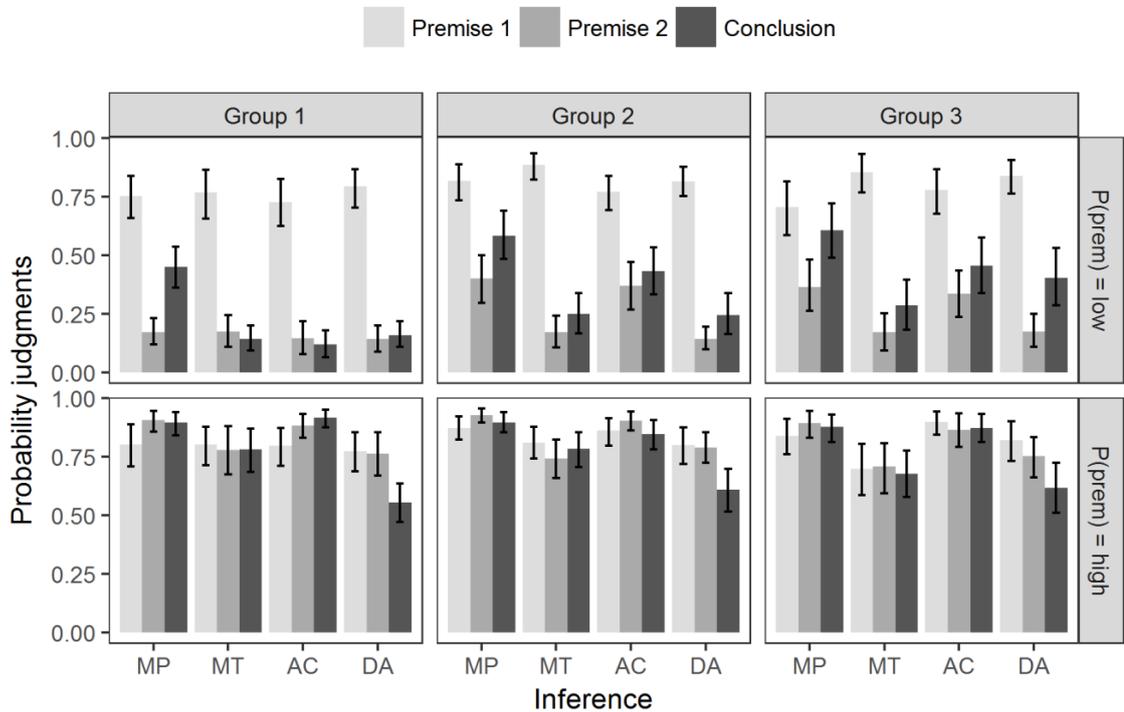


Figure 5.8 Premise and conclusion probabilities in Experiment 3 for the inferences of type C, separately for each group and premise probability condition. Error bars show 95% CIs.

There might be a conflict regarding which component probabilities lead to high probability judgments for both premise and conclusion. This might help explain past findings of lower endorsement rates for this inference than for other valid inferences when using a binary or ternary response format and binary paradigm instructions (Espino & Byrne, 2013; Oberauer, Geiger, & Fischer, 2010). It would be interesting to investigate this explanation further in a subsequent experiment. But however it is interpreted, the lower probability assigned to the IfOr inference has consequences for the computation of coherence. Figure 5.9 zooms in to the probability judgments for the IfOr inference and compares them to those of its converse, the OrIf inference p or q , therefore if not- p then q , showing the coherence values derived from these probabilities for each inference.

Figure 5.9 shows that responses to both inferences tended to be coherent: for the valid IfOr inference, the conclusion was rated more probable than the premise, and for the invalid OrIf inference the conclusion was rated less probable than the premise. The degree to which probability judgments for the premise differed from those for the conclusion was similar for both inferences. Accordingly, the observed coherence rate for both inferences is almost identical. However, the lower ratings of premise probability for the IfOr inference imply a higher chance-rate of conforming to coherence for this inference, which translates into a lower rate of above-chance coherence - despite, as already mentioned, almost identical observed coherence.

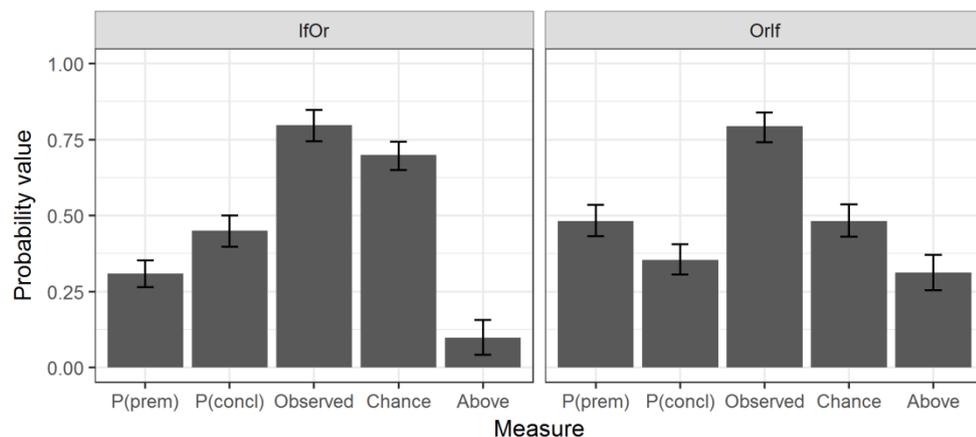


Figure 5.9. Coherence information for the IfOr inference and the OrIf inference of Experiment 3. "P(prem)" = premise probability, "P(concl)" = conclusion probability, "Observed" = observed coherence, "Chance" = chance-rate coherence, and "Above" = above-chance coherence. Error bars show 95% CIs.

Is the lower rate of above-chance coherence observed for the IfOr inference justified, or is it an artefact of the differences in the chance rates for the two inferences? In this case one could argue that it was actually justified, because the lower chance rate for the OrIf inference meant that it required more "effort" to arrive at the same value of observed coherence than was

the case for the IfOr inference. After all, an observed rate of 80% can more readily be considered an achievement when the chance rate is 20% than when the chance rate is 79%. On the other hand, an observed rate of around 80% is already quite high, knowing that 100% is the rate of a hypothetical perfectly coherent person. It could be that participants' responses are already reaching their ceiling, and that much higher coherence is not possible due to measurement imprecision. In that case the lower rate of above-chance coherence for the IfOr inference would indeed be unjustified. The difference in performance between the two inferences seems hard to compare given their different chance rates, and a comparison would have been even more difficult if the observed rates had differed more strongly.

Figures 5.5 to 5.8 suggest that the problem in comparability was also present within inferences because judgments of premise probability differed between groups. This was confirmed by a linear mixed model for the effects of group, premise probability, and inference type on judgments of the probability of the first premise, with random intercepts for participants and scenarios: Judgments of the probability of premise 1 were higher in the inferences group ($EMM = .631$) than in the statements group ($EMM = .571$), $F(1, 1103.54) = 22.09, p < .001$.

For the inferences of type C, a linear mixed model for the effects of group and premise probability on judgments of the probability of the second premise (again with random intercepts for participants and items) showed the same result: probability judgments were higher in the inferences group than in the statements group, $F(1, 368) = 8.89, p = .003$, particularly in the low premise probability condition, where there was more space for an increase, $F(1, 368) = 8.27, p = .004$ (for the low condition: $EMM_{group1} = .158$; $EMM_{group2} = .271$. For the high condition: $EMM_{group1} = .833$; $EMM_{group2} = .841$).

These results show that premise probability judgments differed between groups. This implies that the chance rates will differ between groups, making a quantitative comparison of coherence between groups difficult.

The problem of quantitative comparability of coherence results between inferences and task conditions caused by unequal chance rates has not been considered in past studies investigating coherence statistically (Cruz et al., 2015; Evans et al., 2015; Politzer & Baratgin, 2016; Singmann et al., 2014). There may be some way of rescaling the coherence rate to account both for the presence of the chance rate, and for differences in the chance rate between conditions. If such an adjustment were possible, it would be of great value to the field because it would much increase the type of questions that could be answered with designs like those used in this experiment.

A solution to this problem cannot be reached by computing the proportion of above-chance coherence out of the maximum possible above-chance coherence, because this is equivalent to the observed coherence rate. To see why, consider that:

$$\text{above-chance coherence} = \text{observed coherence} - \text{chance rate coherence}$$

maximum possible above-chance coherence = 1 - chance rate coherence

One cannot obtain the proportion of above-chance coherence out of the maximum possible above-chance coherence by dividing the first by the second, because the division is done on an item-by-item basis, and can therefore lead to impossible outcomes. Consider the example of a response that did not fall in the interval (observed coherence = 0), where the chance rate was .8. Above-chance coherence for this item will then be $0 - .8 = -.8$. Given that the chance rate was .8, $1 - \text{the chance rate} = .2$, and so the proportion of above-chance coherence out of the maximum possible above-chance coherence for this item would then be $-.8/.2 = -4$. This is not a proportion and it makes no conceptual sense.

An alternative way of arriving at the proportion of above-chance coherence out of maximum above-chance coherence is by rescaling the two: if we set maximum above-chance coherence to 1 by adding its complement (i.e., by adding the chance rate) to it, and then also add the chance rate to above-chance coherence, we arrive at the proportion we were looking for without the above problem of nonsensical values. But this is simply the observed coherence rate.

A further ineffective solution, which is nonetheless of theoretical interest, is the following. The problem we are facing is that we want to compare coherence values corrected for chance-rate coherence, but where the chance rate differs between responses. This problem is analogous to that of computing a within subject standard error, where the variance differs between participants (Bakeman & McArthur, 1996, p. 587). The idea is to partial out the variance that is due to interindividual differences from the overall response scores, and compute the standard error on the resulting adjusted scores. Formally:

$$W_{ij} = Y_{ij} - (Y_i - Y_{..})$$

where W_{ij} and Y_{ij} are the adjusted and raw scores, respectively, Y_i is the mean of participant i across repeated measures, and $Y_{..}$ is the grand mean. In analogy to this, one can first subtract the chance rates for each response from the overall mean of the chance rate. The resulting difference scores can then be subtracted from the observed coherence scores to obtain *adjusted observed coherence* scores. The problem is that these adjusted scores are identical to the scores for above-chance coherence obtained through the Evans et al. procedure. The equivalence between the two methods of adjusting for the chance rate is interesting and gives the method of Evans et al. a wider foundation. But it does not solve the problem of quantitative comparability of coherence between conditions.

Without a solution to this problem, experimental designs like that of Cruz et al. (2015), Evans et al. (2015), Politzer & Baratgin (2016), Singmann et al. (2014) and the present experiment seem to be able to make only nominal comparisons between inferences and conditions. That is, assessments of whether or not responses were coherent at above-chance

levels for a series of inferences or conditions, but not whether above-chance performance was higher for one inference or condition than for another.

Nevertheless, this limitation does not undermine the conclusions drawn by Evans et al. (2015) on the effect of an explicit inference task. This is because, although they measured a quantitative increase in above-chance coherence in the inferences task compared to the statements task, and this quantitative change is hard to interpret, they also found a qualitative increase. Coherence was above-chance for two (MP, DA) out of four inferences in the statements task, and for three (MP, DA, AC) out of four inferences in the inferences task. At the same time, the finding in Evans et al. of above-chance coherence for AC in the inferences task is relativized by the finding of no above-chance coherence for this inference in the inferences task of Singmann et al. (2014).

Against this background, it also makes sense that Cruz et al. found no effect of an explicit inference task in Experiments 1 and 2. All they seem to be able to assert with their design is whether or not coherence was above-chance levels for the inferences they investigated, and they found this to be the case for all inferences. Similarly, the limitations in quantitative comparability do not undermine the qualitative findings in Cruz et al. and Politzer et al. that people were reliably coherent above chance levels for a range of one-premise inferences.

The present experiment investigates above-chance coherence in a qualitative form further, and complements this with a method of analysis not based on above-chance coherence.

Above-chance coherence. The results of the coherence analysis for each inference and group are displayed in Figure 5.10. The left panels of the figure show the rate of observed coherence. The horizontal line crossing zero in these panels stands for the rate that would be observed if a person were never coherent, and the maximum of the scale at 1 stands for the rate that would be observed if a person were always coherent. The right panels show the rate of above-chance coherence. The horizontal line crossing the y axis at 0 in these panels represents the likelihood of a response being coherent just by chance. The rows of the Figure display the results for each group. For simplicity, it may be useful to focus first on the middle row: that displaying the results for the inferences group.

Both observed and above-chance coherence were measured with three different degrees of precision: a measurement based on the exact boundaries of the coherence interval, one based on the interval boundaries $\pm 5\%$, and one based on the interval boundaries $\pm 10\%$. These degrees of measurement precision do not entail that the coherence assessment was more lenient in the $\pm 5\%$ and $\pm 10\%$ conditions than in the exact condition. The reason is that the two conditions with lower measurement precision were computed by enlarging not only the width of the coherence interval by 5% resp. 10%, with the trivial result that more responses will lie within it, but by also enlarging the chance rate of a coherent response by the same margin. This means that the rate of observed coherence will trivially be higher in the $\pm 5\%$ and $\pm 10\%$ conditions than in the exact condition, but the rate of above-chance coherence will

only be higher as well if the additional number of coherent responses resulting from a larger interval was higher than expected by chance.

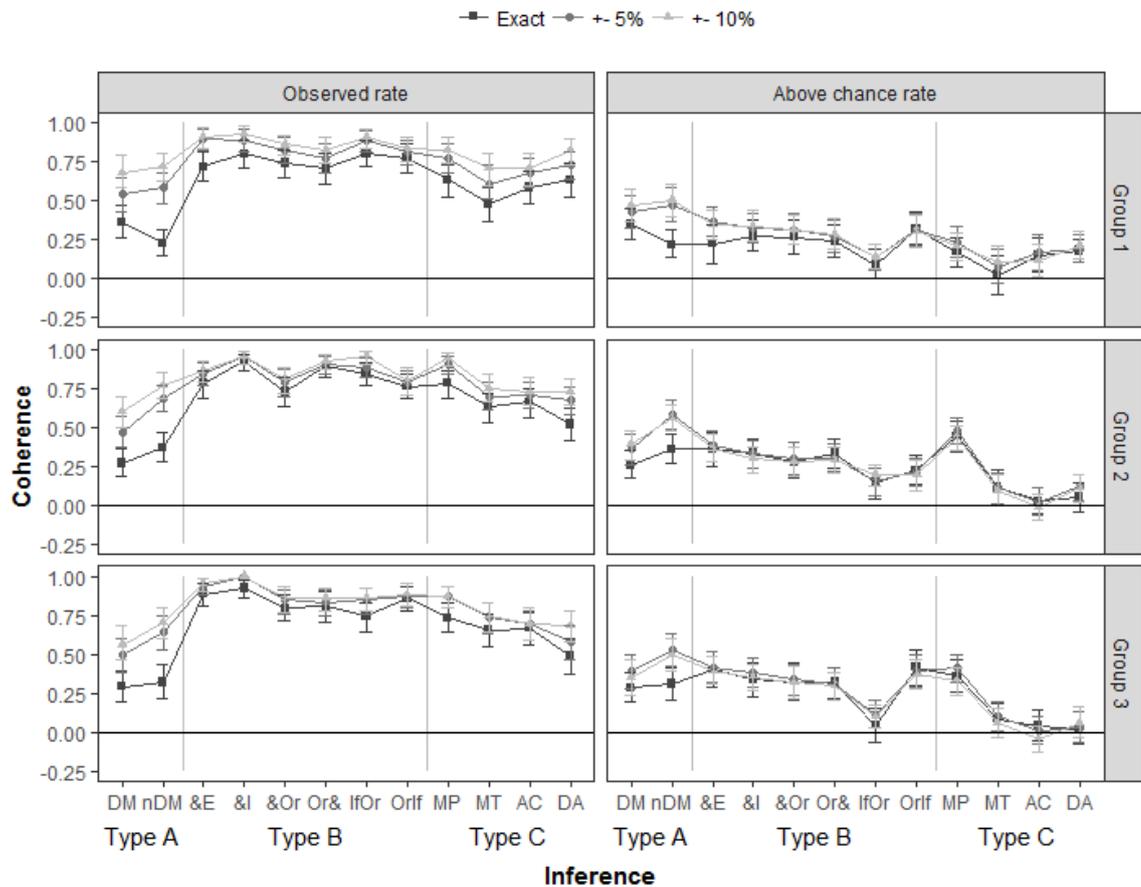


Figure 5.10. Mean values of observed and above-chance coherence for the 12 inferences of Experiment 3, separately for each group and for three levels of measurement precision (see text for details). The black horizontal line represents a coherence rate of 0% in the panels for observed coherence, and it represents the chance rate of a coherent response in the panels for above-chance coherence. Error bars show 95% CIs.

One can see this clearly in Figure 5.10: in the left panels the observed coherence rate is higher in the conditions of lower measurement precision for all inferences, whereas in the right panels the above-chance coherence rate is higher mainly for inferences 1 and 2 (DM and nDM). This makes sense given that for DM and nDM the coherence interval is a point value: It therefore seems likely that we will find a cluster of responses near that point, indicating a sensitivity to coherence, but which do not lie exactly on the point because people's degrees of belief do not have the degree of precision of a point probability. For example, it seems unrealistic that people will think about the likelihood of rain on a particular day as being exactly 79%, such that they would consider it false for it to be 78% or 80%.

The three degrees of measurement precision can be viewed as scales with different numbers of subdivisions. The exact scale has 101 subdivisions, going from 0% to 100%; the $\pm 5\%$ scale has 10 subdivisions, and the $\pm 10\%$ scale has 5 subdivisions.

In order to increase comparability to previous studies (Evans et al., 2015; Singmann et al., 2014), subsequent statistical analyses will be based on exact coherence unless otherwise specified. However, the above measurement of probability judgments for different degrees of precision can be used to test empirically hypotheses about the granularity or coarseness of people's degrees of belief. Such an empirical approach using the tools of probability theory seems preferable to making a priori assumptions about the coarseness of people's beliefs, advocated for instance in qualitative or ranking theoretical approaches (e. g. Spohn, 2013).

To obtain an impression of the statistical pattern in the data relative to that reported in previous studies (Evans et al., 2015), a linear mixed model was computed for the effects of inference type and group on above-chance coherence, with a random intercept for participants. The model intercept was significant, $F(1, 170.39) = 397.34, p < .001$, indicating that overall responses were coherent around 24% more often than expected by chance ($EMM = .237$).

Above-chance coherence differed between inference types, $F(2, 3012.068) = 28.47, p < .001$. It was higher for inferences of type A ($EMM = .296$) and B ($EMM = .274$) than for those of type C ($EMM = .139$) (for A vs. C: $F(1, 1441) = 37.744, p < .001$. For B vs. C: $F(1, 2488.125) = 44.60, p < .001$). Above-chance coherence was similar between the inferences of types A and B ($F(1, 1964.099) = .770, p = .380$).

There was no main effect of group ($EMM_{\text{group1}} = .215, EMM_{\text{group2}} = .250, EMM_{\text{group3}} = .245; F(2, 170.391) = .85, p = .430$), nor an interaction between group and inference type, $F(4, 3012.069) = .71, p = .588$.

The above results are only indicative given the quantitative comparability problem described above. It is therefore useful to complement them with a binary assessment, for each inference, of whether coherence was above chance levels or not. The confidence intervals in Figure 5.10 indicate that this was clearly the case for all inferences except for the inference of IfOr in groups 1 and 3, as well as MT, AC and DA across groups. The results for these four inferences were examined in more detail in what follows.

The group and inference specific analyses included only fixed effects, because attempts to include random effects led to failure of convergence in some cases. That a model for a single inference fails to converge when including random effects, whereas the larger model including all inferences did not, is understandable given that the Maximum Likelihood method used to compute the mixed models is approximate (in contrast to the exact ordinary least squares method of the ANOVA), and so requires larger sample sizes to be effective.

For IfOr, coherence was not above chance levels in Group 1, $F(1, 84) = 3.90, p = .052$; or in Group 3, $F(1, 78) = .64, p = .425$. But coherence was above chance levels in Group 2, $F(1, 100) = 8.38, p = .005$. The same result was obtained for MT: coherence was not above chance

levels in Group 1, $F(1, 84) = .16, p = .690$; or in Group 3, $F(1, 78) = 2.43, p = .123$. But coherence was above-chance levels in Group 2, $F(1, 100) = 4.24, p = .042$.

Thus, for the two p-valid inferences IfOr and MT, coherence was above chance levels in the inferences task but not in the statements task. Further, the presence of an explicit inference task was only associated with significantly higher coherence when participants' attention to the task was not diverted by additional working memory load.

This finding contrasts with that for the p-invalid inferences AC and DA. For AC, responses were coherent above chance levels in Group 1, $F(1, 84) = 7.18, p = .009$. But coherence was not above-chance in Group 2, $F(1, 100) = .49, p = .486$; or in Group 3, $F(1, 78) = .75, p = .390$. The same finding was obtained for DA: responses were coherent above-chance levels in Group 1, $F(1, 84) = 21.10, p < .001$; but not in Group 2, $F(1, 100) = 1.32, p = .253$; or in Group 3, $F(1, 78) = .22, p = .644$.

Thus, for the p-invalid inferences AC and DA, coherence was above-chance levels in the statements task, but ceased to be so in the inferences task. Going back to Figure 5.8, one can see that this was because participants gave higher conclusion probability judgments in the inferences task than in the statements task for all four conditional syllogisms (see Klauer et al., 2010, for a similar finding). Higher conclusion probability increases coherence for the p-valid inferences, but decreases coherence for the p-invalid inferences.

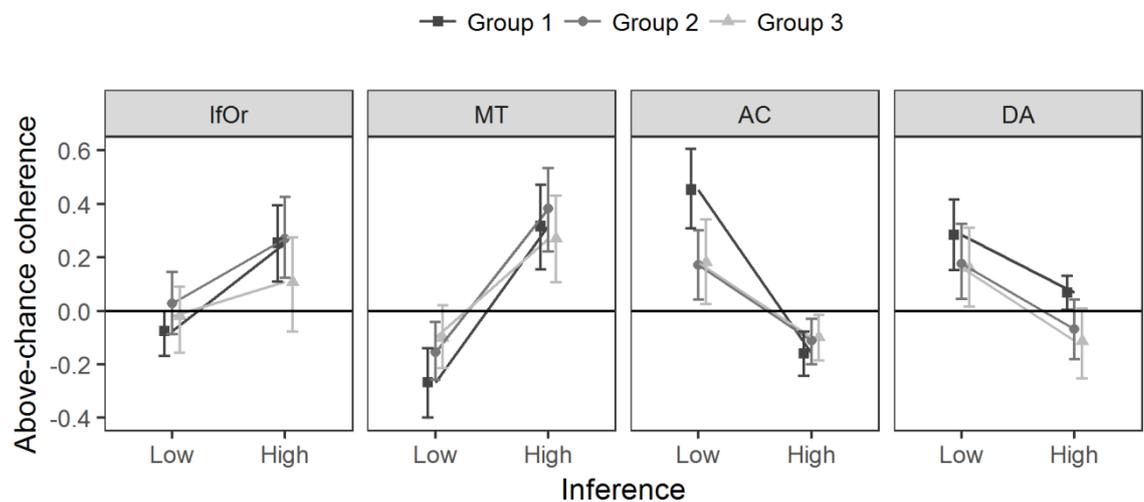


Figure 5.11. Above-chance coherence for the inferences IfOr, MT, AC and DA of Experiment 3, separately for each group and premise probability condition. The horizontal line in the panels represents the chance rate of a coherent response. Error bars show 95% CIs.

Overall, above-chance coherence for IfOr, MT, AC, and DA was not reliably present, but depended on the group, specifically on whether participants received a statements or an inferences task. However, Figure 5.11 shows that responses to these inferences nonetheless tended to be coherent above chance levels in the premise probability conditions in which it

was easier to detect above-chance coherence when it was there. This was the *high* condition for the two p-valid inferences IfOr and MT, and the *low* condition for the p-invalid inferences AC and DA. There was thus a tendency to be coherent across groups also for IfOr, MT, AC and DA, though this tendency was weaker than for the other inferences.

Taken together, responses were coherent above-chance levels across the three groups for 8 out of 12 inferences. Among the 4 remaining inferences, coherence for the two p-valid inferences was above-chance levels in group 2, that is, in the inferences group without working memory load. This suggests that overall people were sensitive to the constraint of coherence even in the absence of an explicit task to draw inferences. Where sensitivity to coherence was less strong, it helped to have an explicit inference task for the valid inferences IfOr and MT, but not for the invalid inferences AC and DA.

The effect of group for inferences IfOr and MT suggest that one of the reasons why coherence was higher in the presence of an explicit inference task was the opportunity to focus attention on the relationships between the statements presented on the screen. Responses were less coherent for inferences when the statements required to be coherent were scattered across trials, or when it was more difficult to focus attention on the inferences because of concurrent working memory load.

Why was coherence found to be less strong for IfOr, MT, AC, and DA, and why did coherence actually decrease with an explicit inference task for AC and DA? In the case of IfOr, this is likely due to the lower probability assigned to the premise, which rendered the test for above-chance coherence less sensitive for this inference. The lower premise probability assignments may be due to a failure of invariance, but this possibility would have to be investigated further in follow-up experiments.

For MT, one way of explaining the lower rate of above-chance coherence is again as the result of a failure of invariance (Oaksford & Chater, 2013). The structure of MT can be compared to a *reductio ad absurdum* argument: a high probability of the minor premise *not-q* is incompatible with a high probability for both the conditional *if p then q* and its antecedent *p*. To maintain coherence after learning the minor premise, we therefore either have to assign a higher probability to the conclusion *not-p*, or a lower probability to the major premise *if p then q*. But the inference itself does not tell us which of these two probabilities to change. The computation of coherence presupposes that the subjective probabilities of the premises remain invariant when judging the probability of the conclusion. If invariance fails, then one can no longer capture the coherence of a response with respect to the original probabilities, and so cannot judge whether the response was coherent or not. One can only establish the new premise probabilities and judge the coherence of the response with respect to them. The extent to which a failure of invariance can explain lower coherence rates for MT would be important to investigate in further experiments.

Coherence for AC and DA. Consider now the results for the invalid inferences AC and DA. The finding that responses to these inferences were coherent at above-chance levels in the statements task, but ceased to be so in the inferences task, raises the possibility that the participants spontaneously held coherent beliefs about the likelihoods of the statements in the inferences, but then interpreted these statements differently in the explicit inferences task, and so not in line with the experimenter's assumptions when measuring coherence. Two alternative interpretations are considered in the following.

The first possible interpretation is that the naturalistic materials, plus the context of an explicit inference, caused some participants to infer that the antecedent and consequent of the conditional premise covaried, so that both $P(q|p)$ and $P(p|q)$ should be taken as high. Supposing this happens, a conjunction of the probability conditionals, (*if p then q*) & (*if q then p*), can be formed (Gilio et al., 2016; Gilio & Sanfilippo, 2014), and Table 5.6 constructed as its Jeffrey table (Jeffrey, 1991; see also Appendix A). There is some evidence that people sometimes form such a biconditional interpretation of conditionals (e.g. Baratgin et al., 2013; Barrouillet & Gauffroy, 2015; Fugard, Pfeifer, et al., 2011).

Let us call the versions of AC and DA for which the major premise is this probability biconditional, "biconditional AC" and "biconditional DA", respectively. Table 5.7 shows the equations for the coherence intervals for these inferences⁷, and Figure 5.12 the location of the intervals for biconditional AC and biconditional DA as a function of their premise probabilities. Note that in contrast to the original AC and DA, biconditional AC and DA are p-valid inferences. One can see that the intervals for the original and for the biconditional versions of the inferences only coincide when $P(\text{major premise}) = 0$, and that the intervals for the biconditional versions are generally stricter, that is, narrower than those of the original inferences. Further, when the probability of the major premise is 1, the intervals for both biconditional inferences are point values equal to the probability of the minor premise.

Table 5.6. The Jeffrey table for the probability biconditional that results from adding the converse, *if q then p*, to the original conditional.

p, q	T
p, not-q	F
not-p, q	F
not-p, not-q	$P(p \ \& \ q) / P(p \ \text{or} \ q)$

Note. T = True, F = False.

⁷ These intervals were derived by the author on the basis of example cases of minimum and maximum coherence bounds, computed using the software package "Check Coherence" by Andrea Capotorti and colleagues (Capotorti, Galli, & Vantaggi, 2003).

Table 5.7. The coherence intervals for biconditional AC (Bic AC) and biconditional DA (Bic DA).

Name	Form	Coherence interval for the conclusion	
		Lower bound	Upper bound
Bic AC	$(if\ p\ then\ q) \& (if\ q\ then\ p),\ q \therefore p$	xy	if $x \geq y$, then y/x if $x < y$, then $x + (1 - y)$
Bic DA	$(if\ p\ then\ q) \& (if\ q\ then\ not-p),\ not-q \therefore not-p$	if $x + y \leq 1$, then $1 - (x + y)$ if $x + y > 1$, then $(x + y - 1)/x$	$1 - x(1 - y)$

Note. \therefore = "therefore", x = the probability of the major premise, y = the probability of the minor premise.

A second interpretation considered is based on the idea that people could use background knowledge or pragmatics to infer the converse of the conditional, *if q then p*, from a given conditional, *if p then q*. Supposing this occurs, people could then use the converse alone for inferences, leaving the original conditional premise aside as inessential (see Adams, 1998, on the degree of essentialness of a premise for a conclusion). In the case of AC, the resulting inference is a version of MP: *if q then p, q, therefore p*. In the case of DA, the resulting inference is a version of MT: *if q then p, not-p, therefore not-q*. The degree of belief people assigned to the converse could vary, but if we assume for simplicity that it is equal to the probability participants assigned to the original conditional premise, then the procedure described is equivalent to swapping the conditional with its converse, or possibly simply misinterpreting the conditional as the converse. There is evidence from past studies suggesting that people sometimes find it difficult to distinguish between a conditional and its converse, e. g. through the conjunctive response in truth table tasks, the treatment of "all A are B" as equal to "All B are A" in reasoning with categorical syllogisms, and the conflation of $P(q|p)$ with $P(p|q)$ in the judgment and decision making literature on the use of Bayes' theorem (Evans & Over, 1996, 2004). Let us call "converse AC" and "converse DA" the versions of AC and DA in which the conditional premise is substituted with its converse.

Figure 5.13 compares the above-chance coherence participants would obtain for an interpretation of the conditional premise in AC and DA as the original probability conditional, as the probability biconditional, and as the converse of the original conditional, respectively. One can clearly see from the confidence intervals in the figure that an interpretation of the conditional premise as a biconditional renders participants' responses to both inferences coherent above chance levels in all groups, whereas an interpretation of the conditional premise as its converse does not. This suggests that a viable possibility for why coherence was not above chance levels in the inferences task for inferences AC and DA, was that some

participants interpreted the conditional premises, in the explicit versions of these inferences, as probability biconditionals.

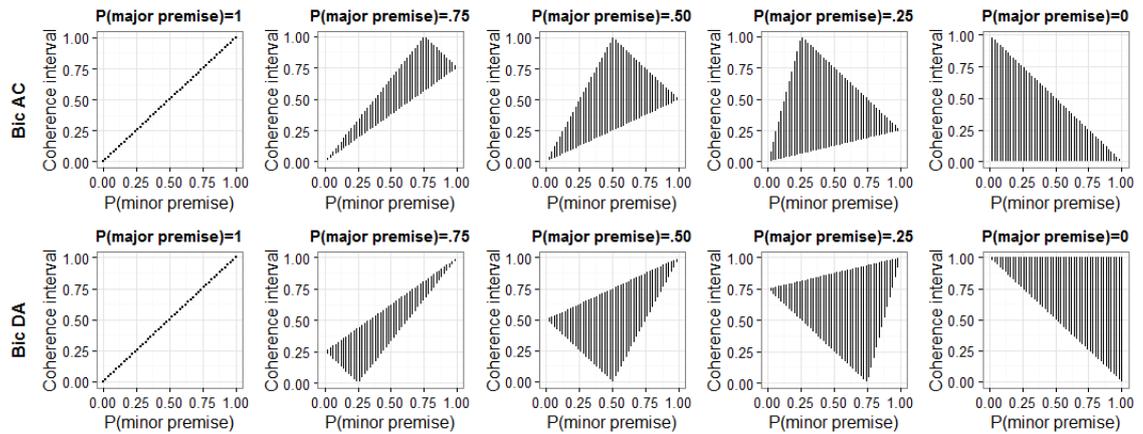


Figure 5.12. Coherence intervals for biconditional AC and biconditional DA as a function of their premise probabilities.

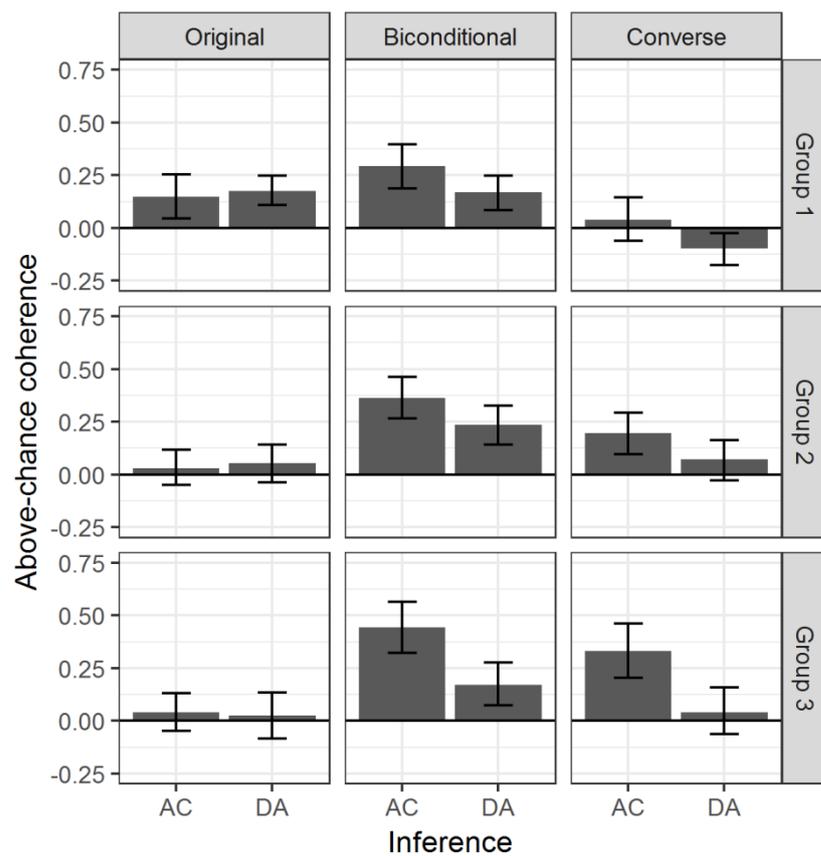


Figure 5.13. Above-chance coherence for AC and DA of Experiment 3, in the original version of the inferences (left), in a version in which the conditional premise is substituted with a biconditional (middle), and in a version in which the conditional premise is substituted with its converse, *if q then p* (right). Error bars show 95% CIs.

For example, suppose we are members of a jury, and the prosecutor makes the following statements as if we are to draw an inference from them: "If the defendant is the murderer, then the defendant's fingerprints are on the knife. The defendant's fingerprints are on the knife." It might then seem natural to us to interpret the above conditional assertion as a biconditional. After all, AC is an invalid inference, but the corresponding inference from the biconditional is valid, as noted above. The biconditional interpretation could also be related to the fact that the materials used in this experiment featured a connection, causal or conceptual, between antecedent and consequent, to allow the construction of the *high-low* premise probability condition. Future research could explore whether the existence of certain connections encourages a probability biconditional interpretation of a conditional assertion in the context of an explicit inference.

Response variance. A second way of assessing people's sensitivity to coherence is by looking not at the frequency with which responses lie within the coherence interval, but at differences in the variance of responses as a function of the width of the interval. This method takes advantage of the premise probability manipulation used. The general prediction is that the variance of responses will be higher when the coherence interval is wide than when it is narrow.

More concretely, for the inferences of type A the coherence interval was always a point value, and so response variance was predicted to be similar across premise probability conditions. For the p-valid inferences of type B, response variance was predicted to be larger when the probability of the premise was low. In contrast, for the p-invalid inferences of type B, response variance was predicted to be larger when the probability of the premise was high.

For the inferences of type C, consider the second column of Figure 5.5. The form of the intervals depicted there leads to the prediction that for the p-valid inferences MP and MT, response variance will be larger in the *high-low* than in the *high-high* condition. In contrast, for the p-invalid inferences AC and DA, response variance was predicted to be higher in the *high-high* than in the *high-low* condition.

The variances for each condition are shown in Table 5.8. The overall pattern looks rather mixed, but in the inferences group the majority of the differences went in the predicted direction. The data were analysed in a series of F tests for equality of variances for each group. For the DM inference of type A, the variance of responses differed in Group 1 ($F(41, 41) = 2.566, p = .003$; 95% CI [1.379, 4.774])⁸; and in Group 2 ($F(49, 49) = 1.836, p = .037$; 95% CI [1.036, 3.218]). But variances did not differ in Group 3 ($F(38, 38) = 1.395, p = .310$; 95% CI [.721, .266]). For the nDM inference of type A, variances differed in all groups (For Group 1:

⁸ To interpret the confidence intervals, it is useful to consider that the F test assesses whether two variances are equal, and that they are equal when their ratio is 1. For this reason, the difference between the variances is significant when the confidence interval for this difference does not include the value of 1.

$F(41, 41) = 4.774, p < .001$; 95% CI [2.566, 8.881]. For Group 2: $F(49, 49) = 3.093, p < .001$; 95% CI [1.755, 5.451]. For Group 3: $F(38, 38) = 6.658, p < .001$; 95% CI [3.491, 12.696].

Thus, the variances of responses to DM and nDM differed between the high and low probability conditions, even though the width of their coherence intervals provided no grounds for this. Rather, the variation observed is likely to be due to inductive factors.

Table 5.8. Variances of conclusion probability judgments in Experiment 3, separately for each group, inference, and premise probability condition.

Inference	Group 1 (n=42)		Group 2 (n=50)		Group 3 (n=39)	
	<i>low</i>	<i>high</i>	<i>low</i>	<i>high</i>	<i>low</i>	<i>high</i>
DM	0.050	0.128	0.145	0.079	0.147	0.105
nDM	0.123	0.026	0.169	0.055	0.206	0.031
&E	0.052	0.033	0.152	0.061	0.144	0.030
&I	0.023	0.023	0.047	0.054	0.007	0.059
&Or	0.043	0.111	0.116	0.135	0.127	0.140
Or&	0.057	0.040	0.073	0.082	0.084	0.067
IfOr	0.033	0.119	0.118	0.155	0.137	0.169
OrIf	0.058	0.151	0.128	0.171	0.087	0.203
MP	0.092	0.028	0.140	0.025	0.144	0.033
MT	0.033	0.095	0.108	0.080	0.131	0.106
AC	0.039	0.016	0.143	0.055	0.152	0.037
DA	0.034	0.070	0.100	0.101	0.161	0.125

For inferences of type B and C, the predictions for the effect of interval width depended on whether the inferences were p-valid or p-invalid. The F tests were therefore performed grouped by validity. For the p-valid inferences of type B, the variance in Group 1 was larger in the high premise probability condition ($EMM = .089$) than in the low premise probability condition ($EMM = .042$) ($F(125, 125) = 2.099, p < .001$; 95% CI [1.476, 2.986]). The variance in Group 2 was similar for the *high* ($EMM = .125$) and the *low* ($EMM = .128$) conditions ($F(149, 149) = 1.024, p = .884$). In Group 3 the variance also did not differ between the *high* ($EMM = .126$) and the *low* condition ($EMM = .134$) ($F(116, 116) = 1.061, p = .752$; 95% CI [.736, 1.529]). None of these three comparisons is in accordance with the predictions.

For the p-invalid inferences of type B, the variance in Group 1 was larger in the *high* ($EMM = .133$) than in the *low* ($EMM = .047$) premise probability condition ($F(125, 125) = 2.855, p < .001$; 95% CI [2.008, 4.061]). In Group 2 the variance was similar in the *high* ($EMM = .111$) and in the *low* ($EMM = .085$) condition ($F(149, 149) = 1.299, p = .112$; 95% CI [.941, 1.793]). In Group 3, the variance was larger in the high ($EMM = .128$) than in the low

($EMM = .062$) condition ($F(116, 116) = 2.074, p < .001; 95\% \text{ CI } [1.439, 2.990]$). Thus, in Groups 1 and 3 where a difference was found, it went in the direction of the predictions, but the absence of a significant difference in Group 2 was not in accordance with the predictions.

For the p-valid inferences of type C, variances in Group 1 were similar in the *high* ($EMM = .064$) and the *low* ($EMM = .086$) premise probability condition ($F(83, 83) = 1.335, p = .190; 95\% \text{ CI } [.865, 2.059]$). The variance in Group 2 was larger in the *low* ($EMM = .151$) than in the *high* ($EMM = .055$) condition ($F(99, 99) = 2.756, p < .001; 95\% \text{ CI } [1.854, 4.096]$). The variance in Group 3 was also larger in the *low* ($EMM = .161$) than in the *high* ($EMM = .079$) condition ($F(77, 77) = 2.047, p = .002; 95\% \text{ CI } [1.305, 3.209]$). This means that in Groups 2 and 3 where a difference was found, it went in the direction of the predictions, but the absence of a significant difference in Group 1 was not in accordance with the predictions.

For the p-invalid inferences of type C, the variance in Group 1 was larger in the *high* ($EMM = .075$) than in the *low* ($EMM = .037$) condition ($F(83, 83) = 2.06, p = .001; 95\% \text{ CI } [1.336, 3.177]$), in accordance with the predictions. In Group 2 the variance was larger in the *low* ($EMM = .129$) than in the *high* ($EMM = .092$) condition ($F(99, 99) = 1.409, p = .090; 95\% \text{ CI } [.096, .129]$), contrary to the predictions. In Group 3 the variance was also larger in the *low* ($EMM = .155$) than in the *high* ($EMM = .097$) condition ($F(77, 77) = 1.610, p = .038; 95\% \text{ CI } [1.026, 2.524]$), contrary to the predictions.

Overall, in contrast to the consistent evidence for sensitivity to the location of coherence intervals shown in the analysis of above-chance coherence, there was no consistent evidence for sensitivity to the width of coherence intervals. The findings generally went in the direction of the predictions for the invalid inferences of type B and for the valid inferences of type C, but were contrary to the predictions for the valid inferences of type B and for the invalid inferences of type C. Further, there were significant differences in the variance of responses to the inferences 1 and 2 of type A, contrary to the predictions.

A lack of sensitivity to the width of the coherence interval, in the presence of a sensitivity to its location, need not be a sign of lower conformance to coherence overall. It can also reflect the fact that the constraint of coherence only limits responses to a certain region of the probability range, and that this range must be narrowed down further using inductive criteria to be able to respond with a point probability as required in the instructions. It may be that among the cases in which people show sensitivity to the location of the interval, people also have sensitivity to the interval width when they perceive that there are not enough inductive criteria to narrow down the interval further. But that when the materials convey additional inductive criteria, people use them, thereby breaking up the relation between interval width and response variance. However, this is just a tentative explanation, and it would be good to see if the findings can be replicated before enquiring further about their source.

In what relation do the results here on above-chance coherence stand to earlier studies? The results replicate the finding of Cruz et al. (2015) and Politzer & Baratgin (2016) that

people's responses are coherent above-chance levels across a wide range of inferences. They also replicate the findings from Cruz et al. and Evans et al. (2015) that coherence is often established intuitively or automatically, in the absence of an explicit reasoning task, when the statements among which coherence is to be established are not presented together on the screen, but in random order. Further, the results support and extend the finding of Evans et al. (2015) that, when coherence is weaker, it can be increased through an explicit reasoning task that allows people to focus attention on the relations between the probabilities of the statements in the inference. This explicit task increased coherence for AC in Evans et al. In the present experiment, it increased coherence for IfOR and for MT.

The present experiment also explored possible causes for a failure of coherence. One factor investigated previously is the use of materials in which prototypicality conflicts with probability, discussed in relation to the conjunction fallacy in Experiment 2 and covered also in Politzer & Baratgin (2016). A further possible cause of the failure of coherence is the presence of negation effects leading to subadditivity, investigated by Evans et al. (2015) as well as in Costello & Watts (2016a). The findings in this experiment suggested three further possibilities. One is the failure of invariance, discussed in relation to MT. The second is the possibility that a failure of coherence can be an artefact caused by a high chance-rate coherence that renders it difficult to detect coherence above-chance levels when it is there. This problem likely affected above-chance coherence for the IfOr inference, although it cannot be ruled out that additional factors (e. g. a negation effect for the conditional in the antecedent) also played a role. A third possibility covered in this experiment, discussed also in Experiment 1 in relation to the OrIf inference, is that people may interpret the statements involved in the inference in a way different from the interpretation the experimenter used as a basis for the computation of response coherence. This may have affected responses to AC and DA, for which coherence was less reliable in the inferences task than in the statements task, and for which coherence became reliably above chance levels across tasks when computed under the assumption of a biconditional major premise.

Taken together, coherence was at above-chance levels across groups for 8 of the 12 inferences investigated, and it was above-chance levels in the inferences group for 10 of the 12 inferences investigated. For the remaining 2 inferences, coherence was above-chance levels across groups under the assumption that the conditional was interpreted as the probability biconditional, i.e. as a statement describing a covariation between p and q . It is striking that coherence was at above-chance levels for the majority of the inferences even in the absence of an inference task, suggesting that there is a tendency for beliefs to be coherent intuitively or automatically. At the same time, it is encouraging that conscious reflection about the dependency relations between statements can increase coherence when it is weaker. Additionally, the specific cases in which coherence was found not to be above chance levels can be informative about possible factors affecting reasoning beyond an assessment of the

extent to which the premises constrain the probability of the conclusion, and these can inform hypotheses for future experiments.

Experiment 4 investigates to what extent these findings can be replicated with a larger internet based sample.

Experiment 4

Experiment 4 was identical to experiment 3, with the exception that it was conducted over the internet instead of in the lab. Both experiments were implemented using the web based programme SoSci Survey (Leiner, 2014), available at www.soscisurvey.de⁹. This assured that the visual display of the task and the formatting of the response options were exactly the same in the two experiments.

Method

Participants. A total of 430 participants from English speaking countries completed the experiment through the online platform Prolific Academic (Peer et al., 2017). Two participants were excluded because they indicated at the end of the experiment that they had not taken part seriously, but instead just "clicked through"; two because they failed a catch trial asking them not to respond but to instead click "next" to continue with the experiment; fourteen were excluded from the statements group because they had two or more trial reaction times of 2 secs or less; and eleven from the inferences groups for having two or more trial reaction times of 3 secs or less. Two participants were excluded for not reporting at least "good" English language skills; and finally, two further participants were excluded because their devices did not support JavaScript, which was relevant to assure that the display conditions of the task were comparable across participants. The final sample consisted of 415 participants (124 in Group 1, 147 in Group 2, and 144 in Group 3). Participants' median age was 30 (range 18-74). They had a diverse formal educational background, with 22% reporting having finished 12th grade, 10% having a technical/applied degree, 49% having an undergraduate university degree, and 15% a postgraduate degree or higher. Participants' median ratings of task difficulty were 17% in Group 1, 56% in Group 2, and 67% in Group 3.

Materials, design, and procedure. These were the same as for Experiment 3. In Group 3, the percentage of correct responses to the 26 trials of the memory task ranged from 15.4% to 100% (median: 65.4%. As for Experiment 3, the chance rate of responding correctly on any trial was 1/3024). The distribution of the percentage of correct responses to the memory task is

⁹ SoSci Survey is a software for creating online experiments and surveys originally developed at the University of Munich, partly in cooperation with the University of Zurich. It is tailored to experimental research and is free of charge when used for academic as opposed to commercial or private purposes.

shown in the lower panel of Figure 5.4 above. One can see that the distribution was very similar for both experiments.

The median duration of the experimental session was 11.5 min for group 1, 11.2 min for group 2, and 9.8 min for group 3.

Results and discussion

Above-chance coherence. Figure 5.14 displays the results for observed and above-chance coherence for each inference. The overall results look very similar to those of Experiment 3. There was again a trivial increase in observed coherence for lower degrees of measurement precision, which remained present for above-chance coherence mainly in inferences 1 and 2. It is also of interest that in the statements task, coherence seemed to be above chance levels for all inferences, whereas in the inferences task above-chance coherence seemed to be less reliably present for IfOr and AC.

The pattern of responses was investigated further in a linear mixed model for the effects of experiment (3, 4), inference type, and group on above-chance coherence, with random intercepts for participants. Inclusion of further random effects led to failure of convergence. With only two levels, experiment was treated as a fixed level-2 predictor rather than as a predictor at level 3.

Overall coherence was around 24% higher than expected by chance ($EMM = .235$, $F(1, 706.606) = 1088.98$, $p < .001$). There was no main effect of experiment, $F(1, 706.606) = .065$, $p = .799$, indicating that overall, above-chance coherence was similar for Experiment 3 ($EMM = .237$) and Experiment 4 ($EMM = .233$). However, there was a small main effect of group ($F(2, 706.610) = 3.40$, $p = .034$): Above chance coherence was higher in the inferences group ($EMM_{\text{group2}} = .259$) than in the statements group ($EMM_{\text{group1}} = .215$; $F(1, 472.324) = 6.488$, $p = .011$). Above-chance coherence for the inferences group with working memory load fell in between ($EMM_{\text{group3}} = .230$) and did not differ significantly from that of the other two groups (group1 vs. group 3: $F(1, 451.205) = .72$, $p = .391$. Group 2 vs. group 3: $F(1, 489.724) = 2.702$, $p = .101$).

As in Experiment 3, there was a main effect of inference type ($F(2, 12557.108) = 41.585$, $p < .001$). Above-chance coherence was higher for the inferences of type A ($EMM = .290$) than for those of type B ($EMM = .251$; $F(1, 8189.152) = 7.04$, $p = .008$) and of type C ($EMM = .163$; $F(1, 6006) = 72.887$, $p < .001$); and it was higher for the inferences of type B than for those of type C ($F(1, 10373.218) = 51.81$, $p < .001$). But this main effect was qualified by an interaction between inference type and experiment, $F(2, 12557.108) = 8.308$, $p < .001$. No other effects were significant.

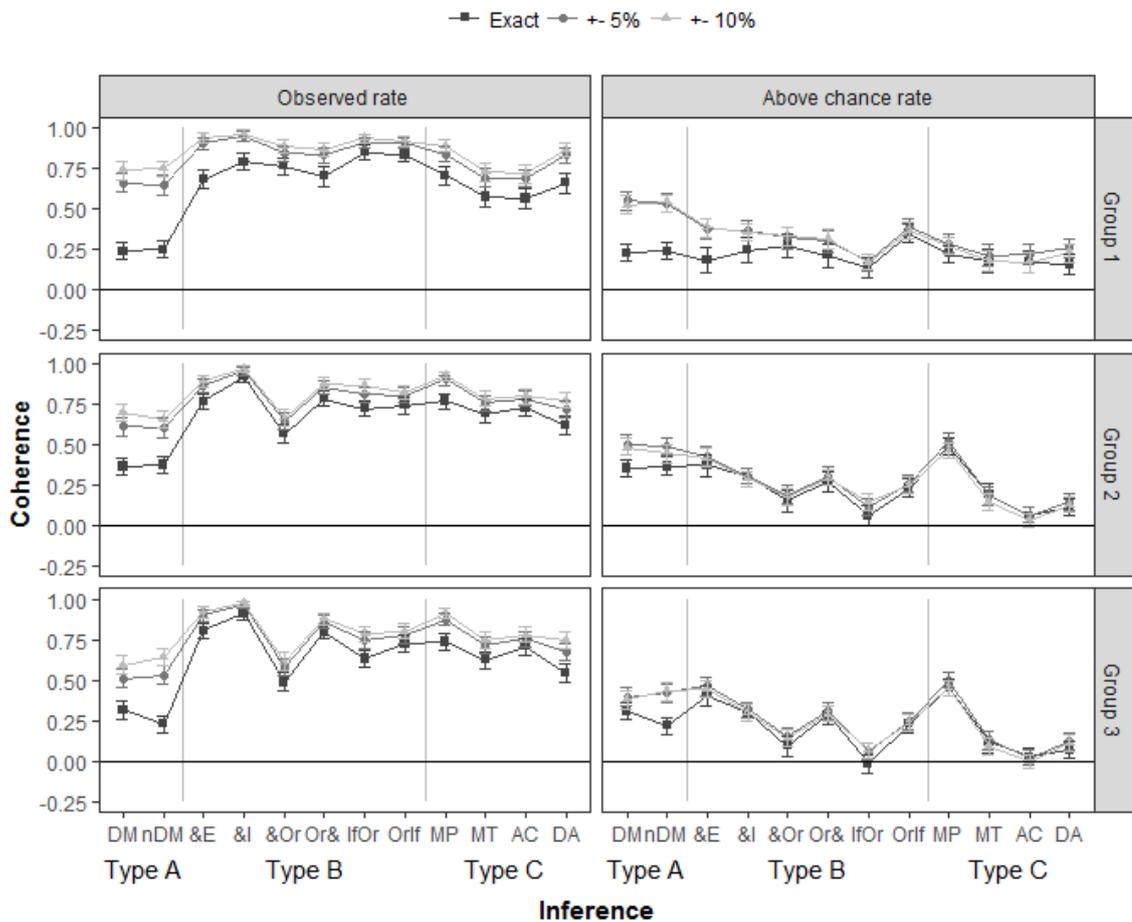


Figure 5.14. Mean values of observed and above-chance coherence for the 12 inferences of Experiment 4, separately for each group and for three levels of measurement precision. The black horizontal line represents a coherence rate of 0% in the panels for observed coherence, and it represents the chance rate of a coherent response in the panels for above-chance coherence. Error bars show 95% CIs.

Follow-up analyses to the interaction between inference type and experiment were conducted including only fixed effects, because some of the models including random effects failed to converge. For inferences of type A, above-chance coherence was similar for Experiment 3 ($EMM = .296$) and Experiment 4 ($EMM = .285$), $F(1, 2184) = .271$, $p = .603$. For inferences of type B, above-chance coherence was higher in Experiment 3 ($EMM = .274$) than in Experiment 4 ($EMM = .227$), $F(1, 6551) = 9.104$, $p = .003$. For inferences of type C, above-chance coherence was higher in Experiment 4 than in Experiment 3, $F(1, 4368) = 7.046$, $p = .009$.

Figure 5.14 suggests that the source of the interaction comes from the less reliable above-chance coherence for ifOr in Experiment 4, and the more reliable above-chance coherence for MT and DA in Experiment 4. Specifically, coherence seemed to have been clearly at above-chance levels across groups in Experiment 4 except for the inferences of IfOr and AC. The inferences &Or, MT, and DA were also near the horizontal line of zero above-chance

coherence in some conditions. Hence further specific analyses were performed for these five inferences. As for Experiment 3, these models included only fixed effects because the inclusion of random effects led to failure of convergence in some of the comparisons.

Responses to &Or were coherent above chance levels in all groups (for Group 1: $EMM = .275$, $F(1, 248) = 58.693$, $p < .001$. For Group 2: $EMM = .155$, $F(1, 294) = 21.780$, $p < .001$. For Group 3: $EMM = .096$, $F(1, 288) = 9.666$, $p = .002$).

Responses to IfOr were coherent above chance levels in Group 1 ($EMM = .138$, $F(1, 248) = 20.011$, $p < .001$). Responses remained marginally coherent above chance levels in Group 2 ($EMM = .063$, $F(1, 294) = 4.014$, $p = .046$). Coherence was not above chance levels for this inference in Group 3 ($EMM = -.014$, $F(1, 288) = .227$, $p = .634$).

Responses to MT were coherent above chance levels across groups (For Group 1: $EMM = .176$, $F(1, 248) = 21.488$, $p < .001$. For Group 2: $EMM = .190$, $F(1, 294) = 32.751$, $p < .001$. For Group 3: $EMM = .119$, $F(1, 288) = 13.491$, $p < .001$).

Responses to AC were coherent above chance levels in Group 1 ($EMM = .170$, $F(1, 248) = 22.308$, $p < .001$) and in Group 2 ($EMM = .061$, $F(1, 294) = 6.071$, $p = .014$). But coherence was not above chance levels in Group 3 ($EMM = .028$, $F(1, 288) = 1.131$, $p = .288$).

Responses to DA were coherent above chance levels across groups (for Group 1: $EMM = .158$, $F(1, 248) = 27.711$, $p < .001$. For Group 2: $EMM = .114$, $F(1, 294) = 18.316$, $p < .001$. For Group 3: $EMM = .073$, $F(1, 288) = 7.838$, $p = .005$).

Overall, the results show that coherence was above chance levels for all inferences in Groups 1 and 2, and that in Group 3 coherence ceased to be above chance levels only for ifOr and AC. This can be considered impressive evidence in favour of people's general sensitivity to the constraint of coherence.

At the same time, the experiment replicates the lower rate of above chance coherent responses for IfOr and AC, found in Experiment 3. Given the strong overall evidence for above-chance coherence, the lower rate for these two inferences is unlikely to indicate a general lack of sensitivity to coherence. It seems more plausible that it is a result of inference specific interpretational or pragmatic factors. In the case of IfOr, the premise was again assigned a lower probability, leading to a higher chance rate, and a correspondingly lower rate of above-chance coherence. However, Experiments 3 and 4 cannot rule out that further factors – such as a negation effect – also play a role in the lower coherence rate for this inference. This possibility would have to be investigated further in experiments in which the chance rate is held constant across inferences.

Figure 5.13 of Experiment 3 displayed above-chance coherence for AC and DA under three interpretations of the mayor premise: as the original probability conditional, *if p then q*, as a probability biconditional *if p then q & if q then p*, and as the converse of the original conditional, *if q then p*. The finding that responses to these inferences, which were not coherent above chance levels under the original conditional interpretation, would have been

above chance levels under a biconditional interpretation, may apply also to the present experiment.

A comparison of Figures 5.10 and 5.14 suggests that the observed coherence rate was lower for &Or in Experiment 4 than in Experiment 3, although it remained above chance levels in all groups in Experiment 4. A possible reason for a lower observed coherence rate for this inference may be that people found it pragmatically infelicitous to infer a disjunction from a conjunction, because this involves making a less informative statement than one could make (Grice, 1989). Such a pragmatic reason would explain why the observed coherence rate seemed to be lower in the inferences tasks than in the statements task, and why this effect was not observed consistently across experiments, appearing to play a role in Experiment 4 but not in Experiment 3. The similar inference of or-introduction, *p, therefore p or q*, has been found to be accepted less often by participants under binary paradigm instructions, asking them to assume the premise to be true to then judge whether the conclusion also has to be true (Braine et al., 1984; Orenes & Johnson-Laird, 2012; Rips, 1983). However, under probabilistic instructions or-introduction is accepted to a similar degree as other valid inferences (Cruz et al., 2015; Cruz et al., 2017; Politzer & Baratgin, 2016; see also Cruz et al., 2016, for a high endorsement rate of *and-to-or* in a study with probabilistic instructions). At any rate, this question cannot be suitably addressed in the present experiment, given the differences in chance rate coherence across inferences and groups. In this experiment it can only be stated that whatever factors led to the lower observed coherence rate, this rate was still reliably above chance across groups.

Overall, the above analysis provides strong evidence for sensitivity to the location of coherence intervals. The following section examines participants' sensitivity to differences in interval width.

Response variance. As in Experiment 3, the question of people's sensitivity to the width of a coherence interval was addressed by comparing the variance in responses between premise probability conditions. The general prediction was that response variance would be larger in the conditions in which the coherence interval was large than in those in which it was narrow. The inference-specific predictions were the same as for Experiment 3. Table 5.9 displays the variances for each condition. One can see that overall, the directions of the differences in Table 5.9 are very similar to those of Experiment 3.

For the DM inference of type A, variances in Group 1 differed between premise probability conditions ($F(123, 123) = 2.762, p < .001; 95\% \text{ CI } [1.936, 3.939]$). Variances in Group 2 were similar between premise probability conditions ($F(146, 146) = 1.455, p = .024; 95\% \text{ CI } [1.051, 2.015]$). Variances in Group 3 were also similar between premise probability conditions ($F(143, 143) = 1.145, p = .418; 95\% \text{ CI } [.824, 1.59]$). Given the prediction of no effect of premise probability for this inference, the results are in accordance with the predictions for Groups 2 and 3, but not for Group 1.

For the nDM inference of type A, variances differed between premise probability conditions in all Groups (In Group 1: ($F(123, 123) = 2.833, p < .001; 95\% \text{ CI } [1.987, 4.041]$). In Group 2: $F(146, 146) = 6.354, p < .001; 95\% \text{ CI } [4.587, 8.800]$. In Group 3: $F(143, 143) = 5.379, p < .001; 95\% \text{ CI } [3.871, 7.475]$), contrary to the predictions.

As in Experiment 3, the predictions for inferences of type B and C depended on whether the inferences were p-valid or p-invalid. The tests were therefore performed grouped by validity. For the p-valid inferences of type B, the variance in Group 1 was larger in the *high* ($EMM = .067$) than in the *low* ($EMM = .039$) premise probability condition ($F(371, 371) = 1.695, p < .001; 95\% \text{ CI } [1.382, 2.078]$). In Group 2 the variance was similar in the *high* ($EMM = .153$) and the *low* ($EMM = .147$) condition ($F(440, 440) = 1.041, p = .674; 95\% \text{ CI } [.863, 1.26]$). In Group 3, the variance was also similar in the *high* ($EMM = .158$) and the *low* ($EMM = .134$) condition ($F(431, 431) = 1.182, p = .083; 95\% \text{ CI } [.978, 1.428]$). None of these three comparisons was in accordance with the predictions.

Table 5.9. Variances of conclusion probability judgments in Experiment 4, separately for each group, inference, and premise probability condition.

Inference	Group 1 (n=42)		Group 2 (n=50)		Group 3 (n=39)	
	<i>low</i>	<i>high</i>	<i>low</i>	<i>high</i>	<i>low</i>	<i>high</i>
DM	0.028	0.077	0.165	0.113	0.127	0.111
nDM	0.081	0.029	0.183	0.029	0.164	0.030
&E	0.034	0.019	0.175	0.057	0.153	0.027
&I	0.026	0.019	0.048	0.087	0.031	0.077
&Or	0.046	0.075	0.112	0.173	0.104	0.180
Or&	0.036	0.006	0.087	0.160	0.064	0.117
IfOr	0.039	0.102	0.144	0.170	0.134	0.162
OrIf	0.048	0.143	0.167	0.173	0.165	0.157
MP	0.089	0.005	0.130	0.052	0.124	0.039
MT	0.026	0.049	0.139	0.129	0.127	0.109
AC	0.027	0.020	0.164	0.074	0.157	0.067
DA	0.033	0.077	0.146	0.138	0.127	0.104

For the p-invalid inferences of type B, the variance in Group 1 was larger in the *high* ($EMM = .103$) than in the *low* ($EMM = .037$) premise probability condition ($F(371, 371) = 2.809, p < .001; 95\% \text{ CI } [2.291, 3.809]$). The variance in Group 2 was also larger in the *high* ($EMM = .147$) than in the *low* ($EMM = .109$) condition ($F(440, 440) = 1.343, p = .002; 95\% \text{ CI } [1.114, 1.620]$). In Group 3 the variance was also higher in the *high* ($EMM = .137$) than in

the *low* ($EMM = .098$) condition ($F(431, 431) = 1.394, p < .001$; 95% CI [1.154, 1.684]). The three comparisons were in accordance with the predictions.

For the p-valid inferences of type C, the variance in Group 1 was larger in the *high* ($EMM = .083$) than in the *low* ($EMM = .028$) premise probability condition ($F(247, 247) = 2.992, p < .001$; 95% CI [2.329, 3.841]), contrary to the predictions. In Group 2 the variance was larger in the *low* ($EMM = .164$) than in the *high* ($EMM = .096$) condition ($F(293, 293) = 1.719, p < .001$; 95% CI [1.367, 2.163]), in accordance with the predictions. The variance in Group 3 was also larger in the *low* ($EMM = .155$) than in the *high* ($EMM = .080$) condition ($F(287, 287) = 1.931, p < .001$; 95% CI [1.531, 2.435]), in accordance with the predictions.

Finally, for the p-invalid inferences of type C, the variance in Group 1 was higher in the *high* ($EMM = .068$) than in the *low* ($EMM = .030$) premise probability condition ($F(247, 247) = 2.273, p < .001$; 95% CI [1.771, 2.919]) in accordance with the predictions. In Group 2 the variance was marginally larger in the *low* ($EMM = .158$) than in the *high* ($EMM = .124$) condition ($F(293, 293) = 1.268, p = .043$; 95% CI [1.009, 1.595]), contrary to the predictions. The variance in Group 3 was also larger in the *low* ($EMM = .148$) than in the *high* ($EMM = .095$) condition ($F(287, 287) = 1.558, p < .001$; 95% CI [1.236, 1.964]), contrary to the predictions.

Overall, the present findings provide no consistent evidence of sensitivity to the width of a coherence interval. The results for DM were in accordance with the predictions in Groups 2 and 3, but not in Group 1. The results for nDM were contrary to the predictions in all groups. The findings for the inferences of type B were in accordance with the predictions for the invalid inferences, but not for the valid inferences. And the findings for the inferences of type C depended on the group: for the p-valid inferences they were in accordance with the predictions in Groups 2 and 3, but not in Group 1. In contrast, for the p-invalid inferences the findings were in accordance with the predictions for Group 1 but not for Groups 2 and 3.

The findings on response variance were very similar in Experiments 3 and 4, and they show a parallel to the findings on above-chance coherence. In both experiments, coherence among the inferences of type B was less reliable for IfOr, and in line with this, the findings on response variance were in accordance with the predictions for the p-invalid but not the p-valid inferences. Similarly, the findings on above-chance coherence for inferences of type C depended on the group: For the p-valid inferences responses were more reliably coherent in the inferences group than in the statements group, whereas for the p-invalid inferences they were more reliably coherent in the statements group than in the inferences group – and we find the exact same pattern for response variance. Although the findings provide no consistent evidence of a sensitivity to the width of the interval, they also suggest that in the cases in which the findings were in accordance with the predictions, they were systematically so, in that they mirrored the results on above-chance coherence, suggesting that the measure of response variance used did tap an aspect of sensitivity to coherence as it had intended.

General discussion

In summary, Experiments 3 and 4 provide strong and consistent evidence that people are sensitive to coherence constraints both for one- and for two-premise inferences. Where coherence was less reliably present, it sometimes helped to have an explicit inference task, at least when no secondary working memory load task diverted attentional resources from it. But the presence of an explicit inference task was not always helpful. Whether it increases the rate of coherent responses or not seems to depend on interpretational and pragmatic factors. If people interpret the statements in the task differently from the experimenters, an explicit inference task can also decrease the rate of coherent responses. This can be just as informative as the reliable presence of coherence, because it can inform hypotheses about which interpretational and pragmatic factors may be playing a role in the reasoning process in addition to the coherence constraints investigated here.

In contrast to the clear evidence for sensitivity to the location of a coherence interval, both experiments provided only equivocal evidence of sensitivity to interval width. As mentioned above, the lack of an effect of interval width need not be seen as evidence for a lower sensitivity to coherence. This is because coherence only constrains responses to an interval, but within this interval inductive criteria may or may not constrain responses further, given the content and context of the inferences. When inductive criteria have a consistent effect on responses across participants, the relation between interval width and response variance can be broken. Where no consistent inductive criteria are perceived that would constrain responses to a certain region of the interval, the variance of responses can be expected to increase with interval width. However, this is only a tentative explanation, and the effect of interval width still has to be investigated further, with a variety of methods, to assess the value of such an explanation.

Experiments 3 and 4 investigated for which of a series of inferences coherence was above chance levels under different conditions. But the differences in the chance rate between conditions prevented the formulation of quantitative judgments about above-chance coherence. For example, judgments for MP were coherent above chance levels across groups. But were they more coherent in the inferences task than in the statements task? The inferences of DM and nDM were also coherent above chance levels across groups. But were they more or less coherent than responses to MP, or was coherence similar for the three inferences? And were responses to &Or really less coherent than for other one premise inferences? Such quantitative comparisons will be addressed in the next series of experiments.

CHAPTER 6. EXPERIMENTS 5 TO 7: QUANTITATIVE COMPARISONS OF DEGREES OF COHERENCE

Contents

- 6.1 Experiment 5: At the edge vs. the centre of the coherence interval
 - 6.1.1 Method
 - 6.1.2 Results and discussion
 - 6.1.3 General discussion
- 6.2 Experiment 6: Higher vs. lower than the premise probabilities
 - 6.2.1 Method
 - 6.2.2 Results and discussion
 - 6.2.3 General discussion
- 6.3 Experiment 7: Certain premises and binary paradigm instructions
 - 6.3.1 Method
 - 6.3.2 Results and discussion
 - 6.3.3 General discussion

EXPERIMENT 5: AT THE EDGE VS. THE CENTRE OF THE COHERENCE INTERVAL

Experiments 1 to 4 provided strong evidence that people's responses to a range of inferences of differing complexity are coherent at above chance levels. Experiments 5 to 7 move beyond this qualitative finding to investigate the degree to which responses are coherent above chance levels for these inferences, and whether people are more coherent for some inferences than for others. For example, is coherence higher for simpler one-premise inferences than for two-premise inferences? And does coherence differ for valid and invalid inferences?

The latter question is important in the investigation of a role of p-validity over and above coherence. As mentioned earlier, the definition of p-validity necessarily rests on that of coherence, and this makes it difficult to test whether p-validity plays an independent role. But if coherence differs for valid and invalid inferences, then this difference is not something that can be explained from within coherence, whereas it would be accounted for naturally by p-validity.

The way such quantitative comparisons of coherence become possible is by holding the chance rate constant between inferences and conditions. This was done in experiments 5 to 7 by using a binary response format, which rendered the chance rate equal to 50% in all cases. The precise task given to participants using this response format differed between Experiments 5 and 7.

Experiment 5 investigated questions about relative coherence between individual inferences, and between groups of inferences, such as those that are valid or invalid, and those of type A, B or C. But in addition, the experiment specifically studied the role of the location of the probability of the conclusion relative to the coherence interval. This was done by making two comparisons. The first was whether the probability of the conclusion was (a) clearly inside, or clearly outside, the interval, or alternatively (b) at the interval edge. The second was whether the probability of the conclusion was (a) inside or (b) outside of the interval.

Suppose people are sensitive to coherence, but their subjective "scale" for degrees of belief is coarser than a point probability, as suggested by Figures 5.10 and 5.14 of Experiments 3 and 4, respectively. Then one can predict that people's judgments would tend to be more coherent when a conclusion probability was clearly inside, or clearly outside, the relevant coherence interval than when it was at the interval edge. If people tend to be coherent but have rather "coarse" degrees of belief, then this would have implications for the development of algorithmic level accounts in reasoning and decision making. No specific prediction was made for whether participants would be more, or less, coherent when assessing whether the conclusion probability was inside, or outside, the coherence interval.

Participants received an inference task in which the probabilities of both the premises and the conclusion were given by the experimenter. The task was to judge whether or not the probability of the conclusion was consistent with the probabilities of the premises.

In contrast to Experiments 1 to 4, the task in this experiment was purely deductive. In Experiments 1 to 4 participants had been asked to generate their own conclusion probabilities, and these conclusion probabilities were constrained deductively by coherence. But coherence only constrained responses to a given interval, and with the exception of the inferences of DM and nDM, this interval was wider than a point value. Given that the task instructions asked for a point value as a response, it was up to inductive criteria or chance to narrow down the interval to a specific point. In the present experiment, in contrast, the question participants were asked could be fully answered using only the deductive constraints of coherence.

Method

Participants

A total of 136 participants from English speaking countries completed the online experiment in exchange for approximately £5 per hour. Participants accessed the experiment through the online platform Prolific Academic. Three participants were excluded because they indicated at the end of the experiment that they had not taken part seriously but had just "clicked through". Further 24 were excluded because they failed one or both of two catch trials designed to check whether participants were reading the materials. The final sample consisted of 109 participants. None of them had trial reaction times of 3 seconds or less, and they all indicated having at least "good" English language skills. Their median age was 31 years (range: 18-73), and most reported having some college education: Around 26.5% indicated having finished 12th grade, 12% reported having a technical/applied degree, 44% reported an undergraduate degree, and 17.5% a postgraduate degree. Participants' median rating of the difficulty of the experiment was 74%.

Design and materials

Experiments 5, 6, and 7 investigated the same 12 inferences as Experiments 3 and 4. These inferences are reproduced in Table 6.1. Each inference was presented in three premise probability conditions. For the one-premise inferences, these were the probabilities of 1, .8 and .5, which were taken as possible instantiations of "certain", "high", and "medium" degrees of belief. For the two-premise inferences, both premises always had the same probability in order to simplify the task for participants. These premise probabilities were matched to those for the one-premise inferences not in terms of their numerical value, but in terms of the sum of their uncertainty (with uncertainty = 1 – probability, Adams, 1998). This implies that the conditions

for the one-premise inferences with a premise probability of 1, .8, and .5 corresponded to the conditions for the two-premise inferences in which both premises had probabilities of 1, .9, and .75, respectively. For example, for a two-premise inference with premise probabilities of .9, the sum of the uncertainties of the premises is $(1 - .9) + (1 - .9) = .2$, which is equal to the uncertainty of a one-premise inference with a premise probability of .8. For ease of exposition, the results section refers to a condition with a *premise probability* of 1, .8, and .5, even though strictly speaking this is only true for the one-premise inferences. For the two-premise inferences the conditions of 1, .8, and .5 really refer to the complement of their uncertainty sum rather than to their premise probability.

Table 6.1. The inferences investigated in Experiments 5 to 7.

Type	Name	Validity	Form
A. One-premise, equivalence and contradiction	1. De Morgan (DM)	1	$\text{not}(p \ \& \ q) \ \therefore \ \text{not-}p \ \text{or} \ \text{not-}q$
	2. not De Morgan (nDM)	0	$p \ \& \ q \ \therefore \ \text{not-}p \ \text{or} \ \text{not-}q$
B. One-premise, valid in only one direction, left to right, or right to left	3. and-elimination (&E)	1	$p \ \& \ q \ \therefore \ p$
	4. and-introduction (&I)	0	$p \ \therefore \ p \ \& \ q$
	5. and-to-or (&Or)	1	$p \ \& \ q \ \therefore \ p \ \text{or} \ q$
	6. or-to-and (Or&)	0	$p \ \text{or} \ q \ \therefore \ p \ \& \ q$
	7. if-to-or (IfOr)	1	$\text{if not-}p \ \text{then} \ q \ \therefore \ p \ \text{or} \ q$
	8. or-to-if (OrIf)	0	$p \ \text{or} \ q \ \therefore \ \text{if not-}p \ \text{then} \ q$
C. Two-premise, conditional syllogisms	9. Modus ponens (MP)	1	$\text{if } p \ \text{then } q, p \ \therefore \ q$
	10. Modus tollens (MT)	1	$\text{if } p \ \text{then } q, \text{not-}q, \ \therefore \ \text{not-}p$
	11. Affirmation of the consequent (AC)	0	$\text{if } p \ \text{then } q, q, \ \therefore \ p$
	12. Denial of the antecedent (DA)	0	$\text{if } p \ \text{then } q, \text{not-}p, \ \therefore \ \text{not-}q$

Note. "1" = "valid", "0" = "invalid", "∴" = "therefore".

Within each premise probability condition, each inference was presented five times, with five different conclusion probabilities. These conclusion probabilities are represented as blue dots in Figure 6.1. The vertical black lines in the figure represent the coherence intervals for each premise probability condition. Notice that for the premise probabilities used, Figure 6.1 illustrates a clear difference between the valid and invalid inferences, with the intervals for the probability preserving valid inferences being restricted to higher values, and those of the

invalid inferences taking low or more probabilistically uninformative values. Figure 6.1 also shows that the standardisation with respect to premise probability came at the cost of a lack of standardisation with respect to interval width. An exception was that all the intervals for the inferences of type B had equal width when the probability of the premise was .5.

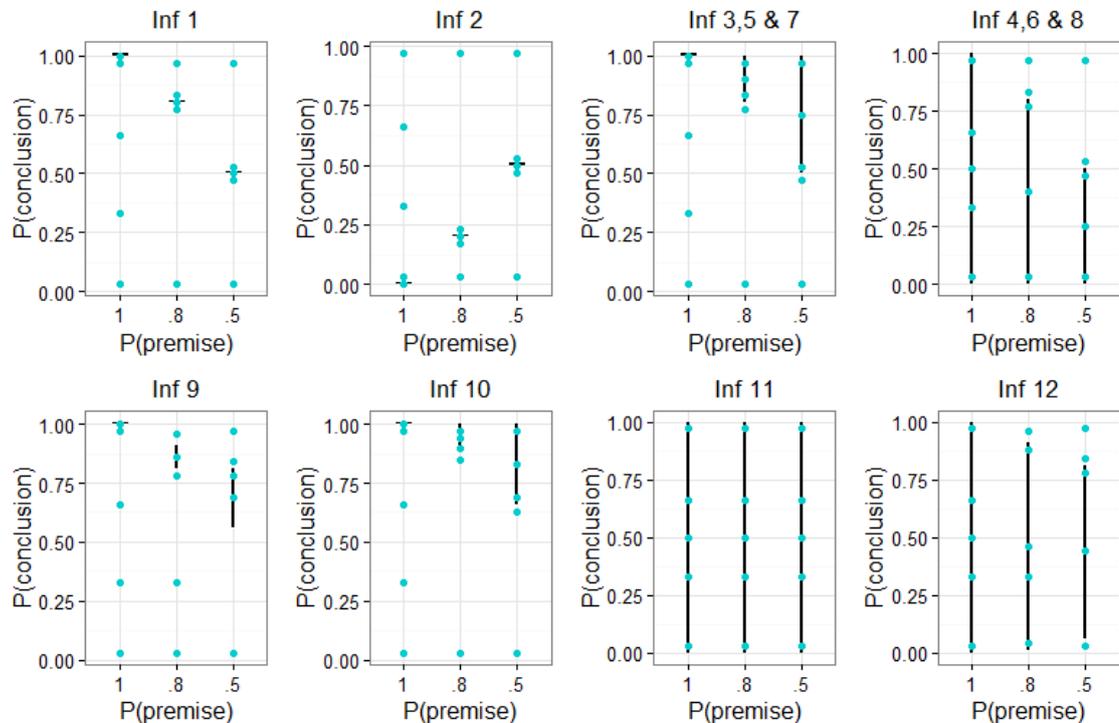


Figure 6.1. The conclusion probabilities used in Experiment 5 for each inference. The dots represent the conclusion probabilities, and the vertical lines represent the coherence intervals for each premise probability condition.

The conclusion probabilities were chosen so as to lie at the centre of the interval, at its inner or outer edge, or clearly outside of the interval. In the latter case they were placed at the upper or lower end of the probability scale. Sometimes this rationale for varying conclusion probabilities relative to the location of the interval required fewer than five conclusion probabilities to be instantiated. For example, the cases in which the coherence interval equalled either the point value of 1 or the unit interval required only three conclusion probabilities: one equal to the point value resp. to the centre of the unit interval, one next to the point or at the top of the unit interval, and one at the bottom of the interval, around zero. In such cases further items were added in order to nonetheless have five items for each combination of inference with premise probability condition. Conversely, there were two cases (out of 36) that required more than five conclusion probabilities to be instantiated. One can see these in Figure 6.1: For inference 9 (MP), the case in which the premise probability condition was .5 (i. e. the case in which the sum of premise uncertainties was .5, because each premise

probability was .75) would have required two additional conclusion probabilities to cover the inner and outer edge of the lower end of the coherence interval. In this case priority was given to cover the upper end of the interval, in order to allow an assessment of overconfidence for MP. Overconfidence for this inference, unlike underconfidence, cannot be measured in the binary approach to reasoning, but it can be in the probabilistic approach.

The second case in which five conclusion probabilities were not enough to cover all relevant positions of the interval was for inference 12 (DA) when the premise probability condition was .5 (i. e., in the condition in which the probability of each premise was .75, and so the sum of premise uncertainties was .5). Figure 6.1 shows that here one further conclusion probability would have been necessary to cover the inner edge of the lower bound of the coherence interval. In this case the outer edge of the lower bound was given priority, because the inner edge of the lower bound is already covered in the condition in which premise probability is .8.

Overall, by fixing the number of conclusion probabilities for each combination of inference with premise probability to five in all cases, it was possible to capture the vast majority of relevant locations on the probability scale, while limiting the number of irrelevant additional items, and preventing some conditions from being more salient than others merely because of differences in their frequency of occurrence. A higher saliency for some conditions than others based on their frequency of occurrence could have led participants to process the oddball items with heightened attention, possibly leading to higher coherence (c. f. the effect of working memory load in Experiments 3 and 4). Differences in coherence due to the logical form of the inferences would then be confounded with differences in coherence due to saliency – a problem that is avoided by equating the number of conclusion probabilities used across conditions.

Table 6.2 provides more detailed information on the conclusion probabilities used. The rightmost column of the table shows that one of the conclusion probabilities was always a point value equal to the centre of the coherence interval. The remaining four conclusion probabilities were randomly selected for each participant and condition out of a range of five values, determined by the location of the coherence interval for the respective condition, and by the upper and lower ends of the probability scale. For example, suppose the lower bound of a coherence interval for some combination of inference and premise probabilities was .6. Then to capture the inner edge of the lower bound of this interval, a random number between .61 and .65 would be selected. And to capture the outer edge of this lower interval bound, a random number between .55 and .59 would be selected. Hence, the edges of intervals were captured by taking a random number within five percentage points of either side of the edge. An exception was when the edge of an interval was very near the lower or upper bound of the probability scale. This was the case for example for inference 12 (DA) in the condition in which the sum of premise uncertainties was .8 (see Figure 6.1). In this case the five percentage

point range for the outer edge of the upper interval bound would have overlapped with the five percentage point range for the upper end of the scale. Therefore instead of using two overlapping ranges, a single five percentage point range was used, which was centred between the upper edge of the coherence interval and the upper end of the probability scale.

Table 6.2. The conclusion probabilities used in Experiment 5 for each inference and premise probability condition.

Inference	P(premise)	Coherence Interval	P(conclusion)
1 (DM)	1	1	1, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
	.8	.8	.8, [.95,.99], [.81,.85], [.75,.79], [.01,.05]
	.5	.5	.5, [.95,.99], [.51,.55], [.45,.49], [.01,.05]
2 (nDM)	1	0	0, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
	.8	.2	.2, [.95,.99], [.21,.25], [.15,.19], [.01,.05]
	.5	.5	.5, [.95,.99], [.51,.55], [.45,.49], [.01,.05]
3 (&E),	1	1	1, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
5 (&Or),	.8	[.8,1]	.9, [.95,.99], [.81,.85], [.75,.79], [.01,.05]
7 (IfOr)	.5	[.5,1]	.75, [.95,.99], [.51,.55], [.45,.49], [.01,.05]
4 (&I),	1	[0,1]	.5, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
6 (Or&),	.8	[0,.8]	.4, [.95,.99], [.81,.85], [.75,.79], [.01,.05]
8 (OrIf)	.5	[0,.5]	.25, [.95,.99], [.51,.55], [.45,.49], [.01,.05]
9 (MP)	1	1	1, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
	.8	[.81,.91]	.86, [.94,.98], [.76,.80], [.01,.05], [.31,.35]
	.5	[.5625,.8125]	.69, [.95,.99], [.82,.86], [.76,.80], [.01,.05]
10 (MT)	1	1	1, [.95,.99], [.01,.05], [.64,.68], [.31,.35]
	.8	[.88,1]	.94, [.95,.99], [.89,.93], [.83,.87], [.01,.05]
	.5	[.66,1]	.83, [.95,.99], [.67,.71], [.61,.65], [.01,.05]
11 (AC)	1	[0,1]	.5, [.95,.99], [.64,.68], [.31,.35], [.01,.05]
	.8	[0,1]	.5, [.95,.99], [.64,.68], [.31,.35], [.01,.05]
	.5	[0,1]	.5, [.95,.99], [.64,.68], [.31,.35], [.01,.05]
12 (DA)	1	[0,1]	.5, [.95,.99], [.64,.68], [.31,.35], [.01,.05]
	.8	[.01,.91]	.46, [.94,.98], [.86,.90], [.02,.06], [.31,.35]
	.5	[.0625,.8125]	.44, [.95,.99], [.82,.86], [.76,.80], [.01,.05]

Note. Conclusion probabilities without brackets denote the centre of the respective coherence interval. Conclusion probabilities in square brackets represent the minimum and maximum of a range of values from which a number was drawn randomly for each participant and condition.

With 12 inferences, 3 premise probability conditions and 5 conclusion probabilities in each premise probability condition, the experiment had $12 \times 3 \times 5 = 180$ trials, plus two catch trials to check whether participants were paying attention. The catch trials were similar in format to the regular trials, but the text for the premises and conclusion of the inferences was replaced with text stating that they were control trials to make sure participants were paying attention. Participants were asked not to respond, and were told that the experiment would continue automatically on the next page in a few seconds. The catch trails remained on screen for 8 seconds.

On each trial, the inference was introduced through a short context story in which a protagonist expressed a given degree of belief in the premises and in the conclusion. Participants' task was to indicate whether or not the likelihood that the protagonist assigned to the conclusion was consistent with the likelihood the protagonist assigned to the premise or premises. Participants were asked to provide their answers using the arrows on their keyboard, the left arrow standing for "no" and the right arrow for "yes". On each trial a picture of two arrows was presented: a red arrow pointing to the left with the word "no" written on it, and a green arrow pointing to the right displaying the word "yes". These arrow pictures were sensitive to mouse clicks, so that it was also possible to respond using a mouse.

The context stories were pseudonaturalistic. That is, they described concrete but fictitious situations in which it would be difficult to draw on world knowledge to judge the probabilities of the events involved. Further, the stories aimed to convey that the conclusions of the inferences were important or consequential, and that at the same time careful thinking as opposed to jumping to conclusions was called for. The reason for this was as follows. One of the purposes of the experiment was to compare coherence for valid and invalid inferences. But it seems implausible that the distinction between deduction and induction would be relevant in all contexts. A context in which it may become relevant is when it is worth being "conservative" because of a higher risk of drawing a conclusion that goes beyond what follows necessarily from the premises. The frame below shows a sample trial for inference 1 (DM) and a premise probability of .8.

The meanings of "premise" and "conclusion" were explained in the instructions. The experiment used six context stories on a range of topics: the research report of a team of zoologists, the murder of a member of parliament, an injured patient in an emergency hospital, a water dam with cracks, a robot mission to mars, and a cholera epidemic. The full description of the context stories can be found in Appendix D.

Each context story was randomly assigned to two of the twelve inferences for each participant. Within each participant the 15 trials in which each inference was involved referred to the same overall scenario, but the events described in the scenario varied slightly between premise probability conditions and between the two inferences to which the scenario was assigned. For example, in the case of the sample scenario above, the inference made reference

to different patients in each premise probability condition (patients P. M., H. D., and R. S.), and the doctor who expressed the beliefs about the premise and conclusion probabilities differed between inferences (Miriam and Leslie). These changes were introduced to avoid carry-over effects between trials, or an attempt to establish coherence across trials when only coherence within a trial was assessed.

Imagine you are part of a team of doctors who are working in an emergency hospital. Several patients are brought in by the ambulance with a variety of severe injuries. It is important that you act carefully on these cases because a wrong diagnosis could be fatal. You are reviewing their files with Miriam, another doctor in the team.

Based on the information gathered until now:

Premise: Miriam thinks it's **80%** likely that:

Patient P. M. does **not** have both a liver injury and a kidney injury.

Conclusion: Miriam thinks it's **16%** likely that:

Patient P. M. does **not** have a liver injury or he does **not** have a kidney injury.

Is the likelihood assigned to the conclusion consistent with the likelihood assigned to the premise?

Procedure

After the instructions and an example, participants went through 15 practice trials involving a different scenario and different inferences from those of the main experiment. The practice trials included five trials with the inference *p, therefore not-p* and a premise probability of 1, which were used to assess whether participants understood the instructions. At the end of the experiment participants provided demographical information and indicated whether they had taken part seriously or had just "clicked through". The last page contained debriefing information. The entire session took on average 18 minutes to complete.

Results and discussion

Figure 6.2 shows the proportion of "yes" and "no" responses to the contradiction *p, therefore not-p* with a premise probability of 1, used in the practice trials. The coherent response was "yes" for the second conclusion probability displayed on the x axis: probability 0, and "no" for the remaining four conclusion probabilities. Although responses appeared to be slightly less coherent for the extreme conclusion probabilities of 0 and 1 than for intermediate

probabilities, overall responses closely conformed to the predicted pattern, suggesting that participants generally had no problem in understanding the instructions.

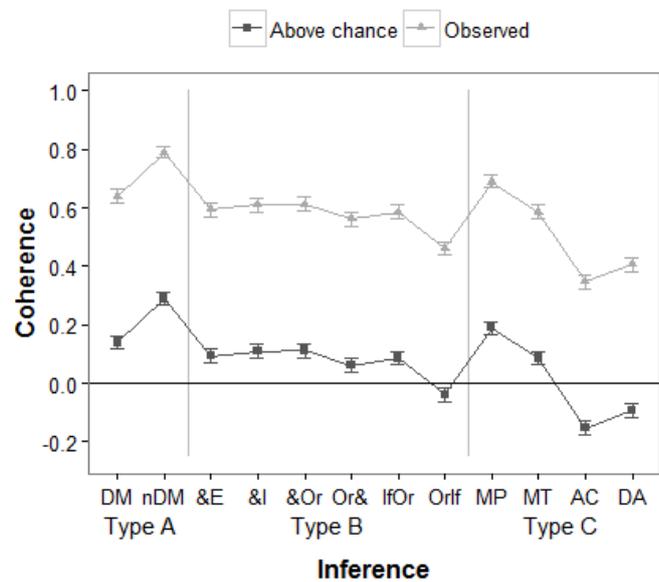
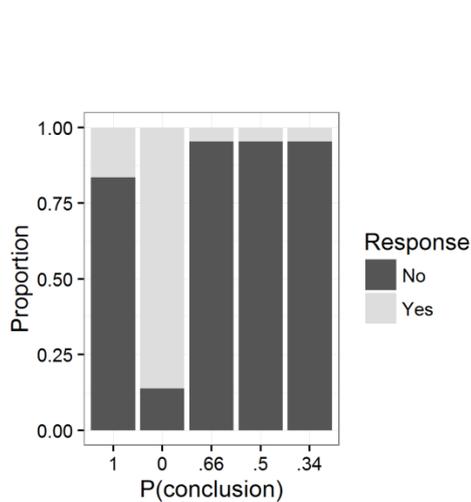


Figure 6.2. Proportion of "yes" and "no" responses to the inference p , *therefore not-p* with a premise probability of 1, observed during the practice trials.

Figure 6.3. Observed and above-chance coherence for the 12 inferences investigated in Experiment 5. Error bars show 95% CIs.

Overall above-chance coherence

Following the same procedure as in the previous experiments, a "yes" response was classified as coherent when the conclusion probability was in the relevant coherence interval, and a "no" response was classified as coherent when the conclusion probability was not in the relevant coherence interval. Then to obtain a first impression of the results, observed and above-chance coherence were computed for each inference across experimental conditions. The results are displayed in Figure 6.3.

One can see that the use of a constant chance rate in this experiment implied that the pattern of coherence between inferences remained the same for observed and above-chance coherence. The confidence intervals in the Figure suggest coherence was above chance levels for 9 of the 12 inferences. It seemed to be highest for nDM and MP, followed by DM, &E, &I, &Or, Or&, IfOr, and MT. In contrast, coherence appeared to be below chance levels for OrIf, AC, and DA. This rough ordering of above-chance coherence between inferences was confirmed in a series of linear mixed models. As before, the procedure for model construction was based on the recommendation of Barr et al. (2013) of including the maximum random effects structure justified by the design. Unless otherwise specified, this led to the computation of random coefficient models with random intercepts for participants and materials, whereas

attempts to include random slopes led to failure of convergence (Bates et al., 2015; Westfall et al., 2014).

Above-chance coherence was higher for nDM than for MP ($F(1, 377.461) = 28.865, p < .001$). Above-chance coherence was also higher for nDM than for DM ($F(1, 218) = 39.527, p < .001$); and it was higher for MP than for MT ($F(1, 218) = 18.425, p < .001$). Among the inferences of type B, above-chance coherence did not differ between inferences &E, &Or, IfOr, &I, and Or& ($F(4, 1256.218) = 1.860, p = .115$). Above-chance coherence was higher for DM than for MT ($F(1, 459.045) = 9.267, p = .002$). But there was no difference in above-chance coherence between MT on the one hand, and inferences &E, &Or, IfOr, &I, and Or& of type B on the other ($F(1, 5836.966) = .614, p = .433$). Further, as suggested by Figure 6.3, coherence was at chance levels for OrIf ($F(1, 109) = 2.995, p = .086$); and coherence was below chance levels for AC and DA ($F(1, 218) = 47.478, p < .001$). There was no significant difference in above-chance coherence between AC and DA ($F(1, 218) = 2.836, p = .094$).

In summary, coherence was above-chance levels for 9 of the 12 inferences. It was highest for the contradiction nDM, followed by MP, followed by the equivalence DM. Above-chance coherence was slightly lower for MT and inferences &E, &Or, IfOr, &I, and Or& of type B. Coherence did not differ from chance levels for OrIf, and it was below chance for AC and DA.

The finding that coherence was clearly above chance levels for most inferences investigated suggests that the cases of OrIf, AC and DA in which coherence was not above chance are not indicative of a general lack of sensitivity to coherence constraints. A better explanation for the absence of above-chance coherence in these cases may be found by looking at more specific features of the inferences involved.

Responses to OrIf were found to be coherent above chance levels in Experiments 1, 3 and 4, as well as in Politzer & Baratgin (2016). The chance-level coherence in this experiment is therefore not a stable finding. Further, the finding cannot be explained through the assumption that the materials used in the experiment suggested a biconditional interpretation of the conditional in the conclusion of the inference, because OrIf is invalid for both interpretations of the conditional. To see why, consider that the biconditional *p if and only if q* is equivalent to a conjunction of the two conditionals *if p then q* and *if q then p*. This means that by and-elimination, the inference from the biconditional to the conditional is valid, and so $P(p \text{ if and only if } q) \leq P(\text{if } p \text{ then } q)$. Hence knowing that OrIf is invalid for a conditional interpretation, $P(p \text{ or } q) \geq P(\text{if } p \text{ then } q)$ it follows that it will also be invalid for a biconditional interpretation, $P(p \text{ or } q) \geq P(p \text{ if and only if } q)$.

A tentative explanation for why coherence was at chance levels for OrIf in this experiment, whereas it was above chance levels in other experiments, is that it reflects a negation effect similar to that suggested in Experiments 3 and 4 for this inference: a negated antecedent in the conditional of the conclusion may make it more difficult to perform a Ramsey test, and any result of the test may be more difficult to relate to the situation in the

premise, in which the statement corresponding to the antecedent of the conditional is being affirmed rather than negated. It may be that such negation effects are more frequent for some types of materials than for others. However, it is important to bear in mind that this is just one possible explanation proposed after observing the data, and follow-up experiments would be needed to test to what extent it can be upheld and acquire predictive value.

The findings for OrIf differ from those of AC and DA. For AC and DA, responses did not merely fail to be coherent above chance levels, but were significantly below chance levels. Moreover, as for Experiments 3 and 4, the same responses that were classified as incoherent in the present analysis would have been classified as coherent under the assumption that the conditional in their major premise was interpreted as a biconditional. A biconditional interpretation could result when the materials used in the inferences suggest that there is a positive correlation between the antecedent and consequent. Evidence for a biconditional interpretation has been observed in previous studies at least for a subset of participants (Baratgin et al., 2013; Barrouillet & Gauffroy, 2015; Skovgaard-Olsen et al., 2016; but see Oberauer, Weidenfeld, et al., 2007; Over et al., 2007; Singmann et al., 2014, for findings of a conditional as opposed to biconditional interpretation of conditionals), and such an interpretation would seem reasonable in cases in which the conditional expresses a causal relation (Oaksford & Chater, 2017; Over, 2017).

Because there is a plausible interpretation of the finding of below-chance coherence for AC and DA as reflecting not a genuine lack of sensitivity to coherence, but instead an alternative interpretation of the sentences contained in it – and this alternative interpretation is backed by previous findings using different tasks and materials – AC and DA were not included in the subsequent analysis. The subsequent analysis looks at factors that may affect the frequency with which responses are coherent above chance levels for different inferences. If AC and DA were included in it, then effects involving them (e. g. the question of differences in coherence between inferences of types A, B, and C) would be difficult to interpret because it would not be clear whether they really reflect differences in above-chance coherence, or instead a mismatch between participants' interpretation of the conditional premises and the interpretation upon which the computation of coherence was based. The results for AC and DA were nonetheless depicted in Figures 6.4 and 6.5. In this way the information for them is not lost and it can be related to that for the other inferences.

In contrast, the results for OrIf were included in the analysis. These results can more readily be seen as a genuine failure to take coherence constraints into account, at least in cases in which the inferences in question contain negations.

Edge vs. clear

The following two sections analyse how above-chance coherence was affected by whether the probability of the conclusion was at the edge of the coherence interval or clearly on one side of it, and by whether the probability of the conclusion was inside or outside the interval.

The conclusion probabilities for the equivalence of DM and the contradiction of nDM that were "inside" the point-value coherence interval, were necessarily "at the edge", with no distinction between an upper and a lower edge. When the probability of the premise was 1, the interval for DM generally had only a lower edge and the interval for nDM only an upper edge. The intervals for the valid inferences of type B only had a lower edge, whereas the intervals for the invalid inferences of type B only had an upper edge. Similarly, the intervals for MT only had a lower edge, and those for AC only had an upper edge. Because of these differences between inferences, the question of whether the edge was on the upper or lower end of a coherence interval was not taken into account in this analysis.

A further point to consider is that by the definition of the intervals for DM, nDM, and AC, as well as for the remaining inferences in the conditions in which their interval was a point value, the question of whether a conclusion probability was "on the edge vs. clearly on one interval side" cannot be crossed with the question of whether the conclusion probability was "inside or outside" the interval, because not all combinations of the values of the two variables exist. For example, for DM and nDM, conclusion probabilities that were inside the interval were always at the interval edge. For this reason, separate analyses were conducted to test the effects of the two factors.

Let us call the distinction between a case in which the conclusion probability was at the edge of a coherence interval, and a case in which the conclusion probability was clearly on one side of the interval (whether this was inside or outside of it), the variable *edge-clear*. The pattern of results for this variable is displayed in Figure 6.4, separately for each inference and premise probability condition. Note that the maximum of the scale is .5 because given the chance rate of .5, this is the maximum possible value of above-chance coherence, corresponding to the responses of a hypothetical maximally coherent person.

A linear mixed model for the effects of edge-clear, inference type, and premise probability on above-chance coherence (excluding AC and DA) showed that overall coherence was 14% higher than expected by chance ($F(1, 752.818) = 325.039, p < .001$). Above-chance coherence differed between inference types ($EMM_A = .229, EMM_B = .077, EMM_C = .121, F(2, 9005.179) = 73.700, p < .001$); and between premise probability conditions ($EMM_5 = .117, EMM_8 = .115, EMM_I = .204, F(2, 15668.488) = 77.334, p < .001$). Above-chance coherence was also found to be higher when the probability of the conclusion was clearly on one side of the interval than when it was on the interval edge ($EMM_{edge} = .118, EMM_{clear} = .173, F(1, 15678.578) = 35.229, p < .001$).

However, these main effects were qualified by a series of interactions. The effect of premise probability differed between inference types ($F(4, 15668.739) = 14.511, p < .001$). The effect of edge-clear also varied between inference types ($F(2, 15679.852) = .27.417, p < .001$). No other effects were significant (highest $F = 1.374$, lowest $p = .253$).

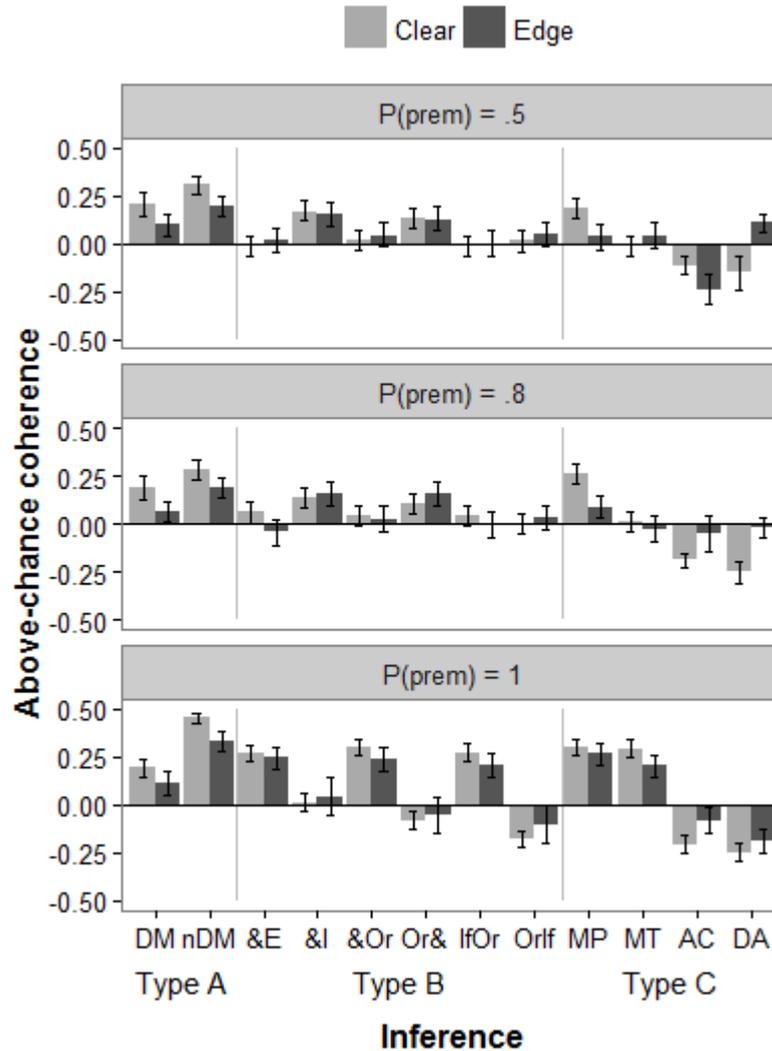


Figure 6.4. Above-chance coherence for Experiment 5 as a function of premise probability and whether the probability of the conclusion was at the edge of the coherence interval or clearly on one side of it. Error bars show 95% CIs.

Type A inferences. A follow-up analysis within each inference type showed that for inferences of type A, above-chance coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the interval edge ($EMM_{\text{edge}} = .165, EMM_{\text{clear}} = .272, F(1, 3052) = 52.392, p < .001$). Above-chance coherence also differed between premise probability conditions ($F(2, 3052) = 14.561, p < .001$); however, this effect was qualified by an interaction between premise probability and inference

($F(1, 3052) = 9.681, p < .001$). Premise probability had no effect for the equivalence of DM ($EMM_{.5} = .151, EMM_{.8} = .128, EMM_1 = .154, F(2, 1526) = .529, p = .589$). Premise probability did have an effect for the contradiction of nDM ($EMM_{.5} = .248, EMM_{.8} = .236, EMM_1 = .394, F(2, 1526) = 28.071, p < .001$): In line with the pattern in Figure 6.4, above-chance coherence for this inference was larger with certain than with uncertain premise probabilities (for 1 vs. .8, $F(1, 981) = 49.398, p < .001$; for 1 vs. .5, $F(1, 981) = 44.080, p < .001$; for .8 vs. .5, $F(1, 981) = .194, p = .660$).

Overall, for both DM and nDM, above-chance coherence was larger when the probability of the conclusion was clearly on one side of the interval than when it was at the edge. In addition, for nDM, but not for DM, above-chance coherence was larger when the probability of the premise was certain than when it was uncertain.

Type B inferences. A follow-up analysis for inferences of type B, differentiating between valid and invalid inferences, showed that above-chance coherence was higher for the valid than for the invalid inferences ($EMM_{valid} = .101, EMM_{invalid} = .049, F(1, 2815.637) = 11.912, p = .001$). Above-chance coherence differed between premise probability conditions ($EMM_{.5} = .061, EMM_{.8} = .064, EMM_1 = .101, F(2, 9278.382) = 6.923, p = .001$); but there was no main effect of edge-clear ($EMM_{edge} = .075, EMM_{clear} = .075, F(1, 9278.382) = .001, p = .978$).

However, the above main effects were qualified by several interactions. The effect of premise probability differed for valid and invalid inferences ($F(2, 9278.382) = 189.342, p < .001$); as did the effect of edge-clear ($F(1, 9278.382) = 8.488, p = .004$). These effects were in turn qualified by a three-way interaction between validity, premise probability and edge-clear ($F(2, 9278.382) = 4.153, p = .016$).

Follow-up analyses to the three-way interaction indicated that for the valid inferences, above-chance coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the interval edge ($F(1, 4611.027) = 4.925, p = .027$). However, above-chance coherence also varied between premise probability conditions ($F(2, 4611.027) = 164.708, p < .001$), and there was an interaction between edge-clear and premise probability ($F(2, 4611.027) = 4.369, p = .013$).

In line with the pattern in Figure 6.4, coherence was at chance levels when premise probability was .5 ($EMM = .007, F(1, 291.042) = .164, p = .686$), and this finding did not differ between inferences ($F(2, 556.795) = .641, p = .527$). No other effects were significant (highest $F = .641$, lowest $p = .527$).

When premise probability was .8, coherence was again at chance levels ($EMM = .020, F(1, 292.241) = 1.257, p = .263$). This result again did not differ between inferences ($F(2, 576.969) = .395, p = .674$). However, in spite being at chance levels in general, coherence nonetheless tended to be higher when the probability was clearly on one side of the interval

than when it was at the interval edge ($F(1, 1338.246) = 5.856, p = .016, X^2(3) = 8.354, p < .05$)¹⁰. No other effects were significant (highest $F = 1.262$, lowest $p = .283$).

Finally, when premise probability was 1, responses were clearly coherent above chance levels ($EMM = .254, F(1, 295.596) = 223.439, p < .001$). This result did not differ between inferences ($EMM_{\&E} = .264, EMM_{\&Or} = .257, EMM_{IfOr} = .240, F(2, 696.544) = .290, p = .749$). Above-chance coherence was again higher when the probability of the conclusion was clearly on one side of the interval than when it was at the interval edge ($EMM_{edge} = .227, EMM_{clear} = .280, F(1, 1341.916) = 9.055, p = .003$). No other effects were significant (highest $F = .533$, lowest $p = .587$).

Overall, we see that for the valid inferences of type B, above-chance coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the edge, and this effect did not differ between the three inferences. However, coherence was only reliably above chance levels in the first place when the probability of the premise was certain. Coherence did not differ from chance levels in the two conditions with uncertain premise probabilities.

An analysis for the invalid inferences of type B showed that above-chance coherence differed between premise probability conditions ($EMM_{.5} = .106, EMM_{.8} = .100, EMM_{.1} = -.060, F(2, 4608.851) = 56.256, p < .001$). Above-chance coherence also differed between inferences ($EMM_{\&I} = .119, EMM_{Or\&} = .057, EMM_{OrIf} = -.030, F(2, 1044.578) = 15.786, p < .001$). Finally, there was a non-reliable trend for above-chance coherence to be higher when the probability of the conclusion was at the edge of the interval than when it was clearly on one side of it ($EMM_{edge} = .062, EMM_{clear} = .035, F(1, 4608.851) = 4.013, p < .045, X^2(9) = 7.087, p > .250$). No other effects were significant (highest $F = .784$, lowest $p = .457$).

Follow-up analyses to the effect of premise probability showed that for the premise probability condition of .5, overall coherence was above chance levels ($EMM = .106, F(1, 285.153) = 51.123, p < .001$), but the extent to which this was the case differed between inferences ($EMM_{\&I} = .159, EMM_{Or\&} = .128, EMM_{OrIf} = .030, F(2, 508) = 8.135, p < .001$). It was above-chance levels for &I ($F(1, 113.737) = 36.946, p < .001$) and for Or& ($F(1, 114.150, p < .001$), but not for OrIf ($F(1, 114.934) = 1.577, p = .212$). There was no main effect of edge-clear ($F(1, 1334.317) = .031, p = .861$) nor an interaction between edge-clear and inference ($F(2, 1334.317) = .497, p = .609$).

A similar pattern was observed for the premise probability condition of .8: overall coherence was again above chance levels ($F(1, 301.428) = 44.758, p < .001$) but the extent of the effect differed between inferences ($EMM_{\&I} = .149, EMM_{Or\&} = .133, EMM_{OrIf} = .022, F(2, 537.852) = 8.446, p < .001$). It was above chance levels for &I ($F(1, 113.761) = 32.349, p <$

¹⁰ Here the more robust likelihood-ratio test was used to try to disambiguate an instance of a discrepancy between a significant F test for the fixed effect, and a non-significant t -test for the parameter (the beta-value of the effect). This can happen because of the different distributional assumptions of the two tests, especially when the number of data points is relatively small, as in the present comparison.

.001) and for Or& ($F(1, 113.809) = 24.811, p < .001$), but was at chance levels for OrIf ($F(1, 116.684) = .670, p = .415$). There was again no main effect of edge-clear ($F(1, 1349.676) = 2.240, p = .135$) nor an interaction between edge-clear and inference ($F(2, 1349.676) = .105, p = .900$).

Finally, in the premise probability condition of 1, overall coherence was below chance levels ($F(1, 342.521) = 7.155, p = .008$), but this result again differed between inferences ($EMM_{\&I} = .031, EMM_{Or\&} = -.072, EMM_{OrIf} = -.143, F(2, 1019.623) = 8.555, p < .001$). Overall coherence did not differ from chance levels for &I ($F(1, 126.833) = .480, p = .490$) nor for Or& ($F(1, 125.562) = 2.709, p = .102$), but coherence was below chance levels for OrIf ($F(1, 129.363) = 14.670, p < .001$). There was again no significant effect of edge-clear ($F(1, 1339.323) = 3.748, p = .053$) nor an interaction between edge-clear and inference ($F(2, 1339.323) = .410, p = .664$).

Overall, for the invalid inferences &I and Or& coherence was above chance levels when the probability of the premise was uncertain, but was at chance levels when the probability of the premise was certain. The effect of premise probability for OrIf was similar, but coherence was lower for this inference: it was at chance levels when premise probability was uncertain, and below chance levels when premise probability was certain. No other effects were significant. In particular, it made no difference for above-chance coherence whether the probability of the premise was at the edge of the interval or clearly on one side of it.

Type C inferences. A follow-up analysis for inferences MP and MT of type C showed that above-chance coherence differed between premise probability conditions ($EMM_{.5} = .060, EMM_{.8} = .084, EMM_{1} = .266, F(2, 3052) = 66.946, p < .001$); and it was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the edge ($EMM_{edge} = .101, EMM_{clear} = .172, F(1, 3052) = 19.859, p < .001$). However, the above main effects were qualified by three interactions. The size of the effect of edge-clear was larger for MP than for MT ($F(1, 3052) = 9.473, p = .002$). The extent to which above-chance coherence was larger for MP than for MP (reported in the previous section on overall above-chance coherence) differed between premise probability conditions ($F(2, 3052) = 7.290, p = .001$). And there was a three-fold interaction between edge-clear, inference, and premise probability condition ($F(2, 3052) = 6.208, p = .002$).

In follow-up analyses for each premise probability condition, the model failed to converge for the case of a premise probability of .5 when random effects were included. Therefore the analyses for the three premise probability conditions were conducted using only fixed effects.

For the premise probability condition of .5, overall coherence was above chance levels ($F(1, 1090) = 15.565, p < .001$), but the extent to which this was the case was higher for MP than for MT ($F(1, 1090) = 10.479, p = .001$). There was no main effect of edge-clear ($F(1, 1090) = 2.303, p = .129$); but edge-clear interacted with inference ($F(1, 1090) = 12.536, p <$

.001). Separate analyses for each inference showed that for MP, overall coherence tended to be above chance levels ($EMM = .109$, $F(1, 545) = 26.922$, $p < .001$); and it was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the edge ($EMM_{edge} = .032$, $EMM_{clear} = .185$, $F(1, 545) = 13.351$, $p < .001$). In contrast, for MT coherence was not above chance levels ($EMM = .011$, $F(1, 545) = .241$, $p = .624$); and it made no difference whether the probability of the conclusion was at the edge of the interval or clearly on one side of it ($EMM_{edge} = .041$, $EMM_{clear} = -.020$, $F(1, 545) = 1.964$, $p = .162$).

A follow-up analysis for the premise probability condition of .8 showed that overall coherence was above chance levels ($F(1, 1090) = 31.361$, $p < .001$); but was again higher for MP than for MT ($F(1, 1090) = 36.728$, $p < .001$). Above-chance coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the edge ($F(1, 1090) = 13.00$, $p < .001$); but this effect was qualified by an interaction between edge-clear and inference ($F(1, 1090) = 4.949$, $p = .026$). Separate analyses for each inference revealed the same pattern as for a premise probability of .5: For MP, coherence was above chance levels ($EMM = .174$, $F(1, 545) = 72.891$, $p < .001$); and it was higher when the probability of the conclusion was clearly on one side of the interval than when it was on its edge ($EMM_{edge} = .087$, $EMM_{clear} = .261$, $F(1, 545) = 18.223$, $p < .001$). In contrast, for MT coherence was at chance levels ($EMM = -.007$, $F(1, 545) = .099$, $p = .753$); and there was no effect of edge-clear ($EMM_{edge} = -.028$, $EMM_{clear} = .014$, $F(1, 545) = .893$, $p = .345$).

A follow-up analysis for the premise probability condition of 1 showed that overall coherence was above chance levels ($EMM = .266$, $F(1, 1090) = 421.890$, $p < .001$). The extent to which coherence was above chance levels did not differ between inferences ($EMM_{MP} = .282$, $EMM_{MT} = .249$, $F(1, 1090) = 1.615$, $p = .204$). Above-chance coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on its edge ($EMM_{edge} = .236$, $EMM_{clear} = .295$, $F(1, 1090) = 5.179$, $p = .023$); and the size of this edge-clear effect also did not differ between inferences ($F(1, 1090) = 1.070$, $p = .301$).

Overall, when the probability of the premises was certain, above-chance coherence was equally high for MP and MT; and it was higher when the probability of the conclusion was clearly on one side of the interval than when it was on its edge. However, coherence differed for the two inferences when the probability of the premises was uncertain. With uncertain premises, coherence for MP remained above chance levels, and it remained higher when the probability of the conclusion was clearly on one side of the interval than when it was on its edge. In contrast, when premise probability was uncertain, coherence for MT was no longer above chance levels, and it no longer made a difference whether the probability of the conclusion was on the edge of the interval or clearly on one side of it.

Discussion. Taken together, the results illustrated in Figure 6.4 provide a number of new insights into quantitative questions about above-chance coherence for the 10 inferences investigated. Although responses to the 10 inferences were coherent above chance levels

when averaging across predictors, the extent to which this was the case differed strongly between inferences and between premise probabilities. Coherence was always above chance levels for the inferences of DM and nDM of type A, and for MP, in line with the finding reported in the first section of the results that when averaging across predictors, above-chance coherence was highest for nDM and MP. For the remaining inferences, above-chance coherence seemed to depend strongly on whether premise probability was certain or uncertain. For nDM, coherence was higher when the probability of the premise was certain, and for inferences &E, &Or, IfOr, and MT coherence was only above chance levels when the probability of the premise was certain. In sharp contrast, for &I and Or& coherence was only above chance levels when the premise was uncertain, while for OrIf coherence was at or below chance levels across conditions.

One possible explanation for this pattern is as follows. For the equivalence and contradiction of type A, whose coherence interval is always a point value, the probability of the premise has no effect on interval width, and therefore made no difference for participants' ability to construct a coherent response. For the inferences of type B and C, for which the coherence interval is generally an area within the probability range whose location is codetermined by premise probability, participants tended to be somewhat careful or conservative in their judgments, not endorsing a high probability for a conclusion unless they were highly confident in it.

For the valid type B inferences in this experiment, a tendency to be careful or conservative would increase the number of responses that are "too low", i. e. that fall below the coherence interval, except when the probability of the premise is 1. In this special case in which there is certainty about the premises, equally high conclusion probabilities are more likely to be viewed as justified.

For the invalid type B inferences in this experiment, the same tendency to be careful or conservative would lead to more frequent coherent responses for uncertain premise probabilities because in these cases the coherence intervals lie lower. In contrast, it would lead to more frequent incoherent responses for a premise probability of 1, because in this case people may more often misclassify higher conclusion probabilities as being too high, when in fact any conclusion probability is coherent.

A tendency to be careful or conservative would arguably makes sense in everyday life, where we often have to make judgments on the basis of little information, because it would help us avoid jumping to conclusions and making unwarranted generalisations. However, it would be premature to draw any conclusion about a general tendency to be cautious or conservative on the basis of the present findings. This is because the materials used in the experiment were explicitly designed to render participants' judgments careful or conservative, as part of an effort to create a context in which the distinction between valid and invalid

inferences becomes relevant. It seems more plausible at this stage to conclude that the manipulation in the materials was successful.

Whatever reason participants may have had to give conservative responses, the observed pattern of differences between inferences provides no evidence that a tendency to be careful or conservative was associated with a higher sensitivity to the difference between induction and deduction, expressed through a difference in coherence between valid and invalid inferences. Leaving aside AC and DA, whose results are difficult to interpret, coherence was higher for the valid inferences &E, &Or and IfOr than for the invalid inferences &I, Or&, and OrIf. But coherence was also higher for the invalid inference of nDM than for the valid inference of DM.

The pattern of results in this and the previous analysis also provides no evidence of a general difference in the degree of above-chance coherence between inference types: Coherence was most reliably above chance levels for DM, nDM, and MP, which differ in the type they were classified into. However, the present analysis does suggest a contrast between inferences of type A on the one side, and inferences of type B and C on the other, regarding the factors that affect above-chance coherence for them. In particular, coherence for the equivalence and contradiction (type A) was less influenced by differences in premise probability than coherence for the one-premise inferences describing set-subset relations (type B), or the more complex two-premise inferences (type C). Differences like this may provide information about the algorithms involved in processing these inferences, but much more research would have to be carried out to be able to pinpoint them.

Regarding the effect of edge-clear, the results show that for 7 of the 10 inferences investigated (DM, nDM, &E, &Or, IfOr, MP, and MT), coherence was higher when the probability of the conclusion was clearly on one side of the interval than when it was on the edge, provided that above-chance coherence was high enough to allow the measurement of relative differences within it. For the remaining three inferences (&I, &Or, and OrIf), no effect of edge-clear was observed. Thus, there was evidence for an effect of edge-clear in the expected direction for the majority of the conditions tested, but not for all conditions. A possible reason for why it was not observed for the invalid type B inferences &I, &Or, and OrIf may be that in the premise probability conditions in which coherence was above chance levels for these inferences, the coherence interval was very wide, and wide intervals may reduce the extent to which it is informative to know whether a conclusion probability is at the edge or clearly on one side. However, this is just a potential explanation for the finding, which would have to be investigated further.

The overall corroboration of the prediction that coherence would be higher when the probability of the conclusion is clearly on one side of the interval than when it is at the interval edge is evidence that the observed effect of coherence is real: that it actually reflects a sensitivity at some level to the constraints of coherence, as opposed to being a coincidental by-

product of some unrelated aspect of reasoning. If people are sensitive to coherence, but have degrees of belief that are less precise than a point probability, then one would naturally expect their coherence performance to change as a function of the location of interval boundaries.

The following section examines a further factor that may play a role for the degree to which participants' responses were coherent above chance levels: whether the probability of the conclusion was inside or outside the interval.

Inside vs. outside

Let us call the effect on above-chance coherence of whether a conclusion probability was inside or outside the coherence interval, the *inside-outside* effect. The pattern of results for this effect is displayed in Figure 6.5, separately for each premise probability condition. As in the previous analysis, the value of .5 on the y axis is the maximum possible value of above-chance coherence, which would be obtained by a hypothetical perfectly coherent person. The figure illustrates that by the definition of their intervals, the probability of the conclusion of the invalid type B inferences (&I, Or&, and OrIf) could not lie outside of the interval when the probability of the premise was 1, because in this case the interval is the probability range. This implies that unlike the analysis of the effect of edge-clear, the effect of inside-outside cannot be fully crossed with premise probability. For this reason, the effect of inside-outside was assessed in two separate analyses. The first was for when premise probability was uncertain, and included the same 10 inferences as in the previous section. The second analysis was for when premise probability was certain, and included the seven inferences DM, nDM, &E, &Or, IfOr, MP, and MT. The latter are the inferences of type A plus the valid inferences of types B and C.

Uncertain premises. A linear mixed model was conducted for the effects of inside-outside, inference type, and premise probability (.5, .8) on above-chance coherence. Given that the independent effects of premise probability and inference type were already examined in the previous two sections, this section reports only effects in which inside-outside was involved. Above-chance coherence was higher when the probability of the conclusion was outside rather than inside the interval ($F(1, 10331.241) = 90.320, p < .001$). However, the extent to which inside-outside had an effect differed between inference types ($F(2, 10341.304) = 120.124, p < .001$). This interaction was in turn qualified by a marginal three-way interaction between inside-outside, inference type, and premise probability ($F(2, 10227.772) = 3.281, p = .038$).

Type A inferences. Follow-up analyses showed that for the inferences of type A, overall coherence was 23% higher than expected by chance ($F(1, 351.519) = 229.918, p < .001$). Above-chance coherence was higher when the probability of the conclusion was inside rather than outside the interval ($EMM_{inside} = .314, EMM_{outside} = .146, F(1, 1962) = 51.069, p < .001$).

The effect of inside-outside did not differ between premise probability conditions ($F(1, 1962) = 1.933, p = .165$).

Type B inferences. For the inferences of type B overall coherence was around 9% higher than expected by chance ($F(1, 497.139) = 63.767, p < .001$). In contrast to the pattern for the inferences of type A, above-chance coherence was higher when the probability of the conclusion was outside rather than inside the interval ($EMM_{outside} = .203, EMM_{inside} = -.032, F(1, 6001.636) = 450.345, p < .001$). The effect of inside-outside was higher for the invalid than for the valid inferences ($F(1, 601.636) = 26.878, p < .001$).

Figure 6.5 suggests that the higher effect of inside-outside for the invalid inferences results from the fact that coherence was higher for the invalid than for the valid inferences when it was above chance levels, and that it was above chance levels only in the condition in which the probability of the conclusion was outside the interval. Follow-up analyses corroborated this pattern. For the valid inferences &E, &Or and IfOr, overall coherence was around chance levels ($EMM = .032, F(1, 292.717) = 4.098, p = .044$). Coherence was higher when the conclusion probability was outside rather than inside the interval ($EMM_{inside} = -.056, EMM_{outside} = .121, F(1, 2974.381) = 131.182, p < .001$), and inside-outside did not interact with premise probability ($F(1, 2974.381) = 1.327, p = .249$). A separate analysis for the inside and the outside condition confirmed coherence to be above chance levels only in the latter (outside condition: $EMM = .124, F(1, 286.962) = 36.083, p < .001$. Inside condition: $EMM = -.059, F(1, 290.187) = 6.810, p = .010$).

For the invalid type B inferences &I, Or&, and OrIf, overall coherence was around 13% higher than expected by chance ($F(1, 293.156) = 87.832, p < .001$). Coherence was again higher when the conclusion probability was outside rather than inside the interval ($EMM_{inside} = -.013, EMM_{outside} = .279, F(1, 2974.315) = 373.288, p < .001$), and inside-outside did not interact with premise probability ($F(1, 2974.315) = .256, p = .613$). A separate analysis for the inside and outside conditions again showed that coherence was above chance levels only in the latter (Outside condition: $EMM = .281, F(1, 284.621) = 235.086, p < .001$. Inside condition: $EMM = -.015, F(1, 288.747) = .499, p = .480$).

Type C inferences. Finally, an analysis for inferences MP and MT of type C showed that overall coherence was 7% higher than expected by chance ($F(1, 270.934) = 26.107, p < .001$). As for the inferences of type B, coherence was higher when the conclusion probability was outside rather than inside the interval ($EMM_{inside} = -.050, EMM_{outside} = .181, F(1, 1962) = 114.309, p < .001$). The size of the effect of inside-outside was higher when premise probability was .5 than when it was .8 ($F(1, 1962) = 7.811, p = .005$).

Follow-up analyses showed that for MP and conclusion probabilities outside the interval, coherence was above chance levels ($EMM = .217, F(1, 111.056) = 92.169, p < .001$); and was higher when premise probability was .5 than when it was .8 ($EMM_{.5} = .252, EMM_{.8} = .181, F(1, 654) = 5.437, p = .020$). For MP and conclusion probabilities inside the interval,

coherence did not differ from chance ($EMM = -.005$, $F(1, 120.267) = .016$, $p = .899$). Coherence was higher when premise probability was .8 than when it was .5 ($EMM_{.5} = -.069$, $EMM_{.8} = .060$, $F(1, 218) = 7.232$, $p = .008$), but coherence was still at chance levels in the premise probability condition of .8 ($F(1, 109) = 1.573$, $p = .212$; this effect was calculated using only fixed predictors because the lower sample size would otherwise have led to failure of convergence).

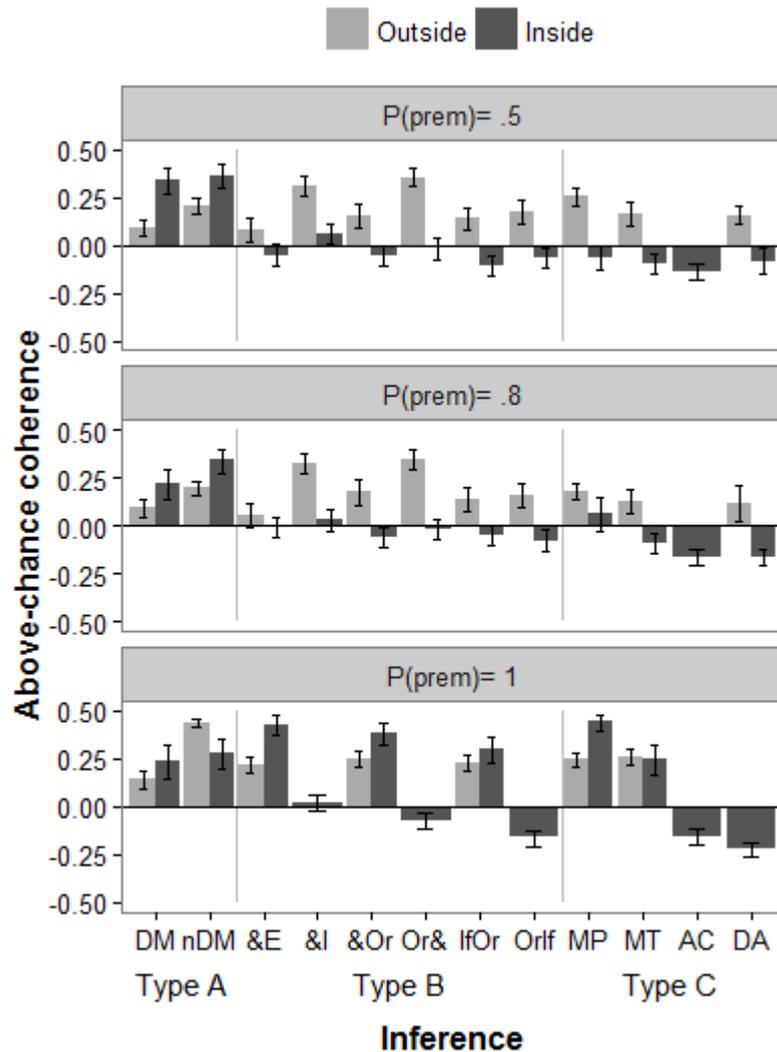


Figure 6.5. Above-chance coherence for Experiment 5 as a function of premise probability and whether the probability of the conclusion was inside or outside of the interval. Error bars show 95% CIs.

For MT and conclusion probabilities outside the interval, overall coherence was above chance levels ($EMM = .144$, $F(1, 109) = 19.212$, $p < .001$), and did not differ between premise probability conditions ($EMM_{.5} = .161$, $EMM_{.8} = .128$, $F(1, 327) = .763$, $p = .383$). For MT and conclusion probabilities inside the interval, coherence was below chance ($EMM = -.095$, $F(1,$

109) = 7.708, $p = .006$), and again did not differ between premise probability conditions ($EMM_{.5} = -.099$, $EMM_{.8} = -.090$, $F(1, 545) = .101$, $p = .751$).

In summary, for inferences DM and nDM of type A, above-chance coherence was larger when the probability of the conclusion was inside the interval. In contrast, for the inferences of type B, as well as for MP and MT of type C, above-chance coherence was larger, and in fact only above chance levels, when the probability of the conclusion was outside the interval.

These findings are in accordance with those of the analysis of edge-clear, and can also be explained through the hypothesis that participants tended to be careful or conservative in their responses. A tendency for participants to respond "no, this conclusion probability is not consistent with the probability of the premise(s)" unless they can be highly confident that it is consistent, would lead to increased performance in cases in which the probability of the conclusion was outside the interval, and decreased performance in cases in which it was inside the interval. As for the analysis of edge-clear, such a tendency seemed to be present for all inferences except those of type A, whose coherence interval is always a point value regardless of premise probability.

Certain premises. A second linear mixed model analysis for the effects of inside-outside and inference type on above-chance coherence was performed for the seven inferences DM, nDM, &E, &Or, IfOr, MP, and MT in the condition in which premise probability was certain. Overall coherence was on average 29% higher than expected by chance ($F(1, 704.015) = 543.347$, $p < .001$). Contrary to the finding for uncertain premises, above-chance coherence was higher when the probability of the conclusion was inside rather than outside the interval ($EMM_{inside} = .322$, $EMM_{outside} = .256$, $F(1, 3263.176) = 20.715$, $p < .001$). The size of the effect of inside-outside differed between inference types ($F(2, 3263.176) = 12.693$, $p < .001$).

A follow-up analysis to the interaction showed that for inferences of type A, overall coherence was above chance levels ($EMM = .274$, $F(1, 299.331) = 207.002$, $p < .001$). There was no main effect of inside-outside ($F(1, 872) = 1.958$, $p = .162$), but inside-outside interacted with inference ($F(1, 872) = 25.379$, $p < .001$). For DM, above-chance coherence was larger when the probability of the conclusion was inside rather than outside the interval ($EMM_{inside} = .234$, $EMM_{outside} = .144$, $F(1, 436) = 5.028$, $p = .025$). In contrast, for nDM, above-chance coherence was larger when the probability of the conclusion was outside rather than inside the interval ($EMM_{inside} = .280$, $EMM_{outside} = .438$, $F(1, 436) = 30.309$, $p < .001$). Looking back at Figure 6.5, this means that whereas the effect of inside-outside remained constant across premise probability conditions for DM, it reversed for nDM between the conditions with uncertain and with certain premise probability.

An analysis for inferences &E, &Or, and IfOr of type B showed that overall coherence was above chance levels ($EMM = .300$, $F(1, 376.419) = 275.954$, $p < .001$). Above-chance coherence tended to be higher when the probability of the conclusion was inside the interval ($EMM_{inside} = .368$, $EMM_{outside} = .232$, $F(1, 1341.842) = 40.905$, $p < .001$), but the size of the

effect of inside-outside differed between inferences ($F(2, 1341.842) = 3.619, p = .027$). Follow-up analyses to this interaction showed that the effect of inside-outside was significant for &E and for &Or (for &E: $EMM_{inside} = .427, EMM_{outside} = .218, F(1, 436) = 36.479, p < .001$; for &Or: $EMM_{inside} = .381, EMM_{outside} = .250, F(1, 436) = 12.764, p < .001$), but marginally failed to reach significance for IfOr ($EMM_{inside} = .298, EMM_{outside} = .229, F(1, 436) = 3.148, p = .077$).

For MP and MT of type C, overall coherence was above chance levels ($EMM = .300, F(1, 291.204) = 209.502, p < .001$). Above-chance coherence was higher when the probability of the conclusion was inside rather than outside the interval ($F(1, 872) = 14.095, p < .001$), but the extent of this effect was larger for MP than for MT ($F(1, 872) = 16.180, p < .001$). A further examination of the interaction revealed that the effect of inside-outside was present for MP ($EMM_{inside} = .445, EMM_{outside} = .245, F(1, 436) = 33.997, p < .001$), but not for MT ($EMM_{inside} = .252, EMM_{outside} = .259, F(1, 436) = .032, p = .857$).

Discussion. Overall, the effect of inside-outside differed markedly between the conditions of uncertain and of certain premise probability. For DM and nDM of type A, where the probability of the conclusion is always a point value equal either to the probability of the premise or to its complement, above-chance coherence was generally higher when the probability of the conclusion was inside (i. e. when it was on) the interval. The only exception to this was observed for nDM when premise probability was 1, where the effect of inside-outside reversed. There seems to be no immediate explanation for this exception, but one way to investigate it further would be to zoom in to the study of equivalences and contradictions, using more than one inference of each kind, and to implement a negations paradigm for them (Evans & Handley, 1999; Oaksford & Stenning, 1992). However, such an analysis goes beyond the scope of this project.

For inferences &E, &Or, IfOr, MP, and MT in the uncertain premise probability conditions, coherence was only above chance levels when the probability of the conclusion was outside the interval. This pattern reversed (except for MT) when premise probability was certain: here responses were coherent above chance levels in both conditions, but more so when the probability of the conclusion was inside the interval. An explanation for this contrast, as described above, is that participants had a tendency to give cautious, conservative responses, not willing to state that the probability of a conclusion was consistent with the probabilities of the premises unless they were highly confident that this was the case. In the exceptional situation of certain premise probabilities, being conservative serves no function because there is maximal information about the occurrence or non-occurrence of the events described by the premises. In this case participants may have found it easier to judge whether a given event followed from the premises than whether it did not follow (c. f. the concept of contrast classes for understanding reasoning with negations, Oaksford et al., 2000; Schroyens, Verschueren, Schaeken, & d'Ydewalle, 2000).

The finding that when premise probability was uncertain and the probability of the conclusion was inside the interval, responses were only coherent above chance levels for DM and nDM, provides additional evidence that coherence was higher and more consistent across conditions for the contradiction of nDM than for MP. Further, the finding that coherence for the equivalence and contradiction (DM and nDM) was affected less and in a different way by whether the probability of the conclusion was inside or outside the coherence interval than the remaining inferences may indicate a difference in the mechanisms involved in processing the two inference types. It may be that the detection of set-subset relations (inferences of type B) requires an ability to detect equivalences and contradictions (inferences of type A), but requires additional processing steps that are not required for understanding equivalences and contradictions. The differences in factors affecting coherence between the one-premise inferences of type B and the two-premise inferences of type C seem less marked. However, much more research would be required to be able to draw conclusions about the generality and form of differences between these inference types.

General discussion

The findings in this experiment are among the first quantitative observations that have been made about response coherence. This necessarily makes them partly explorative, and may result in the generation of more hypotheses than it was possible to test. Nonetheless, the results also provide a number of new insights on people's sensitivity to coherence and on factors affecting it.

Coherence was found to be highest for the contradiction of nDM, followed closely by MP. Coherence remained above chance levels, albeit to a lower extent, for the inferences of DM, &E, &I, &Or, Or&, IfOr, and MT. In contrast, coherence was not reliably above chance levels for OrIf, and it was below chance for AC and DA.

Coherence was more stable across conditions for the equivalence and contradiction of type A (DM and nDM) than for the other inferences, possibly because the location and width of the intervals for the former is more decoupled from premise probability.

Among the one-premise inferences describing set-subset relations (type B) and the more complex two-premise inferences (type C), coherence was higher for the valid inferences when premise probability was certain, and coherence was higher for the invalid inferences when premise probability was uncertain. However, this cannot be explained as an effect of validity, because coherence for the invalid inference of nDM was not higher for uncertain premise probability. If anything, Figures 6.4 and 6.5 suggest this inference to be more aligned with the valid inferences of type B in this regard. A simpler explanation of this pattern seems to be that it reflects a tendency to give cautious or conservative responses. Such a tendency would lead

to the observed differences for inferences of type B and C, where valid responses were associated with a higher and narrower coherence interval, and invalid responses with a lower and wider coherence interval. It would have no effect on inferences of type A, where no such association of premise probability with the location and width of the interval is present. It seems plausible to assume that participants had a tendency to give cautious responses because this was actively encouraged through the way the materials were constructed.

The construction of the materials as encouraging cautious and reflective responding was chosen with the aim of increasing the chances of finding an effect of validity if there is one. It seems implausible to assume that the difference between valid and invalid inferences is relevant in any context. But it may become more relevant in situations in which there is a higher risk associated with inferring conclusions that go beyond what is given by the premises. Such situations were operationalized as situations in which much is at stake and yet careful, cautious reasoning is called for. The pattern of results suggests that this manipulation was successful in inducing cautious responses. At the same time, there was no evidence that these responses were more or less coherent for valid than for invalid inferences. The present experiment was therefore unable to establish whether validity plays a role in reasoning over and above coherence. If it plays a differential role, then there is no evidence that this role is reflected in a difference in the coherence of people's responses to the two groups of inferences.

However, the findings provide strong support for sensitivity to coherence across a range of inferences of different types, including MP and MT and so not only one-premise inferences. This evidence was not just direct through the observation that responses were coherent at above chance levels, but also came from the effect of edge-clear: the observation that the ability to establish whether a conclusion probability is coherent was higher when this probability was clearly on one side, or clearly on the other side, of the interval than when it was on the interval edge. This finding is in line with the observation made in Experiments 3 and 4 that participants' responses were more coherent when the scale of measurement was coarser (e. g. $\pm 5\%$ of an exact probability value) even after adjusting for corresponding changes in chance rate coherence. However, in Experiments 3 and 4 this benefit was observed mainly for DM and nDM, whereas the method used to assess the precision of participants' intuitions about coherence in this experiment revealed an effect more generalised across inferences.

The fact that it is possible to measure the precision or "coarseness" of participants' degrees of belief seems to give probability theory a methodological advantage over alternative non-binary proposals (e.g. Politzer & Baratgin, 2016; Spohn, 2013), which build a theoretically predefined degree of "coarseness" higher than that of a point probability into the instrument used to measure participants' uncertain beliefs. The present findings provide evidence that people's beliefs are indeed less precise than a point probability, but they allow

this degree of precision to be explored empirically. Information on this question can then feed in to algorithmic level proposals for uncertain reasoning.

One limitation of the experiment is that no conclusions could be drawn for AC and DA because coherence was below chance levels for them, whereas it would have been above chance levels under the assumption that the conditional they contain was interpreted as a biconditional. An assessment of coherence for these inferences would require an explicit control of the correlation between p and q , in order to better pinpoint the interpretation participants are making of the conditional, and allow the computation of coherence for each interpretation.

A further limitation was that with only five conclusion probabilities for each combination of inference with premise probability, it was not always possible to capture all theoretically relevant locations of conclusion probabilities relative to the coherence interval. More conclusion probabilities would have been difficult to implement in this experiment given the strong differences in the intervals for the 12 inferences investigated. But further experiments could focus on fewer inferences and in exchange have more trials within each condition of the design. This would also make it possible to test in a meaningful way whether the incoherent responses people make for a given inference are more often over- or underestimations of their conclusion probability.

In follow-up experiments it would also be useful to assess the coherence of responses for more neutral materials that do not encourage cautious or conservative responding – or even materials that encourage liberal responding. This would make it possible to test the explanation suggested here for some of the effects found, and assess their generality.

A further way in which the present results can be complemented and extended is by assessing the coherence of people's responses to the same inferences, while varying the task instructions given.

EXPERIMENT 6: HIGHER VS. LOWER THAN THE PREMISE PROBABILITIES

The present experiment was similar to experiment 5 in that it investigated the same 12 inferences, with the same given premise probabilities and with a binary response format. However, the task given to participants differed. Participants in this experiment were given no conclusion probability to evaluate, but were instead asked to judge whether it was possible for the probability of the conclusion to lie above or below a given value. As in Experiment 5, this task was purely deductive because it could be fully answered on the basis of only knowledge of the coherence intervals for an inference.

No specific prediction was made about possible differences in coherence between the questions of whether the likelihood of the conclusion can be higher and whether it can be lower than a given value. The aim of the experiment was rather to assess differences in coherence between individual inferences, and between groups of inferences, using an alternative method of measurement.

Method

Participants

A total of 42 participants from the recruitment pool of Birkbeck, University of London completed the experiment. Seven cases were excluded because they failed at least one of two test questions aimed to make sure that they were reading the materials. The final sample consisted of 35 participants. None of them had trial reaction times of 3 seconds or less, and all indicated having at least "very good" English language skills. Participants' median age was 23 (range 18-65). The majority of participants were students, 71.4% undergraduate and 14.3% postgraduate. 8.6% stated they had finished 12th grade, 2.9% that they had a technical/applied degree, and 2.9% that they had a doctoral degree. Participants' mean rating of task difficulty was 64%.

Materials and design

The experiment investigated the same 12 inferences as Experiments 3 to 5, shown in Table 6.1 of Experiment 5. Each inference was presented three times, with different premise probabilities. The probabilities used were the same as in Experiment 5: The one-premise inferences (those of type A and B) had premise probabilities of 1, .8, and .5. The two-premise inferences (the conditional syllogisms of type C) had probabilities that were matched to those of the one-premise inferences not in their value, but in the sum of their premise uncertainty (where uncertainty = 1 – probability; Adams, 1998). With both premises of each two-premise

inference having the same probability, this implied that the two-premise inferences had premise probabilities of 1, .9, and .75. As an illustration of the relation between the probabilities for the one- and the two-premise inferences, consider the case of a two-premise inference whose premises both have a probability of .75. Then the sum of the uncertainties of the premises is $(1 - .75) + (1 - .75) = .5$, which is the same as the premise uncertainty $(1 - .5)$ of a one-premise inference with a premise probability of .5.

The three trials for each inference were presented in sequence embedded in the same context story. However, the objects and people referred to in the inferences differed between the three presentations. Each inference was randomly allocated to one of 12 context stories for each participant.

The experiment was divided into two blocks, one for the one-premise inferences and one for the two-premise inferences. The reason for this was that participants received slightly different instructions for the two inferences groups. The instructions for the one-premise inference section read:

In the following you will be shown a series of short stories. In each story an inference is drawn, and your task will be to judge whether it is possible, that is, consistent, for the conclusion of the inference to be more likely than the premise, and whether it is possible for the conclusion to be less likely than the premise.

On each trial participants were asked two questions: "Can the conclusion be more likely than the premise?" and "Can the conclusion be less likely than the premise?" Both questions were used, instead of choosing one of the two to implement in the experiment, in order to obtain a wider picture of the data that could be generalized across possible response biases or negation effects, which if present could affect the two questions differently. The order of the two questions alternated randomly between participants, but was fixed within each participant.

The above instructions were considered unsuitable for the two-premise inferences, because the latter have more complex coherence intervals that stand in a less direct relation to the probabilities of the premises. For the premise probabilities used, the above instructions would have meant that any response to these inferences was correct: it was always coherent for the probability of the conclusion to be either higher or lower than that of the premises, except in the trivial case of a premise probability of 1, for which no higher probability is defined.

For this reason, a threshold probability of .5 was chosen instead for the two-premise inferences. With this threshold the valid inferences become separable from the invalid inferences because for the premise probabilities used, the probability of the conclusion of the valid inferences MP and MT could only be higher than .5, whereas that of the invalid

inferences AC and DA could be both higher or lower (see Figure 6.1 of Experiment 5). The instructions for the two-premise inferences read:

In the following you will be shown a series of short stories. In each story an inference is drawn, and your task will be to judge whether it is possible, that is, consistent, for the conclusion of the inference to be more likely than 50%, and whether it is possible for the conclusion to be less likely than 50%.

On each trial in this section, participants were asked the corresponding two questions: "Can the conclusion be more than 50% likely?" and "Can the conclusion be less than 50% likely?"

In both blocks participants provided their responses by clicking with the mouse on one of two radio buttons with the headings "no" and "yes", respectively.

The tasks for the one- and the two-premise inferences are not equivalent, and it may be advisable to make comparisons between them with caution. However, an advantage of using a threshold of 50% for the two-premise inferences is that it may render responses to these inferences more comparable, in relative terms, to the responses in Experiment 7, which also used a binary response format but differed in its specific task instructions.

The context stories in which each inference was embedded followed the same construction principles as in Experiment 5. Six of the stories were nearly identical to those of Experiment 5, and six new stories were created, again with varying topics (houses after an earthquake, minefield, computer virus, armed robber, poisoned food, and flooded underground system). The stories were pseudonaturalistic, that is, they referred to concrete but fictional situations for which it is difficult to draw on world knowledge to assess the probabilities of the events involved. Further, the stories were constructed with the aim of creating situations in which "a lot is at stake" and which at the same time call for careful, reflective reasoning. It was hypothesized that in such situations it may become more relevant for people to distinguish valid from invalid inferences, if there are situations in which people distinguish between them in the first place. Therefore, using stories with these characteristics may make it easier to detect any differential effect of validity over and above coherence. The frame below shows a sample trial for the inference of nDM and a premise probability of .8. The 12 stories used can be found in Appendix E.

Within each block, the order of occurrence of the three trials for each inference, as well as the order of occurrence of the inferences, varied randomly for each participant. The order of the blocks alternated randomly for each participant.

With 12 inferences and 3 premise probability conditions, the experiment had 36 trials, plus two catch trials (one in each block) to make sure participants were reading the materials. The catch trials were similar in format to the regular trials, but the text of the inferences was

replaced with text stating that they were a control trial to make sure participants were paying attention, and asking them not to respond, but to instead click "next" to continue with the experiment.

Imagine you are part of a team of aid workers who are removing the mines from the Dunlar fields, where a war took place recently. You have to act very thoroughly in order to make sure the area is cleared and safe again for the residents. You are reviewing the latest data on the fields with the team.

Based on the information gathered until now:

Premise: You think it's 80% likely that:

The moss field and the gravel field are cleared.

Conclusion: Therefore, how likely can the following be?

The moss field is not cleared or the gravel field is not cleared.

Can the conclusion be more likely than the premise? (no) (yes)

Can the conclusion be less likely than the premise? (no) (yes)

Procedure

Participants were tested individually in a lab room of the Department of Psychological Sciences of Birkbeck, University of London. The experimenter remained in the room while participants read the instructions and went through three practice trials involving different inferences and scenarios from those in the main experiment. The entire session took approximately 22 min to complete.

Results and discussion

Figure 6.6 shows the values of observed and above-chance coherence for each inference when aggregated over the predictors, excluding a control condition that is treated separately below. In the control condition, the question for the one-premise inferences was whether the probability of the conclusion could be higher than a premise probability of 1. To give a coherent response for this question, it is enough to know that probabilities cannot be higher than 1, and so its inclusion in the figure would have unduly inflated response coherence for these inferences.

Figure 6.7 displays the data separately for each premise probability condition and for the two questions asked (whether the probability of the conclusion can be higher, resp. lower, than

the probability of the premise). As in Experiment 5, the uniform chance rate of 50% implies that the maximum possible value of above-chance coherence was 50%, and that the pattern of coherence across inferences was the same for observed and above-chance coherence.

Overall above-chance coherence

The pattern of results across inferences was similar to that of Experiment 5, with above-chance coherence highest for MP and nDM, and coherence not being above chance for AC and DA. However, the confidence intervals in Figure 6.6 suggest that in the present experiment coherence was above chance levels for OrIf – for which it had been at chance levels in Experiment 5. Conversely, coherence for IfOr now seems to be at chance levels – whereas it had been above chance levels in Experiment 5. In addition, coherence now appears to be at chance levels for DM, for which responses had been reliably above chance in Experiments 3 to 5. Further, among the inferences of Type A and Type B, it appears that coherence was higher for the invalid than for the valid inferences.

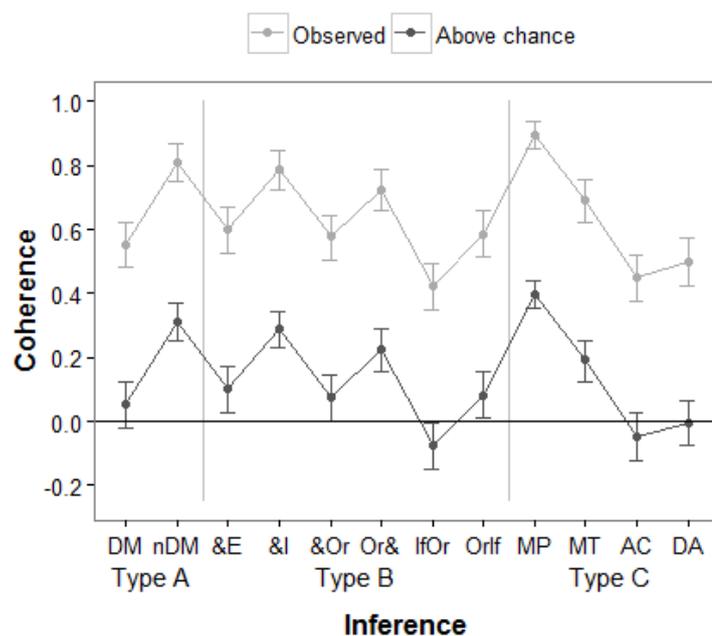


Figure 6.6. Observed and above-chance coherence for the inferences in Experiment 6, excluding the data from the condition in which premise probability was 1 and the question was whether the probability of the conclusion could be higher. Error bars show 95% CIs.

A first set of comparisons between inferences corroborated the above pattern. Above-chance coherence did not differ significantly between MP and nDM ($EMM_{MP} = .400$, $EMM_{nDM} = .343$, $F(1, 70) = 1.595$, $p = .211$). However, the values for nDM, &I, Or&, and MT differed only marginally between each other ($EMM_{&I} = .300$, $EMM_{Or\&} = .262$, $EMM_{MT} = .190$, $F(3,$

140) = 2.391, $p = .071$), and together were lower than that of MP ($EMM = .274$, $F(1, 175) = 7.343$, $p = .007$).

Above-chance coherence did not differ significantly between DM, &E, &Or, and OrIf ($EMM_{DM} = .114$, $EMM_{\&E} = .133$, $EMM_{\&Or} = .114$, $EMM_{OrIf} = .124$, $F(3, 140) = .048$, $p = .986$), and the overall value for these inferences was lower than the overall value for nDM, &I, Or&, and MT ($EMM_{former} = .121$, $EMM_{latter} = .274$, $F(1, 280) = 26.013$, $p < .001$).

Similarly, the overall values for IfOr, AC, and DA did not differ significantly from each other¹¹ ($EMM_{IfOr} < .001$, $EMM_{AC} = .024$, $EMM_{DA} = .048$, $F(2, 595) = .483$, $p = .617$), and together were lower than the values for DM, &E, &Or, and OrIf ($EMM_{former} = .024$, $EMM_{latter} = .121$, $F(1, 1435) = 14.469$, $p < .001$).

Finally, individual tests (including only fixed effects given the smaller sample sizes involved) were conducted for the inferences for which Figure 6.6 suggested above-chance coherence to be marginal. Coherence was above chance levels for DM ($F(1, 210) = 11.576$, $p = .001$), for &E ($F(1, 210) = 16.077$, $p < .001$), for &Or ($F(1, 210) = 11.576$, $p < .001$), and for OrIf ($F(1, 210) = 13.717$, $p < .001$). But coherence was at chance levels for IfOr ($F(1, 210) < .001$, $p = 1$), for AC ($F(1, 210) = .477$, $p = .490$), and for DA ($F(1, 210) = 1.922$, $p = .167$).

As in Experiment 5, the responses that were classified as incoherent for AC and DA would have been classified as coherent under the assumption that their conditional premise is a biconditional, i. e. that there is a correlation between the antecedent p and the consequent q rendering the relation between them bidirectional. This interpretation of the overall lack of above-chance coherence for AC and DA is supported by the pattern in Figure 6.7: The confidence intervals in this figure show that coherence was generally above chance levels for the question of whether the probability of the conclusion can be higher than 50%, and below chance levels for the question of whether the probability of the conclusion can be lower than 50%. The observed difference between the two conditions is exactly what one would expect from a biconditional interpretation. Under a biconditional interpretation AC and DA would be valid, and as a result their conclusion probability could be above 50%, but not below 50%. In contrast, with a conditional interpretation the two inferences are invalid, and their confidence intervals are nearly uninformative (c. f. Figure 6.1 of Experiment 5), rendering conclusion probabilities both above and below 50% coherent. Hence, if participants are being coherent when asked whether the probability of the conclusion can be higher (by saying "yes, it can be higher") but incoherent when asked whether it can be lower (by saying "no, it cannot be lower"), they are responding as if they were following a biconditional interpretation. As mentioned in Experiment 5, evidence for such an interpretation has been found intermittently

¹¹ Whereas the previous comparisons of this section included random intercepts for participants and for materials, this specific comparison included only a random intercept for participants. The attempt to include a random intercept for materials led to failure of convergence.

in previous studies (e. g. Barrouillet & Gauffroy, 2015; Skovgaard-Olsen et al., 2016; but see Oberauer, Weidenfeld, et al., 2007; Singmann et al., 2014).

This explanation does not apply to the absence of above-chance coherence for IfOr. The inference of IfOr is valid whether the conditional it contains is interpreted as uni- or bidirectional. To see why, consider again that a biconditional is equivalent to the conjunction of two conditionals: *q if and only if p = if p then q & if q then p*. By the valid rule of *and-elimination* (inference 3 in Table 6.1), this conjunction implies either of its conjuncts, i. e. the biconditional implies the conditional. By p-validity we know that when a statement *p* implies another statement *q*, $P(p) \leq P(q)$. Thus we have $P(\text{biconditional}) \leq P(\text{conditional}) \leq P(\text{disjunction})$, and with it: $P(\text{biconditional}) \leq P(\text{disjunction})$.

The absence of above-chance coherence observed for IfOr in this experiment suggests that the lower coherence found for the inference in Experiments 3 and 4 was not just an artefact caused by lower premise probabilities assigned to it, but a genuine failure to take coherence constraints into account. Given that findings of an absence of above-chance coherence were the exception rather than the rule in the present experiments, it seems preferable to look for an explanation for why coherence broke down in this specific case rather than to take it as evidence for a lack of sensitivity to coherence in general.

As mentioned earlier, a possible explanation for the absence of above-chance coherence found for IfOr is that the negation in the antecedent of the conditional made the inference more difficult to process. An account of such a negation effect would be given by the contrast class theory of negations together with the definition of the Ramsey test (Evans & Over, 2004; Oaksford, 2002; Oaksford et al., 2000). The contrast class account of negations describes negations in a probabilistic, set-theoretic way: the probability of a negated statement (e. g. "no cup of coffee") is the "contrast class" or "complement set" to the probability of the affirmation of the statement ("a cup of coffee"). Given that concepts, to be informative, often refer to rather specific objects, like a cup of coffee, a chair, or a person, the class of objects that are thereby not referred to, i. e. their complement set, is often larger than the set that is being affirmed. For example, the class of all objects that are not chairs (e. g. cups of coffee, dogs, underground stations) is larger than the class of objects that are chairs. Applied to the conditional in the IfOr inference, this means knowing that *p* is not the case is less informative than knowing that *p* is the case: if *p* is false there is still an infinite number of possibilities of what could be the case instead. As a consequence, it may be more difficult to build a mental representation of the antecedent on the basis of which to perform the Ramsey test. An additional difficulty may arise once the Ramsey test has been performed, because the hypothetical state of affairs in which *p* is false has to be related to the state of affairs in the conclusion in which *p* may be true. This means that the reasoner's focus of attention cannot remain within the mental simulation created through the Ramsey test, but has to set the

outcome of that simulation aside to consider the broader set of possibilities that include the truth of p .

This is however only one possible explanation, whose adequacy would have to be tested in follow-up experiments. Further, for any explanation attempted it is relevant to take into account that coherence for this inference is not reliable: It was found to be above chance levels in Experiments 1 and 5, above chance levels in some conditions of Experiments 3 and 4, and at chance levels in the present experiment. One way to examine further this variability would be to test whether coherence for the inference is higher when using real world materials, and whether it is higher for people with higher working memory capacity. One way to assess the role of negations within it would be through the use of the negation paradigm (Espino & Byrne, 2013; Evans et al., 1995; Evans & Handley, 1999). Part of this paradigm was implemented in Experiment 1 (Cruz et al., 2015) but did not lead to meaningful differences in above-chance coherence between conditions.

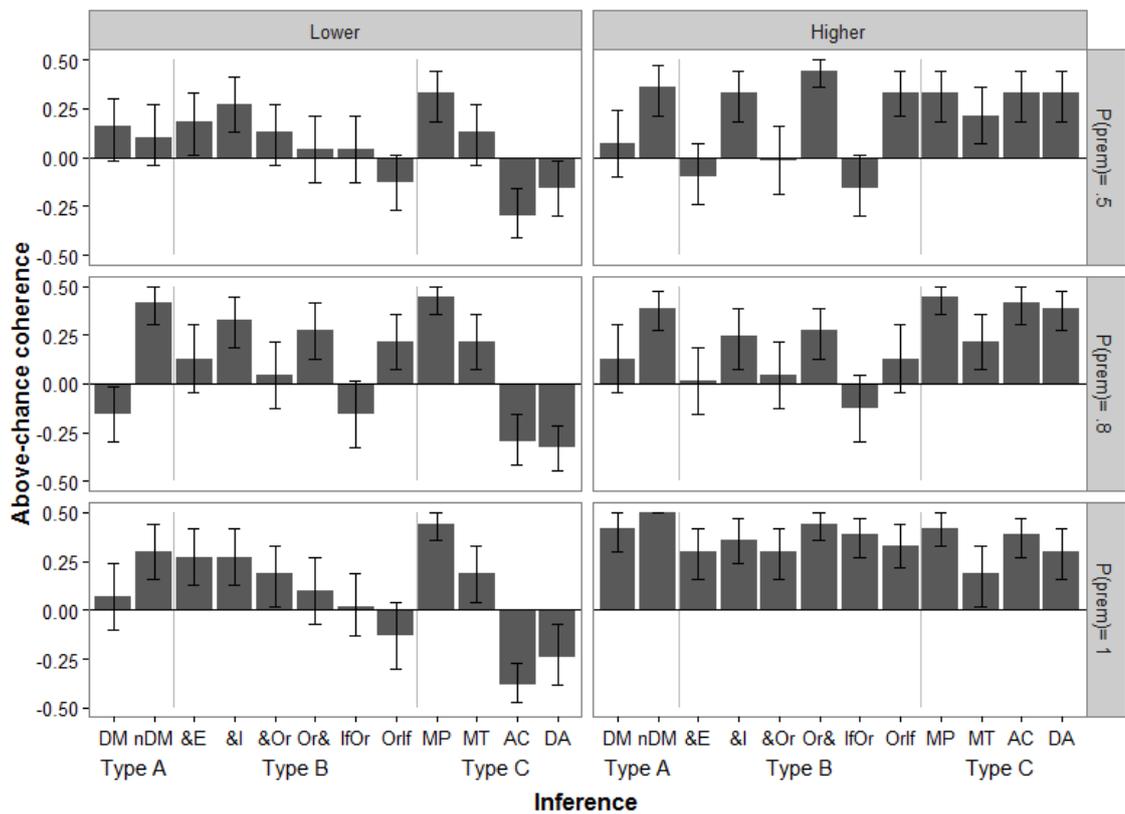


Figure 6.7. Above-chance coherence for Experiment 6, separately for each premise probability and question condition. Higher: question of whether the probability of the conclusion can be higher than that of the premise (resp. for the two-premise inferences, whether it can be higher than .5). Lower: question of whether the probability of the conclusion can be lower than that of the premise (resp. for the two-premise inferences, whether it can be lower than .5). Error bars show 95% CIs.

Higher vs. lower

Figure 6.7 shows the values of above-chance coherence separately for the question of whether the probability of the conclusion can be higher than that of the premise (resp. for the two-premise inferences, whether it can be higher than 50%) and for the question of whether the probability of the conclusion can be lower than that of the premise (resp. for the two-premise inferences, whether it can be lower than 50%). Let us call this variable *higher-lower*.

As mentioned above, the condition in the lower right panel of Figure 6.7 has a special status for the one-premise inferences (types A and B). Here participants were asked to assume that the probability of the premise is 1, and to judge whether the probability of the conclusion can be higher. Of course, no probability can be higher than 1, and so this question can be answered independently of the logical structure of the inferences. It was therefore used as a control condition assessing whether participants understood the instructions and took the upper probability bound into account.

Overall, above-chance coherence in the control condition was high and reliably above chance levels ($F(1, 280) = 392.475, p < .001$), providing evidence that participants had no problem in understanding the instructions. The degree of above-chance coherence did not differ significantly between the one-premise inferences ($F(7, 280) = 1.737, p = .1$, only fixed effects included). However, above-chance coherence was not at ceiling either ($F(7, 280) = 40.379, p < .001$). The relative values of above-chance coherence in the remaining conditions arguably appear more consequential when compared to the benchmark of results in this condition than when compared to the maximum possible value of above-chance coherence of .5.

As before, the absence of above-chance coherence found for AC and DA is difficult to interpret because there is reason to believe that it could reflect a coherent response based on an alternative interpretation of the conditional premises. Responses to these two inferences were therefore not included in the subsequent analysis.

The analysis was divided into three sections. The first section assessed above-chance coherence for the two higher-lower question conditions, focussing on the two uncertain premise probability conditions. The second section compared above-chance coherence between the certain and the uncertain premise probability conditions, focussing on the question of whether the probability of the conclusion can be lower than the probability of the premise. Finally, in a third section above-chance coherence for MP and MT was compared for the three premise probability conditions and the two higher-lower conditions.

1. Analysis for the uncertain premise probability conditions. A linear mixed model was computed for the effects of inference type, premise probability (.8, .5), and higher-lower on above-chance coherence, with random intercepts for participants and scenarios. Overall coherence was 20% higher than expected by chance ($F(1, 350) = 108.509, p < .001$). The degree to which coherence was higher than expected by chance differed between inference

types ($EMM_A = .182$, $EMM_B = .115$, $EMM_C = .289$, $F(2, 350) = 8.485$, $p < .001$): Above-chance coherence was higher for the inferences of type C than for the inferences of type A ($F(1, 525) = 8.891$, $p = .003$, $X^2(4) = 11.182$, $p < .025$) and than for the inferences of type B ($F(1, 1085) = 29.882$, $p < .001$). Above-chance coherence was also slightly higher for the inferences of type A than for those of type B ($F(1, 1085) = 4.086$, $p = .043$). No other effects were significant; in particular, there was no main effect of higher-lower ($EMM_{higher} = .217$, $EMM_{lower} = .174$, $F(1, 1050) = 3.160$, $p = .076$, $X^2(6) = 8.143$, $p < .250$).

Type A inferences. Figure 6.7 suggests that the differences in above-chance coherence between inference types concern not only the overall size of the effect, but that the effect also differed between individual inferences. A follow-up analysis focussing on the inferences of type A (including only a random intercept for participants, as adding a random intercept for items would have led to convergence problems) showed that above-chance coherence was higher for nDM than for DM ($EMM_{DM} = .050$, $EMM_{nDM} = .314$, $F(1, 245) = 26.679$, $p < .001$). Above-chance coherence was also marginally higher for the question of whether the probability of the conclusion could be higher than for the question of whether it could be lower ($F(1, 245) = 4.385$, $p = .037$). However, these two main effects were qualified by an interaction between inference and premise probability ($F(1, 245) = 8.594$, $p = .004$); and by a three-way interaction between inference, premise probability, and higher-lower ($F(1, 245) = 10.309$, $p = .002$).

Further examination of the three-way interaction showed that when the question was whether the probability of the conclusion could be higher, and premise probability was .5, overall coherence was above chance levels ($EMM = .214$, $F(1, 35) = 15.441$); but it was higher for nDM than for DM ($EMM_{DM} = .071$, $EMM_{nDM} = .357$, $F(1, 35) = 8.974$, $p = .005$). The same pattern was observed in the higher question condition with a premise probability of .8 (overall above-chance coherence: $EMM = .257$, $F(1, 35) = 19.352$, $p < .001$. Effect of inference: $EMM_{DM} = .129$, $EMM_{nDM} = .386$, $F(1, 35) = 12.115$, $p = .001$), and in the lower question condition with a premise probability of .8 (overall above-chance coherence: $EMM = .129$, $F(1, 35) = 7.580$, $p = .009$. Effect of inference: $EMM_{DM} = -.157$, $EMM_{nDM} = .414$, $F(1, 35) = 37.838$, $p < .001$): in the three cases coherence was above-chance levels overall, and higher for nDM than for DM. In contrast, in the case of the lower question condition with a probability of .5, overall coherence was at chance levels ($EMM = .129$, $F(1, 35) = 3.571$, $p = .067$), and did not differ between inferences ($EMM_{DM} = .157$, $EMM_{nDM} = .100$, $F(1, 35) = .405$, $p = .529$).

Figure 6.7 suggests that in the three cases in which overall coherence was above chance levels, it was so only due to nDM, and that coherence for DM remained at chance levels across conditions. This was confirmed statistically: the largest effect for DM was in the lower question condition with premise probability of .5, for which it did not reach significance (including only fixed effects: $EMM = .157$, $F(1, 35) = 3.836$, $p = .058$).

Overall, coherence was found to be above chance levels in three out of four conditions for nDM, whereas it remained at chance levels across conditions for DM. Going back to Figure 6.6, coherence for nDM was not only above chance levels in most cases, but the overall degree to which it was above chance levels was also among the highest of the 10 inferences investigated. The overall high coherence obtained for the inference is in accordance with the findings of Experiments 3 to 5.

However, it is interesting that coherence for nDM was at chance levels in the condition in which the question was whether the probability of the conclusion could be lower than the premise probability of .5. In this specific condition, responses to the inference are only coherent when participants always respond "no": even though the inference is a contradiction, the probability of the conclusion has to be equal to the probability of the premise. This procedural requirement for coherence might have irritated participants, leading to the contrast in the results between this and the other conditions.

The finding that coherence was at chance levels for DM across conditions contrasts with that of experiments 3 to 5, in which coherence had been found to be reliably above chance for the inference. An explanation for this difference may again be procedural: in the present experiment, responses to DM were only coherent when participants always responded "no": Given that the inference was an equivalence, the probability of the conclusion could be neither higher nor lower than the probability of the premise. It may have been irritating for participants that the coherent response was always "no" for this inference, while the response allocation varied for the remaining inferences.

The findings for DM and nDM showed no systematic effect of higher-lower on above-chance coherence: Responses for DM generally remained invariant across conditions, and responses for nDM differed only for a particular combination of higher-lower with premise probability.

Type B inferences. An analysis of the effects of premise probability (.5, .8), higher-lower, and validity on above-chance coherence focussing on the inferences of type B (again including only a random intercept for participants) showed that above-chance coherence was higher for the invalid than for the valid inferences ($EMM_{valid} = .002$, $EMM_{invalid} = .229$, $F(1, 805) = 52.152$, $p < .001$). However, the effect of validity was qualified by an interaction between validity and higher-lower ($F(1, 805) = 15.030$, $p < .001$) which was in turn qualified by a three-way interaction between validity, higher-lower, and premise probability ($F(1, 805) = 18.775$, $p < .001$).

Follow-up analyses to the interaction (using only fixed effects due to the small sample sizes involved in the comparisons) showed that in the condition in which the question was whether the probability of the conclusion could be higher and premise probability of .5, overall coherence was above chance levels ($EMM = .138$, $F(1, 210) = 22.412$, $p < .001$), but was larger for the invalid than for the valid inferences ($EMM_{invalid} = .367$, $EMM_{valid} = -.090$,

$F(1, 210) = 61.401, p < .001$). The negative estimated marginal means for the valid inferences imply that coherence was actually only above chance levels for the invalid inferences.

The same pattern of results was obtained in the condition in which the question was whether the probability of the conclusion could be higher and the premise probability was .8 (Overall coherence: $EMM = .095, F(1, 210) = 8.400, p = .004$. Effect of validity: $EMM_{valid} = -.024, EMM_{invalid} = .214, F(1, 210) = 13.125, p < .001$) and in the lower question condition with a premise probability of .8 (overall coherence: $EMM = .138, F(1, 210) = 18.788, p < .001$. Effect of validity: $EMM_{valid} = .005, EMM_{invalid} = .271, F(1, 210) = 17.515, p < .001$). In the three cases overall coherence was above chance, but it was in fact only significantly higher than chance for the invalid inferences.

In contrast, in the lower question condition with a premise probability of .5, overall coherence was above chance ($EMM = .090, F(1, 210) = 7.133, p = .008$), and there was no difference between valid and invalid inferences ($EMM_{valid} = .119, EMM_{invalid} = .062, F(1, 210) = .711, p = .400$).

In summary, overall coherence was above chance levels for the inferences of type B, but it was generally only significantly so for the invalid inferences. The exception was in the lower question condition with a premise probability of .5: in this case coherence was above chance levels and no difference between valid and invalid inferences was found. The observation that above-chance coherence tended to be higher for invalid inferences coincides with the results for the inferences of type A, where above-chance coherence tended to be higher for the invalid inference of nDM than for the valid inference of DM. The observed difference cannot be explained by the fact that for the premise probabilities studied, the coherence interval tended to be wider for the invalid than for the valid inferences. This is because the number of "yes" trials were nonetheless equal for valid and invalid inferences of this type, and because the effect of validity was clearly present also in the condition in which the question was whether the probability of the conclusion could be higher and the premise probability was .5, in which the intervals for valid and invalid inferences were equally wide. However, no clear effect of validity was found in Experiment 5 for these inferences, and Figure 6.7 suggests that in the present experiment, the effect was absent in the lower question condition when premise probability was .5, but also when premise probability was 1.

There was no systematic effect of higher-lower for the inferences of type B. Instead, the form of the effect of higher-lower changed between inferences and premise probability conditions.

Type C inferences. An analysis of the effects of inference, premise probability (.8, .5), and higher-lower on above-chance coherence focussing on MP and MT of type C (again including only a random intercept for participants) showed that above-chance coherence was higher for MP than for MT ($EMM_{MP} = .386, EMM_{MT} = .193, F(1, 245) = 23.093, p < .001$).

There was also a non-significant trend for an effect of premise probability ($EMM_{.8} = .329$, $EMM_{.5} = .250$, $F(1, 245) = 3.833$, $p = .051$).

Overall, above-chance coherence was larger for MP than for MT across experimental conditions, in line with the pattern observed in Figure 6.7 and consistent with the findings of the previous experiment. The higher-lower question condition again played no role for this outcome ($F = .285$, $p = .594$).

2. Analysis for the lower question condition. The second analysis focussed on responses to the question of whether the probability of the conclusion can be lower than that of the premise, comparing the results of the previous analysis for this question with the case in which the premise probability was 1. A linear mixed model was computed for the effects of inference type and premise probability (1, .8, .5) on above-chance coherence, with random intercepts for participants and materials. Overall responses were coherent around 19% more often than expected by chance ($EMM = .185$, $F(1, 350) = 71.563$, $p < .001$). Above-chance coherence differed between inference types ($EMM_A = .148$, $EMM_B = .116$, $EMM_C = .290$, $F(2, 350) = 6.251$, $p = .002$). But there was no difference in above-chance coherence between premise probability conditions ($EMM_1 = .206$, $EMM_{.8} = .198$, $EMM_{.5} = .149$, $F(2, 700) = 1.889$, $p = .152$); nor an interaction between premise probability and inference type ($F(4, 700) = .517$, $p = .723$). Hence, it made no significant difference for the coherence of responses whether the probabilities of the premises were certain or uncertain.

In line with this result, the overall pattern in Figure 6.7 looks similar for the certain and uncertain premise probability conditions. However, for the inferences of type B with certain premise probability, the figure suggests there to be no difference in above-chance coherence between valid and invalid inferences. Given the theoretical interest of this question, it was followed up statistically. The analysis (which included only fixed effects) showed that there was indeed no difference between valid and invalid inferences when premise probability was 1 ($EMM_{valid} = .157$, $EMM_{invalid} = .081$, $F(1, 210) = 1.300$, $p = .255$). This means that the overall results for the effect of validity among inferences of type B are mixed: above-chance coherence was higher for the invalid inferences in three conditions, and there was no effect of validity in two conditions. These mixed results replicate those of the previous experiments, and together provide no evidence of a general effect of validity on above-chance coherence. The differences in above-chance coherence observed seem more likely due to more specific features of the inferences studied.

The results provide no evidence for an effect of conservativeness, as was suggested in Experiment 5. A conservative response tendency would have meant that participants more often said "no, it cannot be higher" and "yes, it can be lower". This would have led to higher coherence for the invalid inferences in both question conditions, contrary to what was observed. The findings also provided no evidence for a general tendency to respond "yes" or to respond "no". A tendency to say "yes" would have led to higher coherence for the valid

inferences in the higher question condition, and a tendency to say "no" would have led to higher coherence for the valid inferences in the lower question condition, neither of which was observed.

3. Analysis for MP and MT across conditions. Finally, the third analysis takes advantage of the fact that for MP and MT, all experimental conditions were comparable, including that in which the question was whether the probability of the conclusion can be higher and the probability of the premises was 1. It also makes sense to have a separate analysis for these two inferences given that the questions asked for them differed slightly from those for the one-premise inferences, having as a reference not the probability of the premise but a probability of .5.

An analysis for the effects of inference, premise probability, and higher-lower on above-chance coherence, including random intercepts for participants and scenarios, showed that overall coherence was around 30% higher than expected by chance ($EMM = .295$, $F(1, 70) = 80.635$, $p < .001$). Above-chance coherence was higher for MP than for MT ($EMM_{MP} = .400$, $EMM_{MT} = .190$, $F(1, 70) = 10.153$, $p = .002$). No other effects were significant (highest $F = 2.587$, lowest $p = .077$, for the effect of premise probability).

The finding of high above-chance coherence for MP, but also generally reliable above-chance coherence for MT, is in accordance with the findings of the previous experiments, and provides a more optimistic picture of role of coherence for complex two-premise inferences than had been suggested by earlier findings (Evans et al., 2015; Pfeifer & Kleiter, 2009; Singmann et al., 2014).

General discussion

This experiment used an alternative task to that in experiment 5, while coinciding with the previous experiment in using a binary response format that makes it possible to perform quantitative comparisons of above-chance coherence between inferences. The results corroborate and extend those of the previous experiment in key respects.

Overall above-chance coherence was again found to be highest for nDM and MP, with coherence for MT and the type B inferences being lower but in their majority still above chance levels. This finding adds to the previous evidence that in general coherence is a psychologically meaningful normative constraint for reasoning.

Coherence for AC and DA again followed a pattern consistent with the assumption that participants are giving coherent responses based on a biconditional interpretation of the conditional premises involved. The lack of above-chance coherence observed for them is therefore difficult to interpret. As mentioned in Experiment 5, a useful next step in assessing above-chance coherence for these two inferences would be to vary the correlation between the

antecedent and consequent of the conditional, or measure participants' subjective assessment of this correlation, and then compute response coherence taking this information into account.

Coherence for DM was found to be at chance levels across conditions in this experiment, in spite of having been reliably above chance levels in experiments 3 to 5. A possible reason for this is that in the present experiment, the only coherent response for DM in all trials was "no". Because it is an equivalence, the probability of the conclusion can never be higher or lower than the probability of the premise. This inference specific constraint might have irritated participants, who perhaps expected the normative response to vary more across trials as it did for the other inferences. Some participants might have also found it difficult to consistently respond "no" in front of an equivalence, which might rather suggest the answer "yes, these two probabilities are the same" or "yes, this inference is correct". The fact that the coherent response to DM was "no" on all trials was a consequence of the design. It would of course not have been a good choice of design if this feature had been present throughout. However, its presence for one inference made it possible to study what happens with the coherence of responses in this case, while still being able to measure coherence for other cases. Further experiments would have to be conducted to assess whether it were such procedural aspects of the present task that led to a failure of coherence for the inference.

The findings provide no evidence that above-chance coherence differed as a function of the higher-lower question condition. It also played no role whether the probability of the premises was certain or uncertain. Furthermore, in line with the results of Experiment 5, there was no consistent effect of validity on above-chance coherence. Responses were higher for the invalid inference of nDM than for the valid inference of DM, but they were among the highest for the valid inference of MP. Similarly, among the inferences of type B above-chance coherence was higher for invalid inferences only in three out of five experimental conditions. The present results suggest that where an effect of validity was observed, it is more likely a result of inference or condition specific factors than the reflection of a general difference in coherence between valid and invalid inferences.

A further finding of interest is that coherence was not lower for uncertain than for certain premise probabilities. This finding goes counter to the belief held in the binary paradigm in reasoning research, that to reasoning deductively, it is necessary to assume the premises to be certain. As mentioned in the introduction, a typical distinction made in previous studies is between an inference being either "deductive" or "probabilistic", implying that what is deductive is not probabilistic, and vice versa (e. g. Evans, Handley, Neilens, & Over, 2010; Singmann et al., 2014; Trippas et al., 2017; Verschueren et al., 2005). The present finding provides empirical support for the proposal that this dichotomy is not necessary when the deductive concepts investigated are generalised to the probability range. Certain truth and falsity are then just two endpoints on a scale for degrees of belief, and it is possible to study deduction and induction together in a unified probabilistic framework.

The use of a binary response format in Experiments 5 and 6 makes it possible to compare their findings more directly to an experimental condition with binary paradigm instructions. This was done in the following experiment.

EXPERIMENT 7: CERTAIN PREMISES AND BINARY PARADIGM INSTRUCTIONS

This experiment explores the coherence of people's responses in a situation in which not only the response format is binary, but in addition the instructions are those of the binary paradigm of reasoning: to assume the premises to be true "for the sake of argument", and decide whether under this assumption the conclusion necessarily also has to be true. The experiment used the same 12 inferences and the same materials as Experiment 6. This made it possible to compare the results of the two experiments more directly, and assess to what extent binary paradigm instructions, inspired on the concept of validity in classical logic, make a difference for people's reasoning over and above the assumption of certainty in the premises. Whereas in Experiment 6 the condition in which the premises had probability 1 represented one particular case among the range of premise probabilities that would have been possible, in the present experiment probabilities do not come into play at all: the only possibilities considered are truth and falsity, and what the assumption of one piece of information being true implies for the truth or the falsity of other pieces of information.

When comparing the data of the present experiment with that of Experiment 6, supporters of the binary paradigm would predict people's responses to be more coherent under binary paradigm instructions (Exp. 7) than under probabilistic instructions (Exp. 6). This is because in the binary paradigm probabilistic reasoning is equated with inductive reasoning, and only reasoning about certain truth and falsity is treated as deductive.

In the probabilistic approach, coherence would be predicted to be similar under probabilistic and binary instructions because from a probabilistic perspective certain truth and falsity are similar to probabilities of 1 and 0: they are not qualitatively different from other probabilities but represent limiting cases on a common scale.

For particular inferences like or-introduction, responses may be more coherent under probabilistic than under binary instructions, because the use of probabilistic instructions asking participants directly for their degrees of belief reduces the role of pragmatic factors involved in asserting something in discourse (Cruz et al., 2017; Grice, 1989; Politzer & Baratgin, 2016). But to the knowledge of the author there are no clear arguments suggesting that similar pragmatic issues would be involved in the inferences investigated here.

Further, there are factors for which a premise probability of 1 does make a difference in the probabilistic approach. For example, in Costello & Watts' (2016) model for estimating the likelihood of events through noisy Bayesian sampling from memory (see also Sanborn & Chater, 2016), it is difficult to estimate a probability of 1 in an unbiased way when the estimation process is noisy. But for present purposes, it seems safe to assume that such a form of probability estimation would nonetheless result in responses being coherent more often than

expected by chance, and so would not be expected to result in qualitatively different conclusions being drawn when comparing the two instruction types.

Method

Participants

A total of 46 participants from the recruitment pool of Birkbeck, University of London completed the experiment. Two participants were excluded because they failed a catch trial aimed to assess whether they were reading the materials. Further six were excluded because they indicated having previous knowledge of formal logic¹². One participant was excluded because of reporting less than "good" English language skills. A further participant was excluded for having one or more trial reaction times of less than three seconds. The final sample consisted of 35 participants. They had a median age of 25 years (range: 18-52). Most participants had some college education, with 33 (91%) indicating that they were students or had completed a university degree (20 an undergraduate, 11 a postgraduate, and 1 a doctoral degree). Participants' median percentage rating of task difficulty was 25%.

Materials and design

The inferences and scenarios were exactly the same as those of Experiment 6 (see Appendix E), as was the overall structure of the experimental design. However, at the beginning of the experiment participants were given the following instructions:

In the following you will be shown a series of short stories, in each of which an inference will be drawn. You will be asked to assume that the premises of the inference are true. Even if the premises may be uncertain in the real world, it is important that you assume, for the sake of argument, that they are certainly true. You will then be asked whether, assuming the premises are true, the conclusion necessarily has to be true as well.

On each trial participants were given the same premises and conclusion as in Experiment 6, but with no information on premise probability. They were then asked: "Assuming that the premise(s) is (are) true, does the conclusion also have to be true?" Participants gave their response by clicking on one of two radio buttons labelled "no" and "yes", respectively. Below is an example of a trial for the &I inference:

¹² Four indicated having visited a university course in formal logic, one read an academic book on formal logic, and one had a Master's in linguistics which included training in formal semantics.

Imagine you are part of a team of technicians surveying the underground system of the city of Limro. A severe flood has started to affect part of the underground. You need to analyse carefully which areas are flooded so that you can evacuate the passengers and avoid the flood spreading further. Consider the following inference:

Premise:

The red line is flooded.

Conclusion:

Therefore, the red line and the yellow line are flooded.

Assuming that the premise is true, does the conclusion also have to be true? (no) (yes)

The experiment consisted of a single block of 12 trials presented in random order, with the pairing of the 12 inferences with the 12 scenarios randomised for each participant. In addition, the experiment included a catch trial aimed to assess whether participants were reading the materials. As in Experiment 6, the catch trial was similar in format to the regular trials, but the text for the premises and conclusion of the inferences was replaced with text stating that it was a test question to make sure participants were paying attention, and asking them not to respond, but instead click "next" to continue with the experiment.

Procedure

Participants took part individually in a quiet testing room of the Department of Psychological Sciences of Birkbeck, University of London. The experimenter remained in the room while participants read the instructions and went through the first practice trials, which referred to different inferences and materials than in the main experiment. The entire session took approximately 15 minutes to complete.

Results and discussion

As in the previous experiments, a response was classified as coherent when it fell within the coherence interval, and as incoherent if it fell outside of it. This was equivalent to classifying "yes" responses as coherent for valid inferences and "no" responses as coherent for invalid inferences. Observed coherence again corresponded to the proportion of coherent responses in each condition. Given the binary response format, the chance rate was again 50%, and above-chance coherence was computed by subtracting this value from observed coherence. The

overall values of observed and above-chance coherence for each inference are shown in Figure 6.8.

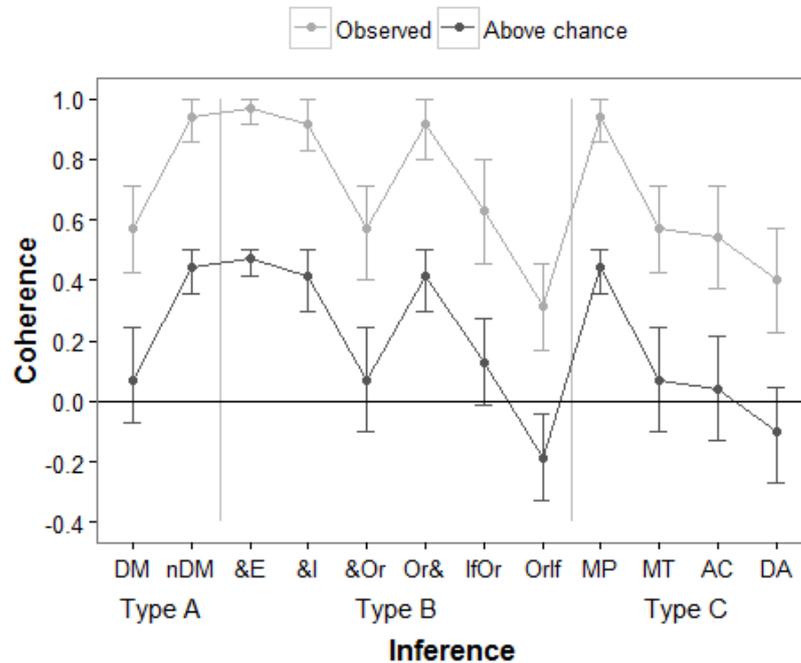


Figure 6.8. Observed and above-chance coherence for the 12 inferences of Experiment 7. Error bars show 95% CIs.

The confidence intervals in the figure show that coherence was clearly above chance levels for 5 of the 12 inferences: nDM, &E, &I, Or&, and MP. It seemed to be at or below chance levels for the remaining 7 inferences.

To assess these differences statistically, two contrasts were created: one for the inferences for which coherence was clearly above chance (the *high coherence* inferences group, composed of inferences nDM, &E, &I, Or&, and MP), and the other for the inferences for which coherence seemed to be at or below chance levels (the *low coherence* group, composed of inferences DM, &Or, IfOr, OrIf, MT, AC, and DA).

An analysis for the effect of the so defined inferences groups on above-chance coherence was conducted using only fixed effects, because inclusion of random effects led to failure of convergence for some comparisons. Overall coherence was around 23% higher than expected by chance ($F(1, 420) = 122.187, p < .001$). The extent to which coherence was above chance levels differed between the two groups ($EMM_{high} = .437, EMM_{low} = .014, F(1, 420) = 107.210, p < .001$). Analyses for each group of inferences confirmed that coherence for the high coherence group was above chance levels ($F(1, 175) = 572.147, p < .001$). The extent to which this was the case did not differ between inferences ($F(4, 175) = .342, p = .849$). Coherence for the low coherence group did not differ from chance ($F(1, 245) = .209, p = .648$); and this result did not differ between inferences ($F(6, 245) = 1.874, p = .086$). Figure 6.8 suggests that

above-chance coherence for IfOr might have been marginal, but its confidence interval included the null, and it was not significant when tested individually ($F(1, 35) = 2.478, p = .124$).

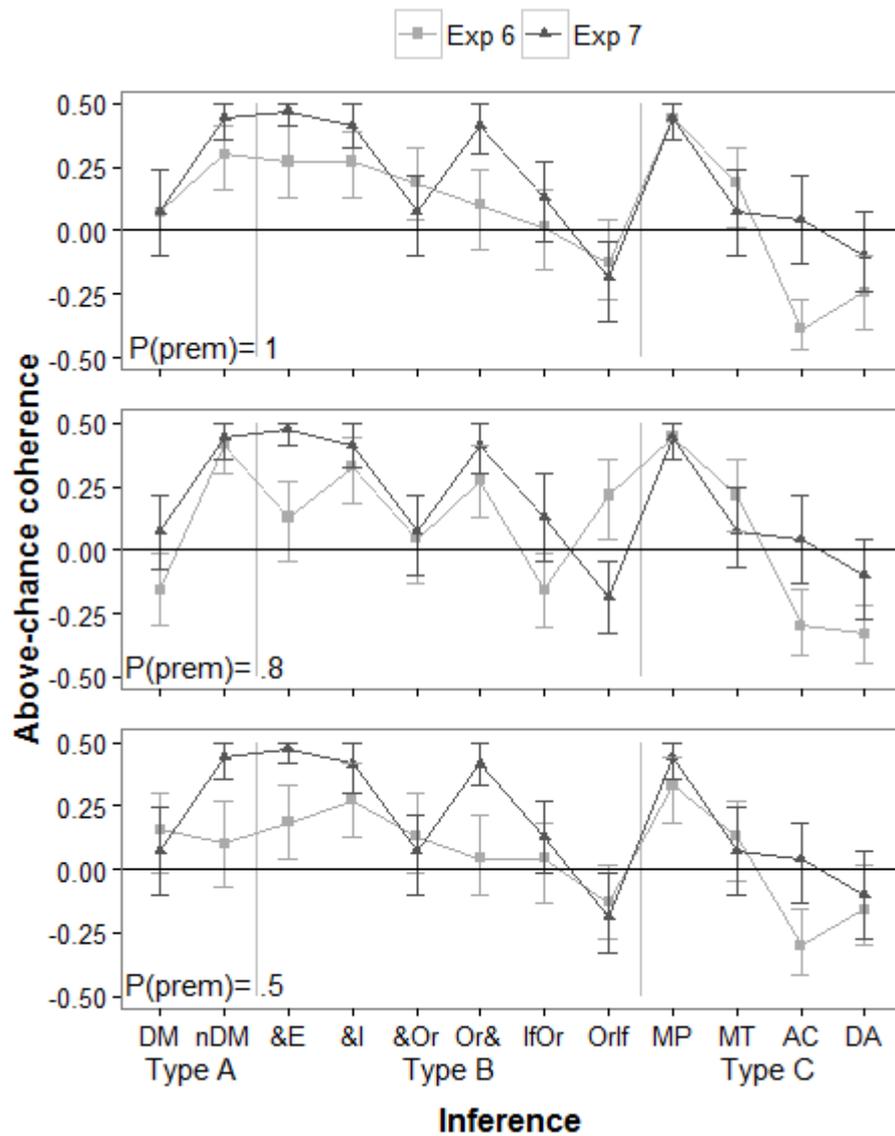


Figure 6.9. Above-chance coherence for binary instructions (Exp. 7) and probabilistic instructions (Exp. 6) when the question was whether the probability of the conclusion can be lower than the probability of the premise (resp. for the two premise inferences, whether it can be lower than 50%). The lower left corner of each panel shows the premise probability condition in Exp. 6 with which the data from Exp. 7 was compared to. Error bars show 95% CIs.

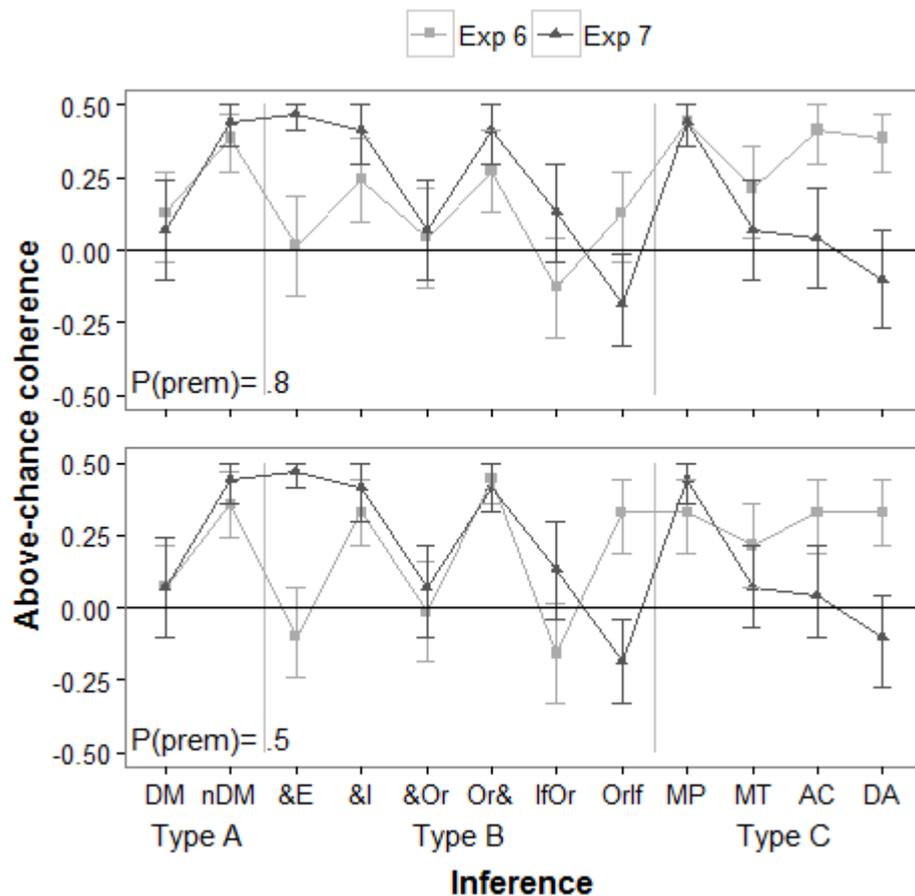


Figure 6.10. Above-chance coherence for binary instructions (Exp. 7) and probabilistic instructions (Exp. 6) when the question was whether the probability of the conclusion can be higher than the probability of the premise (resp. for the two premise inferences, whether it can be higher than 50%). The lower left corner of each panel shows the premise probability condition in Exp. 6 with which the data from Exp. 7 was compared to. Error bars show 95% CIs.

The above results show that overall coherence was above chance levels for fewer inferences under binary paradigm instructions in this experiment, than under probabilistic instructions in any of Experiments 3 to 6. The five inferences for which coherence was above chance include inferences of each of the three types defined, and they include both valid and invalid inferences, so that it can be ruled out that coherence was lower in this experiment specifically for one of these subcategories of inferences. Interestingly, Figure 6.8 also shows that for the five inferences for which coherence was clearly above chance levels, the degree of above-chance coherence was very high, close to the ceiling value of .5. It may be that the use of binary instructions has the effect of making the response pattern more binary as well, such that responses for which above-chance coherence tended to be high in the previous experiments are now coherent even more frequently, and responses for which coherence was

less high but still above chance levels in the previous experiments, are now coherent at chance levels.

Apart from the above observation, the pattern of results replicates the finding in Experiments 3 to 6 that coherence was higher for nDM than for DM, and that it was among the highest for MP. On the other hand, coherence among the two-premise conditional syllogisms was only above chance levels for MP, whereas in the previous experiments it had also been above chance for MT.

Figure 6.9 compares above-chance coherence in this experiment with the three premise probability conditions of Experiment 6 in which the question was whether the probability of the conclusion can be lower than the probability of the premise (resp. for the two-premise inferences, whether it can be lower than 50%). Because there were 35 participants in both experiments, each data point in the Figure is based on the same number of responses. Figure 6.10 shows the corresponding comparison for the question in Experiment 6 of whether the probability of the conclusion can be higher than that of the premise (resp. for the two-premise inferences, whether it can be higher than 50%). The condition in Experiment 6 for the higher question condition and a premise probability of 1 is not included in Figure 6.10, nor in the statistical analysis, because it amounted to the trivial question of whether a given probability can be greater than 1. Its inclusion would thus have unduly inflated above-chance coherence for probabilistic instructions.

The figures show that the pattern of responses to AC and DA in both Experiments 6 and 7 was not coherent above chance levels, but would have been so under a biconditional interpretation, rendering any findings for them difficult to interpret. The results for these two inferences were therefore not included in the subsequent analysis.

The analysis is divided into five parts, one for each pairing of the data from Experiment 7 with one of the non-trivial conditions from Experiment 6. A single analysis comparing the five relevant conditions of Experiment 6 with the results of Experiment 7 would have been difficult because the conditions in Experiment 6 were tested within participants, whereas the data from Experiment 7 come from a different sample of participants. The grouping of participants into only two samples is not enough to create a higher level of analysis in a linear mixed model, and it was not possible to include random slopes for participants in such an analysis either. The five analyses correspond to the five panels of Figures 6.9 and 6.10, respectively. They include only fixed effects, because inclusion of random effects led to failure of convergence for some of the comparisons.

Binary (Exp. 7) vs. lower question with $P(\text{prem}) = 1$ (Exp. 6)

An analysis of the effects of experiment (6, 7) and inference type on above-chance coherence showed that overall coherence was above chance levels ($EMM = .225$, $F(1, 700) = 133.953$, $p < .001$). Above-chance coherence differed between inference types ($EMM_A = .221$, $EMM_B =$

.169, $EMM_C = .286$, $F(2, 700) = 3.636$, $p = .027$). But it did not differ between the two experiments ($EMM_6 = .206$, $EMM_7 = .244$, $F(1, 700) = .957$, $p = .328$); and there was no interaction between experiment and inference type ($F(2, 700) = 1.590$, $p = .205$).

Overall, responses in Experiment 7, under binary paradigm instructions, were not more coherent than responses in Experiment 6 in the condition in which premise probability was 1 and the question asked was whether the probability of the conclusion could be lower.

Binary (Exp. 7) vs. lower question with P(prem) = .8 (Exp. 6)

An analysis of the effects of experiment and inference type on above-chance coherence showed that overall coherence was above chance levels ($EMM = .204$, $F(1, 700) = 129.653$, $p < .001$). Above-chance coherence differed marginally between inference types ($EMM_A = .193$, $EMM_B = .179$, $EMM_C = .038$, $F(2, 700) = 3.414$, $p = .033$). But there was no main effect of experiment ($EMM_6 = .198$, $EMM_7 = .244$, $F(1, 700) = 1.401$, $p = .237$); nor an interaction between experiment and inference type ($F(2, 700) = 1.996$, $p = .137$).

Overall, it made no difference to above-chance coherence whether people were given binary paradigm instructions, or probabilistic instructions with premise probabilities of .8 and the question of whether the conclusion probability could be lower.

Binary (Exp. 7) vs. lower question with P(prem) = .5 (Exp. 6)

An analysis of the effects of experiment and inference type on above-chance coherence again showed that overall responses were coherent above chance levels ($EMM = .197$, $F(1, 700) = 98.193$, $p < .001$). Above-chance coherence tended to be higher in Experiment 7 than in Experiment 6 ($EMM_6 = .149$, $EMM_7 = .244$, $F(1, 700) = 5.747$, $p = .017$, $X^2(3) = 10.912$, $p < .02$) (the effect is significant and adds to the fit of the model, but the confidence interval for the beta value includes the null). No other effects were significant (highest $F = 1.98$, lowest $p = .139$, for the effect of inference type).

Overall, above-chance coherence tended to be higher in Experiment 7 than in Experiment 6 in the condition in which premise probability was .5 and the question was whether the conclusion probability could be lower than this.

Binary (Exp. 7) vs. higher question with P(prem) = .8 (Exp. 6)

An analysis of the effects of experiment and inference type on above-chance coherence showed that overall coherence was above chance levels ($EMM = .236$, $F(1, 700) = 148.345$, $p < .001$). Above-chance coherence differed between inference types ($EMM_A = .257$, $EMM_B = .157$, $EMM_C = .293$, $F(2, 700) = 5.990$, $p = .003$); but it did not differ between the two experiments ($EMM_6 = .227$, $EMM_7 = .244$, $F(1, 700) = .203$, $p = .652$). Further, there was a marginal, non-significant interaction between experiment and inference type ($F(2, 700) = 2.863$, $p = .058$).

An examination of the pattern underlying the interaction trend showed that for the inferences of type A, there was no effect of experiment ($EMM_6 = .257$, $EMM_7 = .257$, $F(1, 140) < .001$, $p = 1$), and experiment did not interact with inference ($F(1, 140) = .722$, $p = .397$). For the inferences of type B, above-chance coherence was higher in Experiment 7 than in Experiment 6 ($EMM_6 = .095$, $EMM_7 = .219$, $F(1, 420) = 7.508$, $p = .006$) but this effect was qualified by an interaction between experiment and validity ($F(1, 420) = 6.397$, $p = .006$). For the valid inferences &E, &Or, and IfOr, above-chance coherence was higher in Experiment 7 than in Experiment 6 ($EMM_6 = -.024$, $EMM_7 = .224$, $F(1, 210) = 14.328$, $p < .001$). But above-chance coherence did not differ between the two experiments for the invalid inferences &I, Or&, and OrIf ($EMM_6 = .214$, $EMM_7 = .214$, $F(1, 210) < .001$, $p = 1$). For inferences MP and MT of type C, there was again no main effect of experiment ($EMM_6 = .329$, $EMM_7 = .257$, $F(1, 140) = 1.283$, $p = .259$), and experiment did not interact with inference ($F(1, 140) = 1.283$, $p = .259$).

In summary, there was no main effect of experiment, and only a marginal, non-significant trend of an interaction between experiment and inference type. An examination, for the sake of completeness, of the pattern underlying the interaction trend revealed that experiment had no effect on above-chance coherence for the inferences of type A, for the inferences of type C, and for the three invalid inferences of type B. It was only among the three valid inferences of type B that above-chance coherence was higher in Experiment 7 than in Experiment 6. Overall, the results for this condition give no indication of a reliable difference in above-chance coherence between the two experiments.

Binary (Exp. 7) vs. higher question with P(prem) = .5 (Exp. 6)

An analysis assessing the effects of experiment and inference type on above-chance coherence showed that overall coherence was above chance levels ($EMM = .226$, $F(1, 700) = 134.966$, $p < .001$). No other effects were significant (highest $F = 2.233$, smallest $p = .108$, for the effect of inference type).

Overall, it made no difference to above-chance coherence whether participants were in an experiment with binary paradigm instructions or in one with probabilistic instructions and premise probabilities of .5, responding to the question of whether the probability of the conclusion could be higher than this premise probability.

General discussion

The present experiment builds on the evidence of experiments 3 to 6, showing that overall, people's responses are coherent above chance levels for a range of inferences of different complexity, across a wide range of materials. Using binary paradigm instructions, Experiment

7 corroborated the earlier findings with probabilistic instructions that coherence is among the highest levels for the contradiction of nDM and for MP. It also echoed the earlier findings of an absence of above-chance coherence for AC and DA. Given the overall clear evidence of sensitivity to coherence constraints, the consistent deviation from coherence for AC and DA suggests that participants are interpreting the materials differently from the interpretation used to compute coherence for them. In particular, they are in line with earlier findings suggesting that people sometimes interpret the conditionals involved in inferences as bidirectional, e. g. as establishing a correlation between the antecedent p and the consequent q (Baratgin et al., 2013; Barrouillet & Gauffroy, 2015; but see Oberauer, Weidenfeld, et al., 2007; Singmann, Klauer, & Over, 2014). Further studies are necessary to test whether this interpretation holds, by establishing whether response coherence varies systematically as a function of changes in the correlation between p and q .

Coherence was found to be above-chance levels for fewer inferences under binary instructions than under probabilistic instructions. For the conditional syllogisms, this was seen by the fact coherence for MT was above chance levels in Experiment 6, but not in Experiment 7. This paints a more positive picture of above-chance coherence for conditional syllogisms than had been suggested by earlier studies (Evans et al., 2015; Singmann et al., 2014), as well as of deductive reasoning under uncertainty more generally.

Of central interest in this section was the direct comparison of above-chance coherence for the two types of instruction, using the same response format and the same materials. Across the five comparisons made, there was no evidence that above-chance coherence was higher under binary instructions than under probabilistic instructions, whether the probabilistic instructions stated the probabilities of the premises to be certain or uncertain. A difference was observed only in two cases, and in both it was only marginal resp. unreliable.

In the first case, above-chance coherence tended to be higher under binary instructions when compared with the condition in which premise probability was .5 and the question was whether the probability of the conclusion could be lower than this. However, note that there was no difference in above-chance coherence when the condition with binary instructions was compared with that in which premise probability was .5 and the question was whether the probability of the conclusion could be higher than this. A possible explanation for this pattern is as a negation effect: it may be more difficult to think of the probability of events being lower than .5, because this is analogous to thinking of the probability of their negation being higher than .5. However, this explanation would have to be tested in further experiments.

The second difference observed was a non-significant trend of an interaction between inference type and experiment when the condition with binary instructions was compared with the condition in which premise probability was .8 and the question was whether the probability of the conclusion could be higher than this. Following up the interaction trend for the sake of completeness, showed that for this comparison there was no difference in above-

chance coherence between the two experiments for the inferences of type A, type C, and the invalid inferences of type B. But above-chance coherence was higher under binary instructions for the three valid inferences of type B. This highly localised and unreliable difference provides no indication of a general effect of instruction type on above-chance coherence.

The general absence of a difference in above-chance coherence between the conditions with binary and with probabilistic instructions constitutes strong evidence against the claim that deduction only occurs in the realm of reasoning about certainty, and that reasoning from uncertain information is inherently inductive.

The experiments in this chapter focussed on comparing the mean values of above-chance coherence between inferences and between inference groups (e. g. between the three inference types defined, and between valid and invalid inferences) using different measurement methods. A further relevant feature of people's sensitivity to coherence is response variance. The previous chapter included a brief investigation of whether participants' conclusion probability judgments vary more strongly when the coherence interval for the conclusion is wide than when it is narrow. The following chapter looks at the question of an effect of response variance from a different perspective, assessing the possibility of it having an effect on response confidence. It also assesses people's general sensitivity to parameters determining the variance of distributions.

CHAPTER 7. EXPERIMENTS 8 and 9: RESPONSE VARIANCE

Contents

7.1 Experiment 8: Coherence interval width and response confidence

7.1.1 Varying location and width of coherence intervals

7.1.2 Measuring people's sensitivity to location and width

7.1.3 Method

7.1.4 Results and discussion

7.1.5 General discussion

7.2 Experiment 9: Sensitivity to the variance of distributions

7.2.1 Method

7.2.2 Results and discussion

7.2.3 General discussion

EXPERIMENT 8: COHERENCE INTERVAL WIDTH AND RESPONSE CONFIDENCE

Investigating the constraint of coherence, it makes sense to ask not only whether people are sensitive to the location of a coherence interval on the probability scale, but also whether they respond differentially as a function of the width of the interval. To the knowledge of the author this has not been attempted before. Experiments 3 and 4 assessed this question by examining differences in the standard deviation of responses as a function of interval width, but found no relation between the two. This form of assessment was based on the hypothesis that there may be a higher variability in responses when the interval is wider.

Experiment 8 addressed the question of a sensitivity to the location and width of coherence intervals using different, further methods. People's sensitivity to the location of an interval was assessed by computing above-chance coherence and by examining the pattern of mean conclusion probability judgments, as in Experiments 3 and 4, but also by examining the distribution of conclusion probability judgments relative to the location of the interval. People's sensitivity to the width of the interval was assessed by examining the pattern of mean conclusion probability judgments, and the distribution of conclusion probability judgments as a function of interval width. Similar to the examination of standard deviations of responses in Experiments 3 and 4, these two methods were based on the assumption that wider coherence intervals are associated with a higher response variance. A third method of assessing people's sensitivity to interval width was based on the idea that the width of the interval may affect people's confidence in their probability judgments.

Experiment 9 was a control experiment to help interpret the results of Experiment 8. It tested whether people are sensitive to changes in the variance of a frequency distribution, using the same response variables as Experiment 8: probability judgments and judgments of response confidence. If in Experiment 9 people's confidence in their probability judgments do not change as a function of the variance of a frequency distribution, then it will be more difficult to attribute any association between judgments of response confidence and interval width in Experiment 8 to a sensitivity to changes in the number of coherent response options.

Experiment 8 investigated five inferences that differed in how the width and location of their coherence intervals changes as a function of the probabilities of the premises. The inferences were MP, DA, and-to-if (&If, also called *two-premise centering*), de morgan (DM), and the negation of de morgan (nDM). Their logical form and the function for their coherence interval are shown in Table 7.1.

Varying location and width of coherence intervals

Figure 7.1 illustrates the way in which the coherence intervals for inferences 1 to 3 change as a function of their premise probabilities. One can see that for MP and DA, the probability of the major premise determines the location of the interval, and the probability of the minor premise determines the width of the interval. However, the effect of premise probabilities on the coherence interval for the conclusion goes in opposite directions for the two inferences. For MP, the higher the probability of the major premise, the higher the location, and the higher the probability of the minor premise, the smaller the width. For DA, in contrast, the higher the probability of the major premise, the lower the location, and the higher the probability of the minor premise, the larger the width.

Table 7.1. The inferences used in Experiment 8.

#	Name	Form	Coherence interval for the probability of the conclusion
1	Modus ponens (MP)	$if\ p\ then\ q,\ p\ \therefore\ q$	$P(q) \in [P(q p)P(p), P(q p)P(p) + (1 - P(p))]$
2	Denial of the antecedent (DA)	$if\ p\ then\ q,\ not-p\ \therefore\ not-q$	$P(\neg q) \in [(1 - P(q p))(1 - P(\neg p)), 1 - (P(q p)(1 - P(\neg p)))]$
3	And-to-if (&If)/Two-premise centering	$p,\ q\ \therefore\ if\ p\ then\ q$	$P(q p) \in [\max\left(0, \frac{P(p) + P(q) - 1}{P(p)}\right), \min\left(\frac{P(q)}{P(p)}, 1\right)]$
4	de morgan (DM)	$not(p\ \&\ q)\ \therefore\ not-p\ or\ not-q$	$P(\neg p\ or\ \neg q) = 1 - P(p\ \&\ q)$
5	not de morgan (nDM)	$p\ or\ q\ \therefore\ not-p\ \&\ not-q$	$P(\neg p\ \&\ \neg q) = 1 - P(p\ or\ q)$

For &If, it is the consequent premise q that determines the location of the interval, whereas the antecedent premise p determines interval width. The direction of the effect is thereby similar to that of MP. The higher the probability of the consequent premise, the higher the location of the interval, and the higher the probability of the antecedent premise, the narrower the interval.

The last two inferences in Table 7.1 have only one premise. The first is an equivalence and the second a contradiction. Therefore, for these two inferences the coherence interval for the conclusion will always be a point value. In the case of the equivalence the point value will

equal the probability of the premise, and in the case of the contradiction it will equal the complement of the probability of the premise.

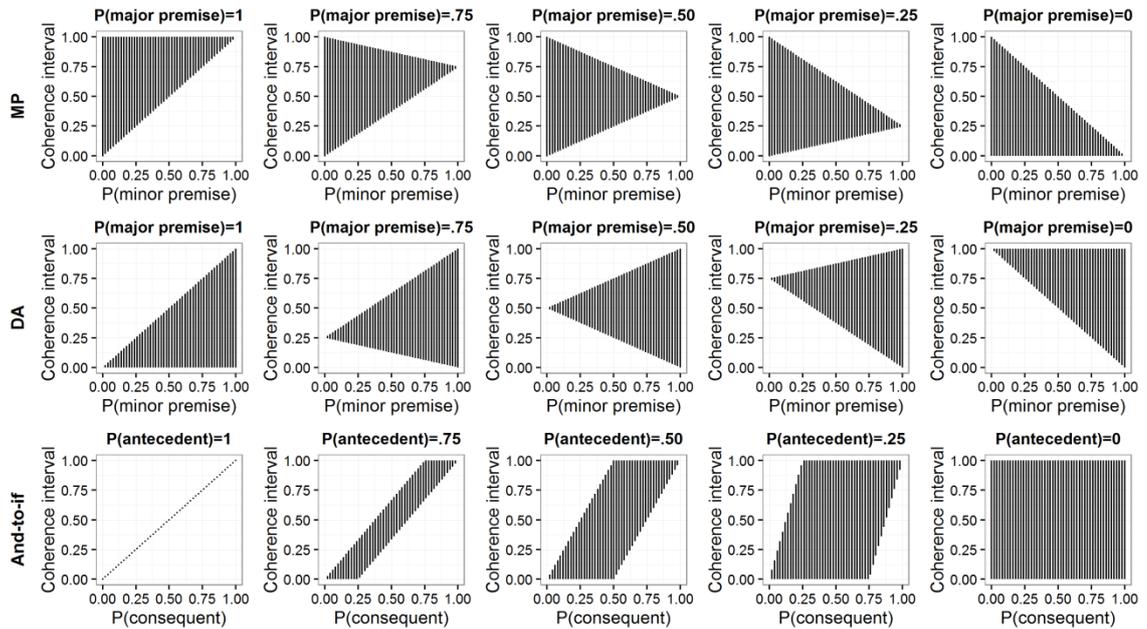


Figure 7.1. Coherence intervals for MP (upper row), DA (middle row), and and-to-if (lower row) as a function of premise probabilities. The shaded areas in the graphs represent the coherence intervals.

Measuring people's sensitivity to location and width

The contrasting relation between the probabilities of the premises and the coherence intervals for the conclusion of these inferences provides a novel opportunity to test the degree to which people are sensitive to coherence intervals. This was done using four methods, two of them targeting both the location and the width of the interval (distribution and mean values of conclusion probability judgments), one targeting only the location (above-chance coherence), and one targeting only the width (response confidence).

Distribution of conclusion probability judgments

The first method consisted in examining the distribution of conclusion probability judgments relative to the width and the location of coherence intervals. It involved assessing whether the location of response distributions followed changes in the location of the interval, and whether the distributions were flatter when the interval was wider, and more peaked when the interval was narrower. This comparison of the distribution of conclusion probability judgments to the coherence interval for an inference was not possible in Experiments 1 to 4, because the fact that in these experiments participants were asked to provide their own estimates of the

probability of the premises implied that the coherence interval differed for each participant response. It was also not possible in Experiments 5 to 7, because these experiments used a binary response scale. Experiment 8 allowed for this by using a continuous response scale while at the same time fixing the premise probabilities that determine the coherence interval to a limited number of values.

The examination of response distributions was descriptive, not inferential: an inferential analysis based e. g. on kurtosis would have been difficult because the distributions were not always unimodal. Beyond the assessment of whether the overall location and width of response distributions varied with the coherence interval, an inspection of response distributions can help uncover possible further factors affecting responses, which may then feed into hypotheses for future experiments.

Above-chance coherence

The second method used to assess people's sensitivity to the location of coherence intervals was the computation of above-chance coherence, using the same procedure as in the previous experiments. The aim here was not to compare coherence between inferences, but merely to establish whether coherence was above chance levels for each inference. This is because any results on interval width can only be interpreted as sensitivity to coherence constraints if people's responses are reliably coherent in the first place.

Mean conclusion probability judgments

The third method of assessing people's sensitivity to the location and width of coherence intervals was the analysis of differences in mean conclusion probability judgments as a function of interval width and location. The predictions for these differences, derived from the patterns observed in Figure 7.1, were as follows.

For MP, sensitivity to the location of the interval leads to the expectation of higher conclusion probability judgments, the higher the probability of the major premise. Taking into account the width of the interval, this effect is expected to be larger when the probability of the minor premise is also high.

For DA, sensitivity to the location of the interval leads to the expectation of higher conclusion probability judgments, the lower the probability of the major premise. Taking into account the width of the interval, this effect is expected to be larger when the probability of the minor premise is also low.

For &If, sensitivity to the location of the coherence interval leads to the prediction that conclusion probability judgments will be higher, the higher the probability of the consequent premise. Taking into account the width of the interval suggests that this effect will be larger, the higher the probability of the antecedent premise.

For DM and nDM, where the coherence interval is a point value, sensitivity to coherence simply leads to the prediction that, for DM, mean conclusion probability judgments will be higher when the probability of the premise is high, and that for nDM, mean conclusion probability judgments will be lower when the probability of the premise is high.

Judgments of response confidence

The third method of assessing people's sensitivity to the width of coherence intervals was based on the idea that interval width may have an effect on people's judgments of response confidence. For MP and DA, one would expect this effect to take the form of a correspondence between response confidence and the probability of the minor premise, because it was the minor premise that determined interval width.

For &If, one would instead expect a correspondence between response confidence and the probability of the antecedent premise p , because it was this premise that determined interval width.

Finally, no correspondence between response confidence and premise probability would be expected for DM and nDM, given that interval width was independent of premise probability for these inferences.

A correspondence between premise probability and response confidence could be positive or negative. In either case it would constitute an additional source of evidence of sensitivity to coherence constraints, but the direction of the effect is of theoretical interest in itself. A positive relation between response confidence and interval width, such that response confidence increases as the interval widens, would suggest that participants' aim in responding is to lie within the interval, and that the exact position within it is less important. That is, it would suggest that participants are mainly trying to be coherent, as opposed to trying to solve also the inductive problem of where best to place their response within the interval. This is because the wider the interval, the easier it is for any given value to lie within it.

Alternatively, there could be a negative relation between response confidence and interval width, such that response confidence decreases as the interval becomes wider. This would suggest that participants are engaging in both the deductive task of rendering their responses coherent, and the inductive task of finding the most plausible point estimate, given the information they have, among those points that are coherent. Finding the single most plausible estimate is more difficult when there are more options available (i. e. when the interval is wide) without more information on which to base one's decision. A search for the most plausible estimate can be viewed as a search for the mean or mode of a distribution, as opposed to a mere search for a region of permissible responses. The search for an optimal point value may have been suggested by the fact that the task required participants to provide a point estimate.

The hypothesis about a relation between interval width and response confidence was undirected in this experiment. If a relation is found, then this could be examined further in follow-up experiments with a directed hypothesis, in which one could also try to test aspects of the possible interpretations given above.

Method

Participants

A total of 45 participants from the participant pool of Birkbeck, University of London completed the experiment. Of these, 4 were removed because they failed a catch trial asking them not to respond but to instead just click "next" to continue with the experiment. Further 5 participants were excluded because they had one or more trial reaction times of less than 3 seconds. The final sample consisted of 36 participants. They had a mean age of 35.32 years (range 20-76), and most had undergone some college education: 47.2% reported having an undergraduate university degree, and 44.4% a postgraduate degree (2.8% reported having a technical/applied degree, and 5.6% to have finished 12th grade). All participants indicated having at least "very good" English language skills. The mean percentage rating of experiment difficulty was 67%.

Design and material

The experiment followed a within participant design with inference (the five inferences in Table 7.1) and premise probability as independent variables, and with point probability estimates for the conclusion, and confidence judgments in these estimates (both measured in percent), as dependent variables.

The premise probabilities used were .9 (high), .5 (medium), and .1 (low). For the two-premise inferences, there were 9 different combinations of premise probability (high-high, high-medium, high-low, medium-high, medium-medium, medium-low, low-high, low-medium, and low-low). Each two-premise inference was presented nine times, once with each combination of premise probabilities. Each one-premise inference was also presented nine-times, three times with each premise probability. Trials with the same one-premise inference and the same probability were still distinguishable from one another through changes in the non-words used in their contents.

Participants were introduced to the task with a pseudonaturalistic scenario about researchers investigating bird species on an island, shown in the frame below. Participants then worked through four practice trials involving different inferences to those tested. With 5 inferences and 9 probabilities, the main experiment had 45 trials, plus two catch trials to ensure participants were paying attention.

Imagine you are part of a team of researchers investigating the birds on the island of Liaku. You want to find out what kind of seeds different species on the island eat. For this, you have attached cups with different seeds to some trees on the island, together with a video camera to monitor which birds eat which kind of seed. The camera does not work very well at night, so that some of your images are more precise than others. You are trying to interpret the data you gathered with the team.

On each trial, participants were presented with an inference and information on the premise probabilities. They were then asked to judge how likely the conclusion could be, by clicking on a percentage scale with the anchors "0% likely, certainly false" and "100% likely, certainly true".

As noted in Experiment 3, the use of a continuous response scale makes it difficult to compare response coherence between inferences and between premise probabilities. However, this was not a problem in the present experiment, which was not so much concerned with a comparison between inferences as with a general assessment of whether responses followed coherence constraints across inferences. Such an assessment is necessary to be able to interpret the results on interval width, in particular those on response confidence, as resulting from coherence constraints. Given this situation, a continuous scale was preferred over a binary one because it made it possible to assess the distribution of responses. It also seemed more natural because it allowed participants to generate their own probability judgments, as opposed to evaluating probabilities only given by the experimenter.

The observations you have gathered until now suggest the following:
Premise 1: It's 90% likely that:
If the next Baila bird you film eats aib seeds, then it will eat dun seeds.

Premise 2: It's 50% likely that:
The next Baila bird you film will not eat aib seeds.

Conclusion: Therefore, how likely can the following be?
The next Baila bird you film will not eat dun seeds.

How much confidence do you have in your answer?

A second question on the same page asked how much confidence participants had in their answer, using a percentage scale with the anchors "no confidence at all" and "complete confidence". An example of a trial for the DA inference is provided in the frame above.

Each trial featured different non-words for the birds and seeds, and the non-words were allocated to the trials randomly for each participant. The order of the trials was also varied randomly for each participant.

The catch trials had the same format as an ordinary trial, but the text of the inferences to be evaluated was replaced with a statement saying they were a control trial to make sure participants were paying attention, and asking them not to respond, but to instead click *next* to continue with the experiment.

Procedure

Participants were tested individually in a quiet testing room of the Department of Psychological Sciences of Birkbeck, University of London. The experimenter stayed in the room while the participants went through the instructions and practice trials in case they had any questions. The entire experimental session took approximately 40 minutes to complete.

Results and discussion

The results are divided into four sections, concerned with the distribution of conclusion probability judgments, above-chance coherence, mean conclusion probability judgments, and judgments of response confidence, respectively.

Distribution of conclusion probability judgments

The use of a continuous response scale makes it possible to examine the distribution of conclusion probability judgments in terms of the extent to which they appear sensitive to changes in the location and width of coherence intervals. In addition, an exploratory examination of response distributions can sometimes provide relevant information about possible factors affecting responses, and so help construct hypotheses for subsequent experiments.

The distributions for each inference are displayed in Figures 7.2 to 7.5. The horizontal lines beneath the distributions indicate the location of the coherence interval that matches their colour. Note that the y axis has a maximum of 5 in Figures 2 and 5, but a higher maximum in Figures 3 and 4 to accommodate the sharp peak in responses in the latter conditions.

Figure 7.2 shows the distributions of responses for MP. Overall, the peaks of the distributions fall within the coherence intervals, albeit leaned somewhat towards their lower bounds. In the middle and the right panel there also seems to be a correspondence between interval width and distribution variance: the distributions have sharper peaks when the interval is narrow. In contrast, in the left panel, where the probability of the major premise was .1, mean responses were generally low regardless of the probability of the minor premise.

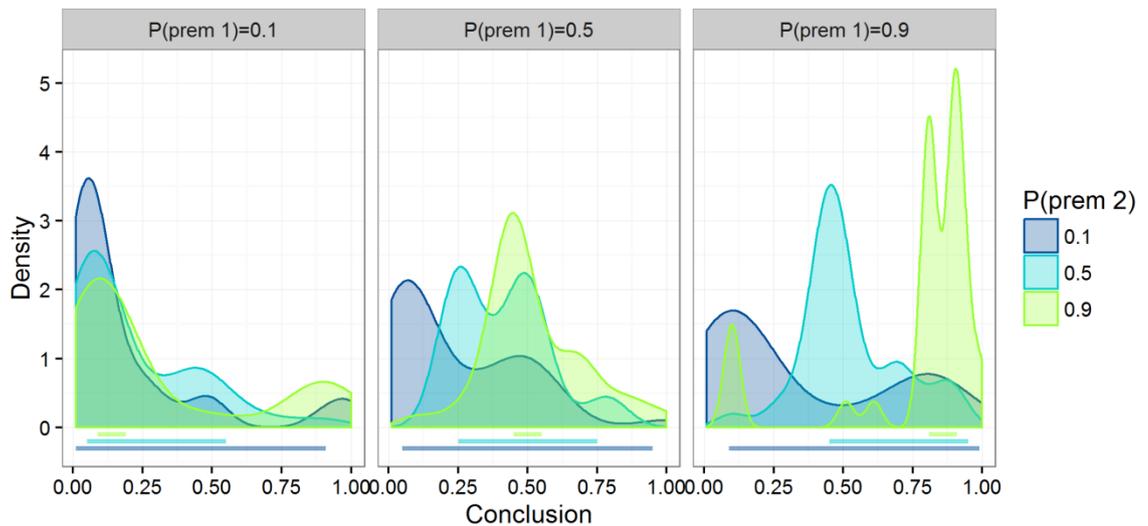


Figure 7.2. Distribution of conclusion probability judgments for MP as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals.

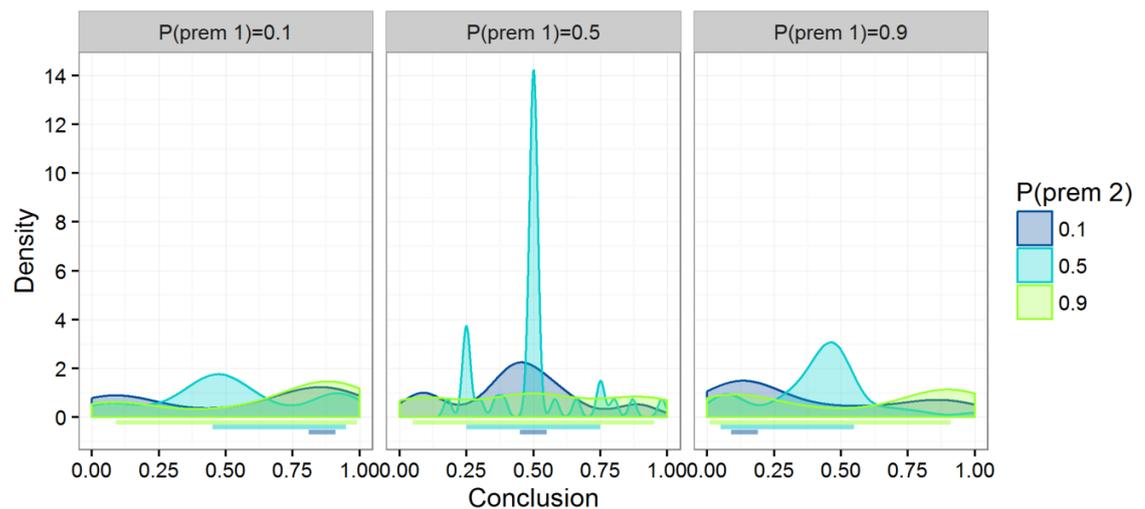


Figure 7.3. Distribution of conclusion probability judgments for DA as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals.

Figure 7.3 displays the distributions of responses for DA. One can see that there was also a correspondence between the distribution peak and the location of the coherence interval. However, a correspondence between distribution variance and interval width appears to be present only when the interval was very wide, being reflected in flat distributions for these cases. Further, there seemed to be a matching effect such that when both premises had a probability of .5, the majority of responses were also close to .5. This tendency to match had no negative effect on the coherence of responses because a conclusion probability of .5 was coherent when both premises had a probability of .5.

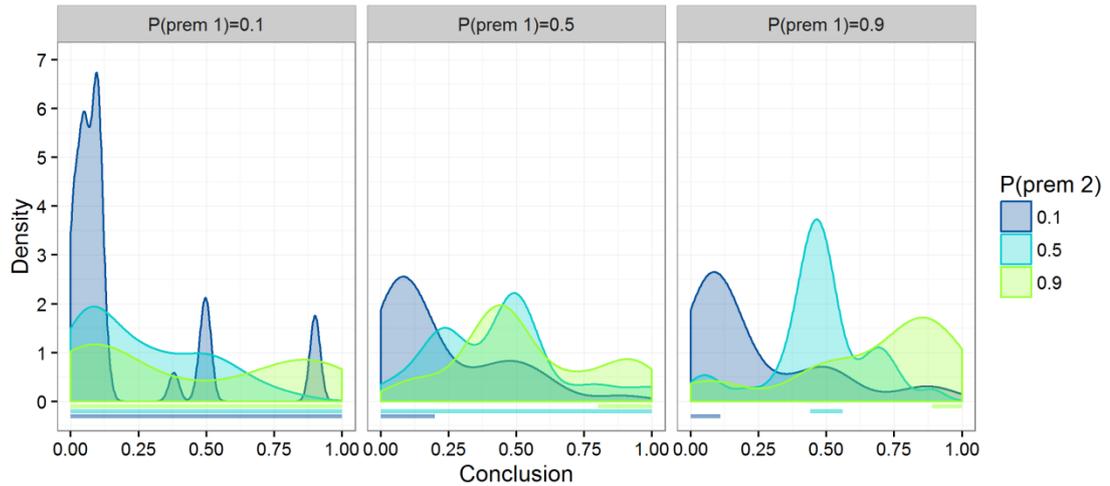


Figure 7.4. Distribution of conclusion probability judgments for &If as a function of premise probabilities. The horizontal lines beneath the distributions indicate the location of the respective coherence intervals.

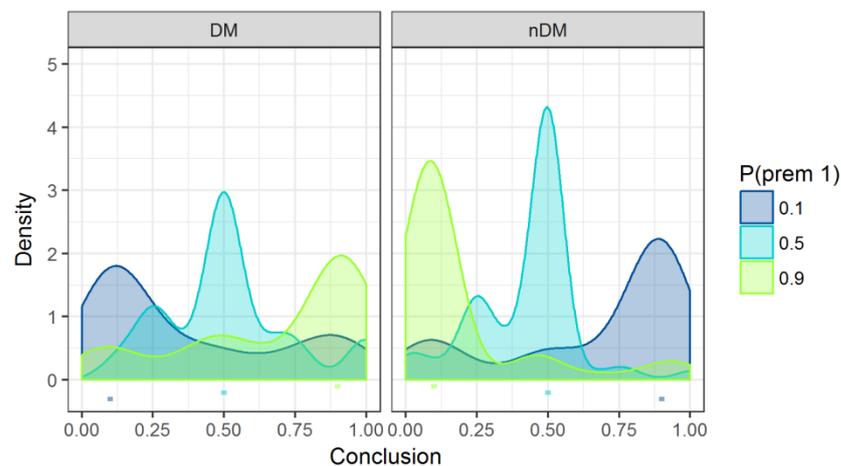


Figure 7.5. Distribution of conclusion probability judgments for the inferences of DM and nDM, as a function of premise probabilities. The horizontal lines below the distributions show the location of the coherence interval for the condition that matches their colour.

The distributions in Figure 7.4 are for &If. As in the previous figures, the distribution peaks generally lie within the coherence interval. An exception was the condition in which the antecedent premise had probability .5 but the consequent premise had probability .9. In this condition, participants tended to assign to the conclusion a probability near .5, even though the coherence interval demanded a probability nearer to .9.

Of particular interest in Figure 7.4 is the contrast observed between the distributions in the left and right panels. In the left panel, the coherence interval was the unit interval, and so any response was coherent. If people's responses were only guided by coherence, one would

expect a flat distribution in this case, as was found for DA. Instead, we see that judgments of conclusion probability were more often on the lower than on the upper half of the scale, and that the form of the distributions differed strongly between conditions. Given the absence of deductive constraints, the differences in the form of the distributions must have been determined by inductive considerations.

In contrast, in the right panel, where the coherence intervals are very narrow, the distributions are very similar to one another and seem to be guided mainly by the location of the intervals.

The same pattern as in the right panel of Figure 7.4 was found in Figure 7.5, which shows the distributions of responses for DM and nDM, whose coherence intervals are point values. Here too we see that although many responses missed the only coherent value on the scale, their distribution appeared to be determined mainly by the location of this value, which was inverted for the valid and the invalid inference.

This again suggests that the scope for inductive determinants of conclusion probability judgments is limited by the scope of deductive considerations of coherence: inductive considerations can play a large role when there is room for them, as in the right panel of Figure 7.4, but the deductive constraint of coherence seems to take precedence.

The pattern in Figure 7.5 is slightly more pronounced for nDM than for DM, in line with the finding from Experiments 5 and 6 of more frequent coherent responses for invalid than for valid one-premise inferences.

Above-chance coherence

The procedure for computing observed and above-chance coherence was the same as in the previous experiments. The mean values of observed and above-chance coherence for each inference are displayed in Figure 7.6. The figure collapses responses across premise probability conditions because as outlined in Experiments 3 and 4, it is difficult to compare coherence rates across these conditions when they differ in chance rate coherence. The analysis of above-chance coherence in this experiment is limited to the question of whether coherence was above-chance levels at all for each inference.

The confidence intervals in Figure 7.6 indicate that coherence was indeed above-chance for all inferences. The fact that the subtraction of the chance rate from the observed rate had a far larger effect for MP, DA, and &If than for DM and nDM makes sense given that the width of the interval (and with it the chance rate) for DM and nDM was a point value.

The pattern in Figure 7.6 was corroborated in a linear mixed model for the effect of inference on above-chance coherence, with random intercepts for participants and scenarios: overall responses were coherent above-chance levels ($F(1, 180) = 181.75, p < .001$). The degree of above-chance coherence differed between inferences ($F(4, 180) = 3.67, p = .007$). However, individual comparisons showed that it was significant for each inference (for MP:

$F(1, 36) = 43.47, p < .001$; for DA: $F(1, 36) = 47.60, p < .001$; for &If: $F(1, 36) = 59.73, p < .001$; for DM: $F(1, 36) = 27.47, p < .001$; for nDM: $F(1, 36) = 51.61, p < .001$).

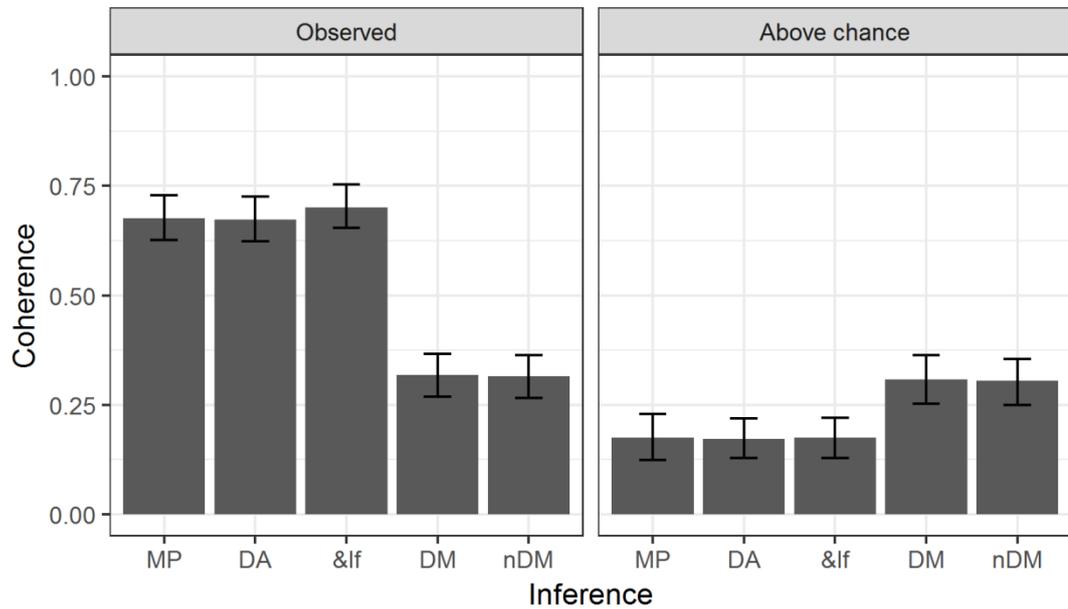


Figure 7.6. Observed and above-chance coherence for the 5 inferences of the Experiment. Error bars show 95% CIs.

Mean conclusion probability judgments

A series of linear mixed models assessed, for each inference, the effects of premise probabilities on conclusion probability judgments, with random intercepts for participants and scenarios. Mean conclusion probability judgments for each condition are shown in Figures 7.7 and 7.8.

For MP, conclusion probability judgments were higher for higher values of the major premise ($F(2, 288) = 50.35, p < .001$); and for higher values of the minor premise ($F(2, 288) = 40.40, p < .001$). The effect of the probability of the major premise on conclusion probability judgments was stronger when the probability of the minor premise was also high ($F(4, 288) = 5.12, p = .001$). These findings are in accordance with the predictions.

For DA, conclusion probability judgments were lower for higher values of the major premise ($F(2, 288) = 5.27, p < .001$), in accordance with the predictions. However, conclusion probability judgments were higher for higher values of the minor premise ($F(2, 288) = 4.28, p = .02$). If anything this effect was expected to go in the opposite direction. There was no interaction between the effects of the first and the second premise on conclusion probability judgments ($F(4, 188) = .55, p = .70$), contrary to the predictions. This pattern of results suggests that participants were sensitive to the location of the coherence interval for this inference, but not to interval width – an interpretation in line with the response distributions shown in Figure 7.3.

For &If, conclusion probability judgments were higher for higher values of the consequent premise ($F(2, 288) = 55.32, p < .001$). Conclusion probability judgments were also higher for higher values of the antecedent premise ($F(2, 288.18) = 13.29, p < .001$). The interaction between the effects of the first and the second premise on conclusion probability judgments went in the expected direction, but did not reach significance ($F(4, 288.18) = 2.03, p = .09$). Overall these results are in accordance with the predictions.

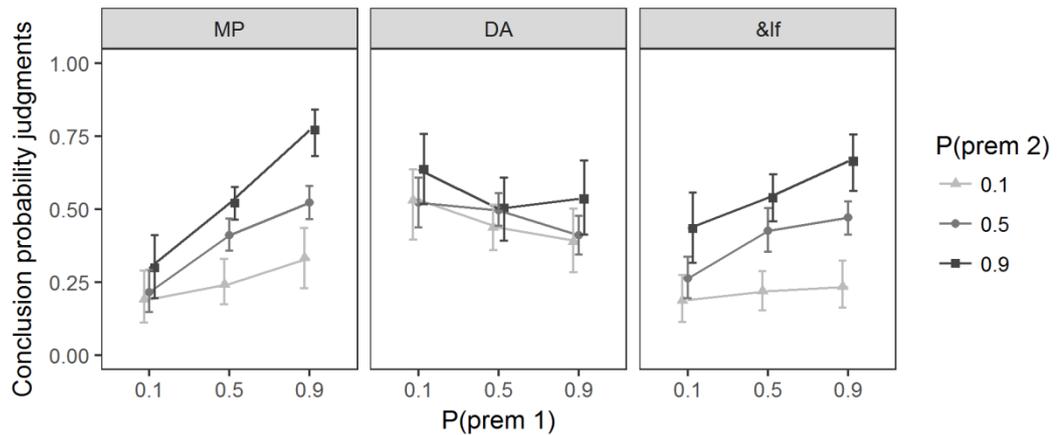


Figure 7.7. Conclusion probability judgments for MP, DA, and &If, as a function of premise probabilities. Error bars show 95% CIs.

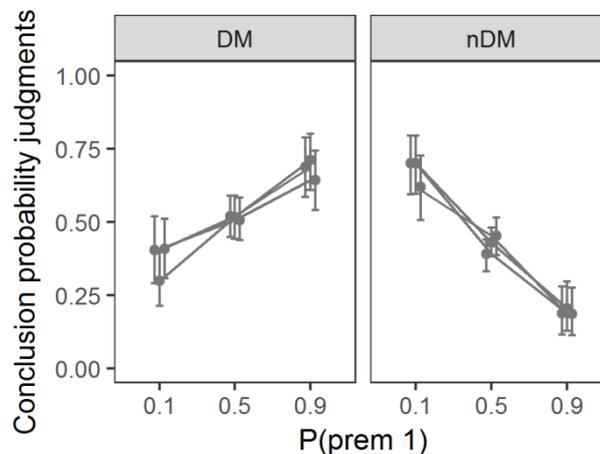


Figure 7.8. Conclusion probability judgments for DM and nDM as a function of premise probabilities. The three lines in each panel display the three repetitions of each premise probability condition for these inferences. Error bars show 95% CIs.

For DM, conclusion probability judgments were higher, the higher the probability of the premise ($F(2, 288) = 36.77, p < .001$). For nDM, conclusion probability judgments were lower, the higher the probability of the premise ($F(2, 288) = 115.72, p < .001$). The findings

for both inferences are in accordance with what would be predicted on the basis of sensitivity to coherence. But they provide no evidence for a role of interval width over and above a role of interval location.

Overall, the results for mean conclusion probability judgments suggest that people were reliably sensitive to the location of the coherence interval. Evidence for a sensitivity to interval width was more equivocal, being present for MP and marginally for &If, but not for DA.

Judgments of response Confidence

The relation between interval width and judgments of response confidence was assessed in a series of linear mixed models for each inference, with random intercepts for participants and scenarios. Mean judgments of response confidence for each inference and premise probability condition are shown in Figures 7.9 and 7.10.

For MP, there was no relation between response confidence and the probability of the minor premise ($F(2, 288) = 4.82, p = .009$), contrary to the hypothesis. However, there was a relation between response confidence and the major premise ($F(2, 288) = 4.82, p = .009$), which was qualified by an interaction between the effects of the major and the minor premise ($F(4, 288) = 3.38, p = .01$). Follow-up analyses showed that when the probability of the major premise was .9, confidence was higher for higher values of the probability of the minor premise ($F(2, 72) = 9.36, p < .001$). But there was no effect of minor premise probability on confidence judgments when the probability of the major premise was .5 ($F(2, 72) = 1.99, p = .14$) nor when the probability of the major premise was .1 ($F(2, 72) = .13, p = .88$). Overall, the findings provide no evidence for a relation between interval width and response confidence for MP.

For DA, no effects were significant (main effect of the minor premise: $F(2, 288) = 2.496, p = .08$; main effect of the major premise, $F(2, 288) = 2.38, p = .10$; interaction: $F(4, 288) = .924, p = .45$). Hence there was also no evidence for a relation between interval width and response confidence for this inference.

For &If, there was no relation between the antecedent premise and response confidence ($F(2, 288) = 1.88, p = .16$), nor a relation between the consequent premise and response confidence ($F(2, 288) = 1.29, p = .28$). An interaction between the effects of the probability of the antecedent and the consequent premise ($F(4, 288) = 3.46, p = .009$) indicated that there was no relation between the probability of the antecedent premise and response confidence when the probability of the consequent premise was .9 ($F(2, 72) = 1.09, p = .34$) nor when it was .1 ($F(2, 72.178) = 2.16, p = .122$). However, when the probability of the consequent premise was .5, then response confidence was higher when the probability of the antecedent premise was also .5 (for the difference between the conditions with probabilities of .5 and of .1: $p = .024$; for the difference between the conditions with probabilities of .5 and of .9: $p = .001$; for the difference between the conditions with probabilities of .1 and of .9: $p = .59$).

Hence there was also no evidence for a relation between interval width and response confidence for &If.

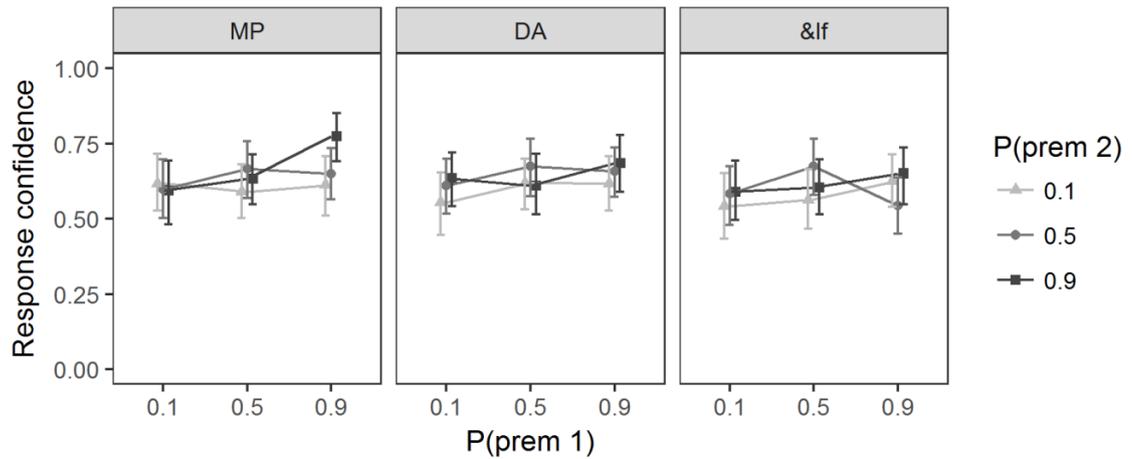


Figure 7.9. Mean judgments of response confidence for MP, DA, and &If, as a function of premise probabilities. Error bars show 95% CIs.

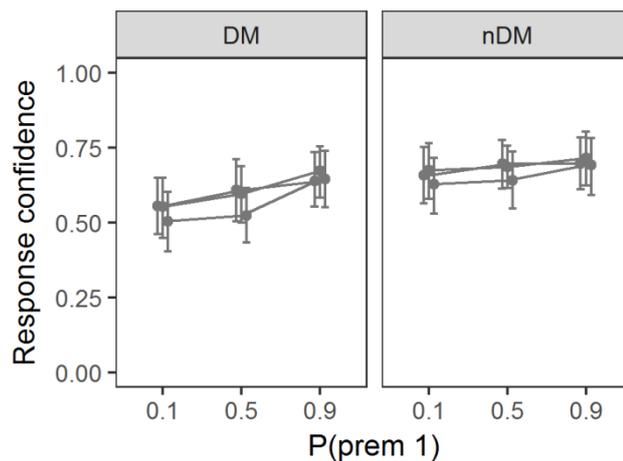


Figure 7.10. Mean judgments of response confidence for DM and nDM as a function of premise probabilities. The three lines in each panel display the three repetitions of each premise probability condition for these inferences. Error bars show 95% CIs.

For DM and nDM, whose coherence intervals are a point value, no relation between premise probability and response confidence was predicted. For DM, response confidence was higher for higher values of the probability of the premise ($F(2, 288) = 10.78, p < .001$). For nDM no relation between premise probability and response confidence was observed ($F(2, 288) = 1.81, p = .17$).

It was not possible to calculate random slopes for participants in these analyses. However, a graphical examination of the pattern of individual responses for each inference gave no indication of the presence of subgroups with different profiles that could have cancelled out an effect in the overall analysis.

In summary, confidence judgments varied with premise probability in some conditions and in different ways, but they were not found to be related to the width of the coherence interval.

General discussion

There was consistent evidence that people are sensitive to the location of the coherence intervals of the five inferences investigated, across the inspection of response distributions, the analysis of above-chance coherence, and the analysis of the pattern of mean conclusion probability judgments.

The evidence from conclusion probability judgments for a sensitivity to the width of the coherence interval was more mixed. The response distributions suggested that when the interval was very narrow, as in the right panel of Figure 7.4 and in Figure 7.5, responses are guided mainly by the location of the interval, with little difference in the form of response distributions. When the interval is wider, there is more space for inductive criteria to play a role in narrowing down responses to a certain area on the probability scale. Such inductive criteria sometimes play a role, as in the left panel of Figure 7.4, but not invariably so – as exemplified by the flat distributions in Figure 7.3 for the wide interval conditions, and by the distributions in Figure 7.2. Coherence intervals seem to constrain responses to within their range, and within those constraints, inductive criteria may or may not narrow down responses further in a systematic way, depending on the contents and context of the inference.

The pattern of mean conclusion probability judgments provided evidence for a role of the width of the coherence interval for MP, no role for DA, and a marginal role for &If, consistent with the observed response distributions for these inferences.

In general, there was reliable evidence of sensitivity to the location of coherence intervals, but only partial evidence of sensitivity to interval width. However, a lack of sensitivity to interval width by no means implies a lower sensitivity to coherence. Rather, it suggests the hypothesis that coherence is an important constraint that takes precedence over inductive considerations, but that within the constraints of coherence, the distribution of responses may be narrowed down to a higher or lower degree by inductive factors – provided that the coherence interval is wide enough for this.

However, an alternative interpretation of the lack of a consistent effect of interval width on either conclusion probability judgments or judgments of response confidence is that people

have general difficulties in processing information on the variance of a distribution. This alternative was investigated further in Experiment 9.

EXPERIMENT 9: SENSITIVITY TO THE VARIANCE OF DISTRIBUTIONS

This experiment set out to investigate whether people are sensitive to the variance of a distribution, using the same outcome variables as Experiment 8: probability judgments and response confidence. It was hypothesised that if people are sensitive to the variance of a distribution, they should be sensitive to changes in factors that affect this variance, such as sample size.

The experiment had two conditions, a *sample condition* and a *population condition*. In the sample condition, participants' judgments of the probability that an instance from a population had a certain feature, given the proportion of instances with that feature in a sample, was expected to equate this proportion, regardless of sample size. The reason is that, in the absence of further information, the proportion in the sample is the best estimate of the proportion in the population, even if it is only a small sample. However, participants' confidence in the above probability judgments was expected to be higher for higher values of sample size. For larger sample sizes, the proportion of cases with a certain feature in a sample is a better estimator of the proportion of instances with that feature in the population, given the lower error variance it is associated with.

The population condition was a control condition in which participants' were asked to judge the likelihood that an instance from a population has a certain feature, as before, but the information provided as a basis for this judgment was not the proportion in a sample, but the proportion in the population itself. Participants' probability judgments in this condition were expected to equate the proportion in the population, as before. But participants' confidence in the above probability judgments was not expected to vary as a function of population size, because regardless of the size of the population, the information on which to base probability judgments is equally precise. This condition was included in the design to control for possible confounding factors related to the size of a set itself, whether the set refers to a sample or to a population. For example, it could be that in general, people are more confident when making judgments about small than about large set sizes. Such an effect would be visible in the population condition, and could then be used to help interpret the results for the sample condition.

Method

Participants

A total of 255 participants completed the online experiment (125 in the sample condition and 130 in the population condition). All indicated at the end of the experiment that they took part seriously, as opposed to just clicking through, but 6 were excluded because they failed a catch trial asking them not to respond but to instead click next to continue with the experiment. Seven participants were excluded because they had trial reaction times of 3 seconds or less; and five because they indicated having less than "good" English language skills. The final sample had 237 participants (114 in the sample group and 123 in the population group). Their mean age was 36 years (median = 33, range = 18 - 74). Most of them had some college education, with 19.4% indicating that they finished 12th grade, 10.5 reporting a technical/applied degree, 47.7 an undergraduate university degree, and 15.2 a postgraduate degree (5.5% reported a degree lower than 12th grade, and 1.3% a doctoral degree). Participants' median rating of task difficulty was 14% in the sample group, and 9% in the population group.

Design and materials

The experiment followed a mixed design with condition (sample, population) as between participant variable, and set size (10, 100, 1000) and proportion (.9, .7, .5) as within participant variables. With three set sizes and three proportions, the experiment consisted of 9 trials, plus a catch trial that was similar in format to the other trials, but in which the content of the statement to be evaluated was replaced by a statement saying that it was a control question to make sure participants' were paying attention, and asking them not to respond but to instead just click *next* to continue with the experiment.

Participants were given a short pseudonaturalistic scenario in which a protagonist wanted to find out the proportion of a certain feature in a population, and gathered information on nine occasions for this. The set size and proportion information for each trial corresponded to the information gathered by the protagonist on a particular occasion. Participants' task was to assess how likely it was for an instance of the set to have a certain feature. This judgment was made on a continuous scale with the anchors "0% likely/certainly false", and "100% likely/certainly true". Further below on the same page, participants were asked to rate how much confidence they had in their answer. This judgment was also made on a continuous scale, this time with the anchors "no confidence at all", and "complete confidence". The two response scales differed in their visual appearance to prevent carry-over effects from one to the other. The frame below is an example of a trial in the sample and in the population conditions, for a set size of 10 and a proportion of .9:

Sample:

A tree disease has spread to the orange plantations of the farmers of Orisau. An agronomist went to some of the fields and took random samples of trees on each field, to record the number of affected trees among them.

On Field one the agronomist took a sample of 10 trees, and observed that 9 of them were affected.

How likely was it for a random tree on Field one to be affected?

How much confidence do you have in your answer?

Population:

A tree disease has spread to the orange plantations of the farmers of Orisau. An agronomist went to some of the fields and recorded the number of trees on each field, and the number of affected trees among them.

Field one had 10 trees and 9 of them were affected.

How likely was it for a random tree on Field one to be affected?

How much confidence do you have in your answer?

Within each participant, the scenario remained constant across the 9 trials. There were 9 scenarios in total, which were randomly allocated to each participant, and can be found in Appendix F. The order of occurrence of the set size and proportion information also varied randomly for each participant.

Procedure

Participants received general instructions followed by three practice trials, involving a different scenario and partly different proportions from those in the main experiment. After going through the nine experimental trials, participants provided demographic information, and indicated whether they had taken part seriously or had just "clicked through". The final page provided debriefing information. The experimental session lasted on average 5.6 minutes.

Results and discussion

Mean probability judgments and judgments of response confidence are displayed in Figure 7.11 for each condition. The pattern of responses is in accordance with the predictions: probability judgments closely conformed to the proportions given in both the sample and the population condition. Whereas confidence judgments were uniformly high in the population condition regardless of set size, confidence judgments in the sample condition seemed to increase with increasing sample size.

The pattern in Figure 7.11 was assessed in two linear mixed model analyses for the effects of sample size, proportion, and condition (sample vs. population), with a random intercept for participants. The first model assessed the effect of these variables on probability judgments, and the second on judgments of response confidence.

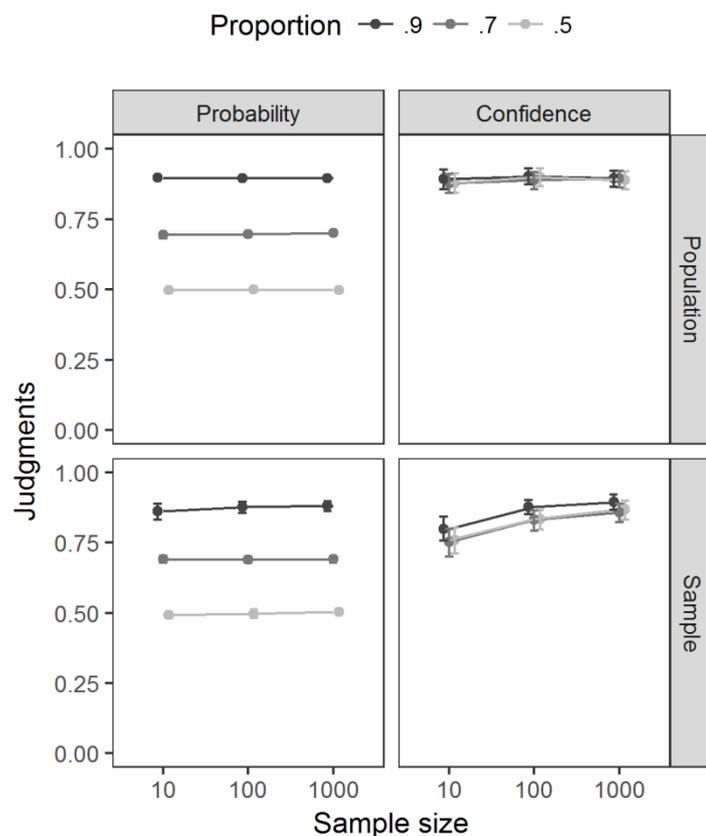


Figure 7.11. Mean probability judgments and judgments of response confidence for the conditions of Experiment 9. Error bars show 95% CIs.

Probability judgments

The analysis for probability judgments revealed no main effect of sample size ($F(2, 1896) = 1.866, p = .16$), and sample size did not enter into any interaction. This is in accordance with the predictions.

Trivially, there was a main effect of proportion ($F(2, 1896) = 78943.94, p < .001$), such that probability judgments were higher when the proportion of instances for which a certain feature held was higher. There was also a marginal effect of condition ($F(1, 237) = 4.44, p = .036$): probability judgments were slightly higher in the population ($EMM = .70$) than in the sample condition ($EMM = .69$). Further, there was an interaction between proportion and condition ($F(2, 1896) = 6.77, p = .001$): For a proportion of .9, judgments were slightly lower in the sample condition ($EMM = .874$) than in the population condition ($EMM = .897$). Judgments for a proportion of .7 and of .5 were almost identical in both conditions (for a proportion of .7: $EMM_{popu} = .698$; $EMM_{sample} = .691$. For a proportion of .5: $EMM_{popu} = .499$; $EMM_{sample} = .498$).

Confidence judgments

The analysis for confidence judgments revealed a main effect of condition: mean confidence judgments were higher in the population condition than in the sample condition ($F(1, 237) = 8.46, p = .004$). There was also a main effect of sample size ($F(2, 1896) = 57.11, p < .001$): confidence judgments were higher for larger sample sizes. However, these two main effects were qualified by an interaction between condition and sample size ($F(2, 1896) = 36.21, p < .001$): Confidence judgments increased with increasing sample size in the sample condition ($F(2, 912) = 61.32, p < .001$) but not in the population condition ($F(2, 984) = 2.68, p = .69$). This is exactly what had been predicted on the assumption that participants are sensitive to changes in the variance of distributions as a function of sample size.

Further, there was a main effect of proportion ($F(2, 1896) = 13.26, p < .001$): response confidence was higher when the proportion of cases in which a feature held was .9 ($EMM = .877$) than when it was .7 ($EMM = .850, p < .001$) or .5 ($EMM = .855, p < .001$). There was no difference in confidence judgments between the proportions of .7 and .5 ($p = .85$).

Finally, there was a three-way interaction between proportion, condition, and sample size ($F(2, 1896) = 5.13, p = .006$) indicating that the difference in the size of the effect of sample size between conditions was smaller when the proportion was .9.

General discussion

Overall, these findings are fully in accordance with the hypothesis that people are sensitive to differences in the variance of distributions as a function of sample size. They therefore provide evidence against the alternative hypothesis for the findings of Experiment 8 raised in the discussion section of that experiment. The lack of a consistent effect of the width of the coherence interval in Experiment 8 does not seem to be due to a failure to process information on the variance of a distribution. Rather, the evidence is that coherence intervals are not

processed as distributions, but as indicators of boundaries within which responses must lie to satisfy certain rationality constraints. Within those constraints, inductive criteria can affect the distribution of responses, narrowing it down further, depending on the presence of relevant factors from the content or context of the materials.

CHAPTER 8. EXPERIMENT 10: PROBABILITY PRESERVATION PROPERTIES

Contents

8.1 Method

8.2 Results and discussion

8.3 General discussion

The aim of this experiment is to assess the role that deductive validity plays in reasoning, over and above any role of coherence. More specifically, it follows Adams (1996) in conceptualising p-validity as a form of probability preservation. When introducing the theoretical background of the probabilistic approach, it was mentioned that Adams distinguishes four forms of probability preservation, which in order of strictness are certainty preservation (= binary validity), high probability preservation (= p-validity), positive probability preservation, and minimum probability preservation.

As mentioned earlier, the probabilistic account of conditionals advocated in this thesis proposes that the conditional probability in the Equation, $P(\text{if } p \text{ then } q) = P(q|p)$, is primitive within the probabilistic logical system, and is arrived at through the psychological process of the Ramsey test, rather than being derived from unconditional probabilities using the ratio formula, $P(p \ \& \ q)/P(p)$. This treatment of the conditional probability as primitive is in line with the coherence based probabilistic logic proposed by followers of de Finetti (Coletti & Scozzafava, 2002; de Finetti, 1970/1974; Gilio, 2002; Pfeifer & Kleiter, 2009; see also the concept of Popper functions, Adams, 1996, note 1; Adams, 1998, Appendix 2; Popper, 2002). But it differs from Adams' (1996, 1998) approach, who followed Kolmogorov (1933/1950) in defining conditional probabilities through the ratio formula.

The ratio formula has the disadvantage that the conditional probability $P(q|p)$ is undefined when $P(p) = 0$. Adams proposed that the probability of the conditional *if p then q* be set to 1 in this case. Let us call this Adams' stipulation. It implies that $P(q|p) = P(\text{not-}q|p) = 1$ when $P(p) = 0$, which seems contradictory (Adams, 1998, Appendix 2). It also renders it impossible to model counterfactuals. Clearly, counterfactuals are not always certainly true, but can have the full range of probabilities.

A third consequence of Adams' stipulation that is directly relevant to this experiment is that the ordering of the strictness of probability preservation properties depends on it. The coherence based approach has a more principled way of dealing with false-antecedent cases than this stipulation. It proposes that when the probability of the antecedent is 0, the probability of the conditional can be anywhere in the probability range. The value it will take for a person on any particular occasion will be determined by the content of the conditional and the person's associated background knowledge. This conceptualisation of the conditional corresponds to that in the Jeffrey table, and allows the use of the Ramsey test to assess the probability of the conditional. A consequence is that, under this definition of the conditional, certainty preservation coincides with high probability preservation. To see why, consider the inferences discussed in Edgington (1995, p. 286), shown in Table 8.1.

In Adams' probabilistic logic, the inferences in the first column are both certainty and high probability preserving. The inferences in the second column are also certainty preserving, because whenever the premises have probability 1, the conclusion also has probability 1. But

they are not high probability preserving, because whenever the probability of the premises lies below 1, by however small an amount, the probability of the conclusion can be 0.

The reason for why the inferences in the right column are certainty preserving is that the case in which they would have failed to be so is that in which the antecedent of the conditional has probability 0. Since in this case the probability of the conditional is set to 1 by default, the possibility of a failure of probability preservation is precluded.

Table 8.1. P-valid inferences with categorical conclusions and their p-invalid counterparts with conditional conclusions. Taken from Edgington (1995).

Binary valid & p-valid	Binary valid but p-invalid
<i>p, q, therefore p</i>	<i>p, therefore if q then p</i>
<i>p or q, not-p, therefore q</i>	<i>p or q, therefore if not-p then q</i>
<i>not(p & q), p, therefore not-q</i>	<i>not(p & q), therefore if p then not-q</i>
<i>if p then q, if q then r, p, therefore r</i>	<i>if p then q, if q then r, therefore if p then r</i>
<i>if p then q, not-q, therefore not-p</i>	<i>if p then q, therefore if not-q then not-p</i>

In contrast, because in the coherence approach the conditional can take any value when the probability of the antecedent is 0, the case in which certainty preservation would fail for the inferences in the right column is not precluded: these inferences are not certainty preserving in the coherence based approach. Consider for example the first inference: *p, therefore if q then p*. We may judge that *p* certainly holds in the actual world. But the world in which the probability of *q* is zero may be different from the actual world, and *p* may not hold in this different world. Suppose it is certain that I am in the lab and doing research. From this it does not follow that if I were in the café upstairs, then I would be doing research, because although the probability that I am now in the café is 0, if I were to be in the café, it is unlikely I would be doing research.

A consequence of this is that, whereas in Adams' approach high probability preservation is a stricter criterion than certainty preservation (because all inferences that are high probability preserving are also certainty preserving, but not all inferences that are certainty preserving are also high probability preserving), in the coherence based approach certainty preservation and high probability preservation are equivalent. Therefore, the four probability preservation properties of inferences outlined in Adams (1996) collapse to three probability preservation properties in the coherence based approach: certainty preservation/high probability preservation, positive probability preservation, and minimum probability preservation.

The present experiment tested two inferences from each of these three probability preservation categories, together with two invalid but probabilistically informative inferences, and two invalid and probabilistically uninformative inferences. Table 8.2 lists the 10

inferences investigated, ordered according to the increasing degree of strictness of the probability preservation properties that hold for them.

All the inferences in Table 8.2 have two premises. The reason for this is that any one-premise inference that is high probability preserving will also be positive and minimum probability preserving. Therefore any effect of the difference between high, positive, and minimum probability preservation can only be studied with inferences that have more than one premise.

Table 8.2. The 10 inferences investigated, grouped by their probability preservation properties.

Property	#	Name	Form	Coherence interval
(1) Probabilistically uninformative	1	Paradox	$r, q, \text{ therefore if } p$ $\text{ then } (r \ \& \ q)$	[0, 1]
	2	Paradox	$\text{ not-}r, \text{ not-}p,$ $2 + r$ $\text{ therefore if } (r \ \& \ p)$ $\text{ then } q$	[0, 1]
(2) Not probability preserving	3	not-MP	$p, \text{ if } p \text{ then } q,$ $\text{ therefore not-}q$	$[(y - xy), (1 - xy)]$
	4	not-MT	$\text{ not-}q, \text{ if } p \text{ then } q,$ $\text{ therefore } p$	$[0, \min\{y/(1-x), (1-y)/x\}]$
(3) Certainty and high probability preserving	5	And-intro	$p, q, \text{ therefore } p \ \& \ q$	$[\max\{0, P(p) + P(q) - 1\}, \min\{P(p), P(q)\}]$
	6	MT	$\text{ if } p \text{ then } q, \text{ not-}q,$ $\text{ therefore not-}p$	$[\max\{(1-x-y)/(1-x), (x+y-1)/x\}, 1]$
(4) Positive probability preserving	7	MP	$\text{ if } p \text{ then } q, p,$ $\text{ therefore } q$	$[xy, xy + (1-y)]$
	8	or-MP	$\text{ if } p \text{ or } q \text{ (or both)}$ $\text{ then } r, p, \text{ therefore } r$	$[xy, xy + (1-y)]$
(5) Minimum probability preserving	9	Proof by cases	$\text{ if } p \text{ then } q, \text{ if not-}p$ $\text{ then } q, \text{ therefore } q$	$[\min(x, y), \max(x, y)]$
	10	or-intro	$p, q, p \text{ or } q \text{ (or both)}$	$[\max\{P(p), P(q)\}, \min\{P(p) + P(q), 1\}]$

Note. Where x and y are used, x = premise 1, y = premise 2.

Further, the invalid inferences 3 and 4, not-MT and not-MP, are similar to the valid inferences 6 and 7, MT and MP, but they negate the conclusions of inferences 6 and 7. This means that they are not merely probabilistically informative in a way that does not preserve probability, like the inferences AC and DA would be. Inferences 3 and 4 have the additional constraint that the probability of their conclusion must equal 1 – the probability of the

conclusion of their valid counterparts 6 and 7, respectively. Thus, whenever the probability of 6 and 7 is high, the probability of 3 and 4 must be low. Contrast this with inferences 1 and 2: For them the probabilities of the premises pose no constraints at all on the probability of the conclusion, which may then be high or low on inductive grounds depending on background knowledge and the specific content expressed by the inferences.

The additional constraint in inferences 3 and 4 singles them out as a specific subclass of informative but invalid inferences. One could call the subclass of informative but invalid inferences whose conclusion negates a deductive consequence of the premises, or negates one or more of the premises themselves, *contradictory* to convey this additional constraint. Put briefly, this subclass of inferences can be said to have conclusions that are the "opposite" of the premises.

The simplest example of a contradictory inference is arguably p , *therefore not-p*. Here one can see clearly that the probability of the conclusion is informative because it has to be $1 - p$ – the probability of the premise. It cannot be any probability, and not even any probability lower than the probability of the premise. In the special case in which the premise p has probability 1, the conclusion must have probability 0, which is the "opposite" of its valid counterpart, p , *therefore p*, whose conclusion must in this case have probability 1.

For the present experiment this means that the "opposite" of the valid inferences 5 and 6 were the contradictory inferences 3 and 4 rather than the uninformative inferences 1 and 2. Contradictory inferences were chosen to represent the class of informative but invalid inferences in order to establish a larger contrast in the probability preservation properties between valid and invalid inferences. Inferences AC and DA were not used because their status as valid or invalid depends on the subjective correlation between the antecedent and consequent of the conditional in the major premise. As mentioned in previous experiments of the present thesis, this makes it difficult to interpret findings on AC and DA without also controlling for this correlation.

The experiment set out to investigate whether the coherence of people's responses changes as a function of the strictness of the probability preservation properties of an inference, and whether inferences for which stricter probability preservation properties hold are considered of higher quality, specifically "more correct", than inferences for which less strict probability preservation properties hold. Any effect of probability preservation on these two measures would not be explainable through coherence alone. The information on response coherence was of interest in itself, but it also served as a form of control condition to aid in the interpretation of people's judgments of inference quality. If people's judgments of inference quality varied systematically with the strictness of the probability preservation properties of the inferences, but at the same time people's probability judgments were clearly incoherent for these inferences, then there would be reason to believe that the quality judgments were based on information other than the actual coherence constraints. Judgments of inference quality can

only be interpreted as informed by probability preservation properties if the conclusion probability judgments for the corresponding inferences are coherent.

Method

Participants

A total of 140 participants from English speaking countries completed the online experiment on the platform Prolific Academic, in exchange for approximately £5 per hour. All participants indicated at the end of the experiment that they had taken part seriously, as opposed to just clicking through. However, 15 were excluded because they failed to pass one or both of two catch trials asking them not to respond but to instead click next to continue with the experiment. Further 12 were excluded because they had one or more trial reaction times of 3 seconds or less. The final sample consisted of 113 participants. Among them were 7 cases of a repeated IP address, but none of these used the same Prolific Academic ID. All participants indicated having at least "good" English language skills. Participants' mean age was 34 years (range 18 - 64), and most had some college education: 21% indicated that they had finished 12th grade, 11% a technical or applied degree, 53% an undergraduate university degree, and 19% a postgraduate degree (6% indicated not having finished 12th grade, and 3% that they had a doctoral degree). Participants' median rating of the difficulty of the experiment was 66%.

Design and materials

Participants were shown two inferences from each of the five probability preservation categories shown in Table 8.2. Each inference was presented on three consecutive trials. The first two trials displayed the inference with given premise probabilities, and the task was to judge on a percentage scale how likely the conclusion could be given the probabilities of the premises. The response scale had two anchors labelled "0% likely/definitely false" and "100% likely/definitely true" on the left and right ends, respectively. On one of these two trials, both premises had a probability of .9, and on the other both premises had a probability of .75. The order of the two trials was varied randomly for each inference and participant.

On the third trial the inference was shown again without information about premise probabilities. Participants were asked to judge the quality of the inference in general, by indicating on a continuous visual analogue scale to what extent they thought that the inference was correct. This scale had three anchors ordered from left to right: "definitely incorrect", "don't know", and "definitely correct".

Each inference was embedded in a different context story. The context story remained the same across the three trials of the inference, but minor details in the story (e. g. the names of

objects or locations) changed. An example of the contents of the three trials for the *and-elimination* inference is shown in the frame below.

Imagine you are part of a team of aid workers who are removing the mines from the Dunlar fields, where a war took place recently. You have to act very thoroughly in order to make sure the area is cleared and safe again for the residents. You are reviewing the latest data on the fields with the team. The information you have gathered until now suggests the following:

[trial 1:]

Premise 1: It's **90%** likely that:
The oat field is cleared.

Premise 2: It's **90%** likely that:
The barley field is cleared.

Conclusion: Therefore, how likely can the following be?
Both the oat field and the barley field are cleared.

[trial 2:]

Premise 1: It's **75%** likely that:
The moss field is cleared.

Premise 2: It's **75%** likely that:
The gravel field is cleared.

Conclusion: Therefore, how likely can the following be?
Both the moss field and the gravel field are cleared.

[trial 3:]

Premise 1:
The stone field is cleared.

Premise 2:
The rye field is cleared.

Conclusion:
Therefore, both the stone field and the rye field are cleared.
Given the premises, to what extent is it correct to infer the conclusion?

There were 10 different scenarios, which were randomly allocated to the 10 inferences for each participant. The full list of scenarios can be found in Appendix G. The order of occurrence of each inference varied randomly for each participant. With three trials per inference and ten inferences, the experiment had 30 trials, plus two catch trials. The catch

trials were similar in appearance to the experimental trials, but in place of the text of the inferences they stated that they were control trials to make sure participants were paying attention, and asked them not to respond but to instead click *next* to continue with the experiment.

Procedure

Participants received task instructions with an example, together with a request to take their time to think through the questions and answer as carefully as they could. They then worked through two practice trials involving inferences different from those assessed in the experiment. After completing the experimental trials, participants were asked for demographical information and whether they had taken part seriously. The last page provided debriefing information. The experimental session took on average 22.52 minutes to complete.

Results and discussion

Coherence of conclusion probability judgments

The analysis started with the computation of the coherence of participants' responses to each inference, because the interpretation of the judgments of inference quality depended on whether these judgments were made against a background of sensitivity to coherence constraints.

The panels of Figure 8.1 display the distribution of conclusion probability judgments for each inference, giving an overall impression of the proportion of responses falling within the coherence interval in each condition. Note that the y axis ranges from 0 to 10 for all inferences except inference 10 in the condition with a premise probability of .75. In this latter case, the axis ranges from 0 to 80 to accommodate the sharp peak in participants' responses.

The overall pattern in Figure 8.1 shows a clear sensitivity to coherence constraints for all inferences except inference 6, which is MT. Inferences 1 and 2 (in the upper left corner of Figure 8.1) are probabilistically uninformative, so that any response to them is coherent. Responses to these inferences are in accordance with this fact, showing wide distributions with various shallow peaks whose location changes across conditions.

Inferences 3 and 4 (in the upper right corner of Figure 8.1) are informative but invalid. Responses to these inferences show a clear sensitivity to the location of the coherence interval, including a shift in the mean towards higher probability judgments when the premise probability was lower. This shift follows the shift in location of the interval, and shows that participants are not just giving heuristic responses based on a positive correlation with premise probabilities.

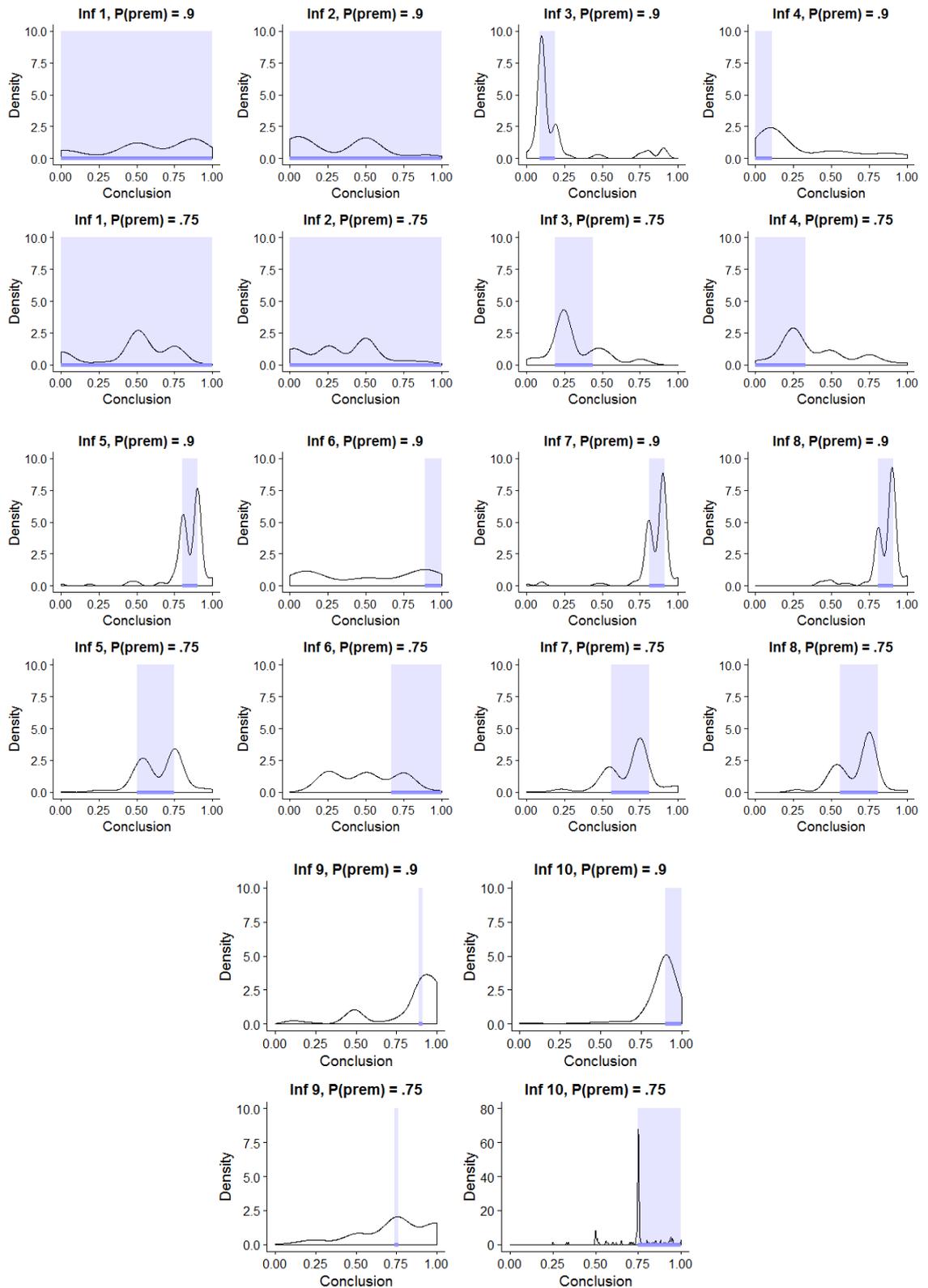


Figure 8.1. Density curves showing the distribution of conclusion probability judgments. The shaded area represents the coherence interval for each inference and premise probability.

Inferences 5 to 10 are valid. With the exception of Inference 6, responses to these inferences also show a notable correspondence with the location of the coherence interval, including a responsiveness to small shifts in interval location that go in the opposite direction

to those for the invalid inferences. Inferences 3, 4, 7, and 8 also show evidence of sensitivity to the width of the coherence intervals, having steeper distributions when the interval was narrower. In contrast, responses to inference 10 seem to be concentrated on the lower interval bound rather than across its range.

The response pattern for inference 6, MT, seems to be the only one showing no apparent correspondence with the coherence interval. As mentioned in earlier, the observation of lower coherence for MT in some conditions and experiments but not others, in the presence of reliable coherence for other inferences across experiments, might reflect a failure of invariance for MT (Oaksford & Chater, 2013). In general, MT can be interpreted as a case of a *reductio ad absurdum* inference. The negation of the consequent stands in conflict with the joint truth of the conditional and its antecedent, so one's degree of belief in one of the two must be lowered. But which of the two is lowered may depend on the content and context of the utterance. Coherence constraints hold only statically, for a single point in time, or equivalently, they hold only under the assumption of invariance. If this assumption is not met, then one cannot speak of an incoherent response, but rather of a change in the set of probabilities for which coherence can be assessed. Nonetheless, for the purposes of the present experiment it was enough to determine whether coherence was above chance level for each inference.

The statistical assessment of coherence followed the same procedure as in previous experiments. The observed rate of coherent responses is shown in the upper panel of Figure 8.2 for the informative inferences. The rate of above-chance coherence for these inferences is shown in the lower panel of Figure 8.2.

Drawing on the betting analogy, the heights of the bars shown in Figure 8.2 for observed coherence tell us how often participants won a bet on each inference, the bet being won (lost) when the probability given for the conclusion was coherent (incoherent). The height of these bars, relative to the upper end of the y axis (at 1), indicates the proportion by which responses were coherent compared to a maximally coherent person.

The heights of the bars for above-chance coherence tell us how much participants earned through their bets. Zero earnings imply that performance was at chance levels. For example, if the chance rate was 50% and a person responded randomly, then observed coherence for that person would be .5, and above-chance coherence would be 0, in the long run. In other words, the person would win about half of the bets, but would not earn anything in the long run. This means that when coherence is above-chance levels there is evidence of "inside knowledge", which in this case is knowledge (at some level) of coherence constraints. But the amount of the earnings depends not only on the amount of insight knowledge, but also on the likelihood of winning a bet by chance. The higher the likelihood of winning by chance, the lower the maximum possible pay-out for a performance that goes beyond chance. Hence higher above-chance coherence does not necessarily amount to higher insight knowledge. The same holds

for observed coherence: A higher frequency of bets that are won does not necessarily indicate higher insight knowledge, because the chance rate of winning could be high as well. That makes it difficult to compare performance between inferences and experimental conditions, unless their chance rates are equated as in Experiments 5 to 7. This was not necessary in the present experiment because the question here was merely whether performance was above chance levels at all, i. e. whether or not participants showed insight knowledge for each inference, regardless of any differences in the amount of insight knowledge between inferences.

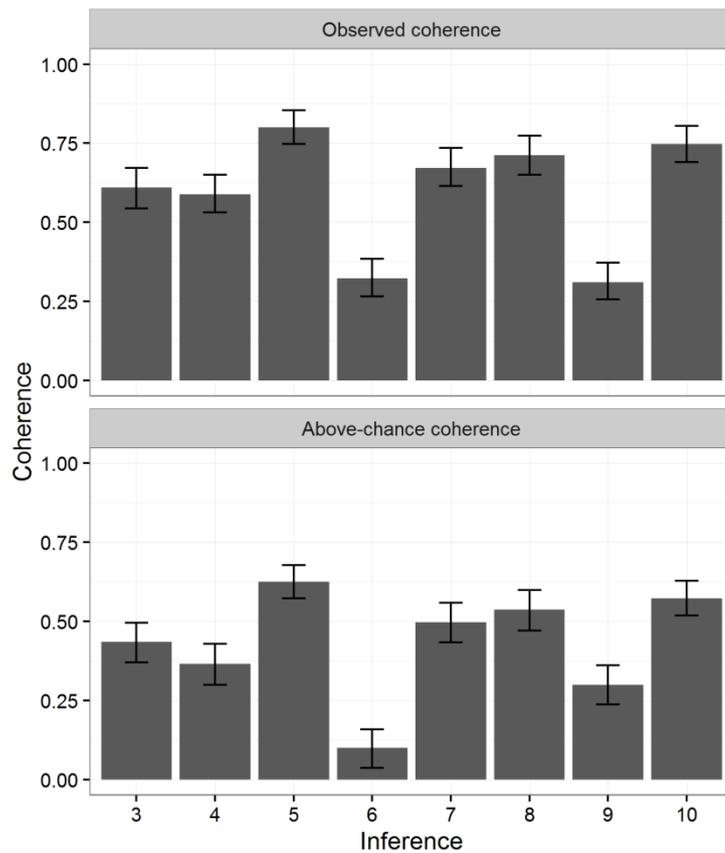


Figure 8.2. Mean values of observed and above-chance coherence for the probabilistically informative inferences. Error bars show 95% CIs.

The confidence intervals in the Figures show that coherence was indeed above-chance levels for each inference, even for inference 6, MT. This was corroborated through a linear mixed model for the effect of inference on ratings of above-chance coherence, with random intercepts for participants and scenarios. The analysis showed that coherence was above chance levels overall ($F(1, 904) = 1027.50, p < .001$), and that it was significant also in the case of the lowest value of above-chance coherence: that for inference 6 ($F(1, 113) = 6.50, p = .012$).

The significant effect for MT suggests that a subset of participants did interpret the inference as intended and provided invariant probability assignments to its premises. Nonetheless, the lower rate of coherence for this inference replicates findings from other Experiments in the present thesis, as well as findings from previous studies (e. g. Evans et al., 2015, Singmann et al., 2014), rendering it worthwhile to investigate it further. In particular, it would be an advance to have an empirical assessment of the extent to which the lower coherence of MT can be explained through a failure of invariance.

Overall, the finding that participants' conclusion probability judgments were sensitive to the coherence constraints of the inferences investigated makes it possible to interpret the ratings of inference quality as ratings informed by knowledge (at some level) of the probability preservation properties of the inferences.

Judgments of inference quality

The mean ratings of the extent to which participants considered the inferences correct are shown in Figure 8.3. The pattern in the figure shows a clear divide between quality judgments for valid and invalid inferences. Judgments for the two probabilistically uninformative inferences, 1 and 2, lie around the lower half of the scale. In contrast, judgments for the two informative, but invalid inferences are in the lower 25% of the scale, whereas judgments for the six valid inferences lie in the upper 25% of the scale. There seem to be only small differences in quality judgments between the six valid inferences, compared to the large distance between the judgments for them and those for the invalid inferences.

The pattern in the figure was corroborated through a linear mixed model analysis for the effects of inference type - the five probability preservation properties in Table 8.1, labelled (1) to (5) - on judgments of inference quality, with a random intercept for participants (inclusion of further random effects led to failure of convergence). Judgments of inference quality differed between inference types ($F(4, 2147.00) = 774.56, p < .001$), with all individual comparisons between types significant (all $ps \leq .001$) except that between the positive and the minimum probability preserving inferences ($p = .16$). The lowest judgments were given to the invalid, informative inferences 3 and 4 ($EMM = .172$), followed by the invalid, uninformative inferences 1 and 2 ($EMM = .457$), followed by the high probability preserving inferences 5 and 6 ($EMM = .828$), followed by the highest quality judgments for the positive probability preserving ($EMM = .933$) and minimum probability preserving ($EMM = .893$) inferences 7 to 10.

It is also apparent from the confidence intervals in Figure 3 that the difference between the invalid, informative inferences 3 and 4 on the one side, and the valid inferences 5 to 10 on the other, was larger than the differences between the different subcategories of valid inferences (high, positive, and minimum probability preserving). These differences in the size of the differences were corroborated in a linear mixed model in which a series of difference

scores for above-chance coherence were computed between inference categories, and used in place of the original scores for above-chance coherence. However, the outcome of this latter analysis is not reported in detail here (see Cumming & Finch, 2005; Masson & Loftus, 2003).

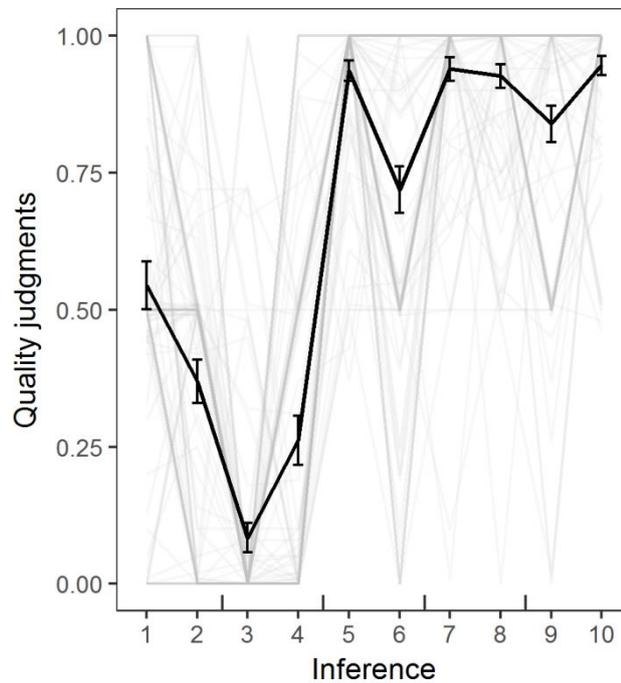


Figure 8.3. Judgments of inference quality for the inferences investigated. The error bars show 95% CIs, and the grey lines in the background show the individual participant values.

General discussion

The results showed that using the same materials, quality ratings for the uninformative inferences 1 and 2 (which may vary on inductive grounds) lay around the lower half of the probability scale. In contrast, ratings for inferences 3 and 4, which minimised probability preservation, were in the lower 25% of the scale, and those for inferences 5 to 10, which preserved high probability, were in the upper 25% of the scale. The specific contrast found between the valid and invalid inferences provides strong evidence for the psychological significance of deductive validity as a feature of inference quality, and singles it out against other measures of probabilistic quality, particularly coherence alone. The philosophical literature, from classical times to recent work on probabilistic logic (Adams, 1996, 1998), has held that the property of deductive validity is of particular importance. But the present findings provide independent empirical justification for this special treatment in an account of the judgments of ordinary people.

The finding that the strongest contrast to the valid inferences 5 to 10 was given by the invalid but informative inferences 3 and 4, rather than by the uninformative inferences 1 and 2, is in accordance with the fact that unlike inferences 3 and 4, whose coherence intervals were constrained to low probabilities, the uninformative inferences were completely unconstrained deductively. This allowed them to vary freely in strength on inductive grounds depending on their specific content.

In general, the study of the inferences in the present thesis suggests a more fine grained classification than the previously used categories of valid vs. invalid, and informative vs. non-informative allow. Among the informative invalid inferences, there are contradictory ones that negate one or more of the premises, or negate deductive consequences of the premises. And there are others whose coherence interval is constrained by the premises to a similar degree as for valid inferences, but not in a way that preserves probability. Examples of this second subset of informative invalid inferences are AC and DA: As illustrated in Experiments 3 and 4, the coherence intervals for these inferences allow for a high conclusion probability in some cases, and a low conclusion probability in others. Their subjective quality is therefore more strongly influenced by inductive criteria, and is more likely to vary from one content to the next. This variability may render them more similar to the uninformative than to the contradictory inferences. Conversely, among the valid inferences some are only high probability preserving, and others are also positive or even minimum probability preserving. It would be worth investigating further the extent to which these more specific categorisations play a role in people's judgments of inference quality.

In summary, the results of this experiment provide a strong and new form of evidence for the empirical relevance of deduction in reasoning, in addition to the more general constraint of probabilistic coherence.

A possible limitation of the experiment is that it did not choose premise probabilities that maximised the difference in the probability intervals between inference types (Adams, 1996). This was difficult to achieve in a comparison of 10 inferences. But in follow-up studies it may be useful to perform pairwise comparisons between inferences with differing probability preservation properties, in a way that maximises the differences in their coherence intervals. Such comparisons would allow for a more fine-tuned assessment of the effects of probability preservation on subjective inference quality.

PART 4. GENERAL DISCUSSION

CHAPTER 9. GENERAL DISCUSSION

Contents

- 9.1 The findings obtained across experiments
 - 9.1.1 Coherent responses to MT
 - 9.1.2 Changing responses to AC and DA
 - 9.1.3 Conditionals, or-introduction, and the conjunction fallacy
 - 9.1.4 Comparing above-chance coherence between inferences
 - 9.1.5 The effect of an explicit inference task and working memory
 - 9.1.6 Certain vs. uncertain premises, probabilistic vs. binary paradigm instructions
 - 9.1.7 Factors with no systematic effect on above-chance coherence
 - 9.1.8 The precision of people's degrees of belief
 - 9.1.9 The variance of belief distributions
 - 9.1.10 P-validity matters over and above coherence
- 9.2 Conclusions
- 9.3 Implications for belief bias and dual-component theories
- 9.4 Limits of deduction and dynamic reasoning
- 9.5 Where next?
 - 9.5.1 Dynamic reasoning
 - 9.5.2 Counterfactuals, generals, and universals
 - 9.5.3 Coherence and rationality

With the advent of the probabilistic paradigm in the psychology of reasoning (Over, 2009), the question of what role, if any, deduction plays in real world reasoning, arose anew (Evans, 2012; Evans & Over, 2013). The present thesis argues that the answer to this question depends in part on how deduction is defined. The definition of deduction from classical logic is binary, and produces what de Finetti called "the logic of certainty" (de Finetti, 1972; Elqayam & Over, 2013). When the premises are assumed to be certain, the conclusion of a classically valid inference must be certain as well. But suppose people have uncertain degrees of belief in the premises of an inference. What degree of belief would be reasonable for them to have in the conclusion? Classical logic cannot answer this question, for it cannot be applied to reasoning under uncertainty. If we retain its binary definition of deduction, the central hypothesis of the probabilistic approach, that most reasoning takes place from uncertain premises, implies that deduction has only a small role to play in most reasoning situations (Oaksford & Chater, 2007).

Using the classical definition of deduction, a number of empirical studies have found an effect of deduction in reasoning tasks, in addition to what has been called an effect of "belief" (Evans et al., 2010; Klauer et al., 2010; Markovits et al., 2015; Rips, 2001; Singmann & Klauer, 2011; Thompson, 1994; Trippas et al., 2017). In particular, people more often accept the conclusion of valid than of invalid inferences, in addition to accepting believable conclusions more often than unbelievable ones. The latter effect has been called *belief bias*, as it constitutes a departure from the normative response under binary logical instructions.

The findings of effects of both "logic" and "belief" in reasoning have contributed to the development of dual-component theories (Evans, 2006; Klauer et al., 2010; Markovits et al., 2015; Verschueren et al., 2005). When these theories assume a binary logical process that is contrasted with a "belief based" process, the "logical" process questions the basic hypothesis in the probabilistic approach that most real world reasoning is from uncertain premises.

The present thesis argues for an alternative way of accounting for the findings on an effect of deduction that does not question the basis of the probabilistic approach. Rather than integrating a binary logical component on the one side, and a probabilistic, belief based component on the other, into a dual-component framework, it explores the integration of logic and probability itself.

The central deductive concepts of classical logic, consistency and validity, can be generalised to cover degrees of belief: binary consistency can be generalised to coherence, and binary validity to probabilistic validity, p-validity (Adams, 1998; Coletti & Scozzafava, 2002; Gilio, 2002). This generalisation makes it possible to study deduction from uncertain premises (Stevenson & Over, 1995), so that there is no need to qualify the basis of the probabilistic approach.

With this generalised notion of deduction, many findings on belief bias can be reinterpreted as effects of coherence. Participants in an experiment may be violating the

classical deductive *instructions* to assume the premises to be certain, and nonetheless be reasoning deductively, by estimating a coherent degree of belief in the conclusion of an inference on the basis of its logical form and of their uncertain degrees of belief in the premises. Not to follow binary instructions can be said to be a fault, but it is not necessarily a logical one.

However, the fact that the definition of deduction can be generalised to cover degrees of belief does not imply that people actually use deduction in a probabilistic way. It could still be that when people engage in deduction, they do so in the classical binary way, and that reasoning from uncertain premises is inductive in practice even when it does not have to be so in theory. Whether people are sensitive to the constraints of coherence and of p-validity is an empirical question that has only recently started to be investigated, but which was the main focus of this thesis.

The thesis investigated the role of coherence and p-validity in uncertain reasoning through ten experiments. Previous studies on the coherence of people's conclusion probability judgements had focussed on conditional syllogisms (Evans et al., 2015; Singmann et al., 2014). Responses to these two-premise inferences were found to be coherent above chance levels mainly for MP, and less reliably for DA. However, the chance rates in these studies were often very high (e.g. Singmann et al., 2014, Figure 3) and varied between inferences, making it more difficult to detect above-chance coherence when it is there, and to compare this coherence between inferences.

The findings obtained across experiments

The present experiments extended these findings, using different methodologies, across 23 inference forms, summarised in Table 9.1. Some of these forms were investigated in different variants (e. g. inference 5), which were not counted as separate inferences. The table presupposes, in its validity column, that *if* refers to the probability conditional (Adams, 1998; Jeffrey, 1991).

Coherent responses to MT

For the conditional syllogisms, the results corroborated the earlier finding of above-chance coherence for MP, but found coherence to be reliably above chance levels for MT too. Of the six experiments in which both inferences were investigated, overall coherence was above chance in all cases for MP. For MT, coherence was above chance levels in all but two instances. The first was in Experiment 3, where it was at chance levels in the statements task but above chance levels in the inferences task. In Experiment 4, which replicated lab Experiment 3 on the internet, coherence was above chance levels across conditions. The

second instance was in Experiment 7: coherence for MT was at chance levels when participants were given binary paradigm instructions to assume the premises to be true, and then judge whether the conclusion also has to be true. This was in spite of the fact that coherence for MT was above chance levels for the same materials when given probabilistic instructions (Experiment 6).

Table 9.1. The inferences investigated in the 10 experiments of the thesis.

Type	#	Name	Form	Valid	Exp
A. One-premise equivalences and contradictions	1	de morgan	$\text{not}(p \ \& \ q) \ \therefore \ \text{not-}p \ \text{or} \ \text{not-}q$	1	3-8
	2	not de morgan	$p \ \& \ q \ \therefore \ \text{not-}p \ \text{or} \ \text{not-}q$	0	3-8
B. One-premise inferences with set-subset relations between premise and conclusion	3	if-to-or	$\text{if } p \ \text{then } q \ \therefore \ \text{not-}p \ \text{or } q$	1	1
			$\text{if not-}p \ \text{then } q \ \therefore \ p \ \text{or } q$	1	1,3-7
	4	or-to-if	$p \ \text{or } q \ \therefore \ \text{if not-}p \ \text{then } q$	0	1,3-7
			$\text{not-}p \ \text{or } q \ \therefore \ \text{if } p \ \text{then } q$	0	1
	5	one-premise or-introduction	$p \ \therefore \ p \ \text{or } q$	1	1
			$\text{not-}p \ \therefore \ \text{not-}p \ \text{or } q$	1	1
			$q \ \therefore \ p \ \text{or } q$	1	1
			$q \ \therefore \ \text{not-}p \ \text{or } q$	1	1
	6	and-elimination	$p \ \& \ q \ \therefore \ p$	1	2-7
			$p \ \& \ q \ \therefore \ q$	1	2
7	one-premise and-introduction	$p \ \therefore \ p \ \& \ q$	0	3-7	
8	and-to-or	$p \ \& \ q \ \therefore \ p \ \text{or } q$	1	3-7	
9	or-to-and	$p \ \text{or } q \ \therefore \ p \ \& \ q$	0	3-7	
10	and-to-if/ one-premise centering	$p \ \& \ q \ \therefore \ \text{if } p \ \text{then } q$	1	2	
C. Two-premise inferences	11	and-to-if/ two-premise centering	$p, q \ \therefore \ \text{if } p \ \text{then } q$	1	2,8
	12	MP	$\text{if } p \ \text{then } q, p \ \therefore \ q$	1	3- 8,10
	13	MT	$\text{if } p \ \text{then } q, \text{not-}q \ \therefore \ \text{not-}p$	1	3- 7,10
	14	AC	$\text{if } p \ \text{then } q, q \ \therefore \ p$	0	3-7

	15	DA	$if\ p\ then\ q,\ not-p\ \therefore\ not-q$	0	3-8
	16	two-premise and-introduction	$p,\ q\ \therefore\ p\ \&\ q$	1	10
	17	or-MP	$if\ p\ or\ q\ (or\ both)\ then\ r,\ p\ \therefore\ r$	1	10
	18	two-premise or-introduction	$p,\ q\ \therefore\ p\ or\ q\ (or\ both)$	1	10
	19	Proof by cases	$if\ p\ then\ q,\ if\ not-p\ then\ q\ \therefore\ q$	1	10
Two-premise uninformative	20	Paradox 1 + r	$r,\ q\ \therefore\ if\ p\ then\ (r\ \&\ q)$	0	10
	21	Paradox 2 + r	$not-r,\ not-p\ \therefore\ if\ (r\ \&\ p)\ then\ q$	0	10
Two-premise contradictory	22	not-MP	$p,\ if\ p\ then\ q\ \therefore\ not-q$	0	10
	23	not-MT	$not-q,\ if\ p\ then\ q\ \therefore\ p$	0	10

Note. "Valid" = Validity, "Exp" = Experiment.

These findings constitute strong evidence for a basic sensitivity to coherence constraints for MT. They also provide direct evidence for the descriptive adequacy of the proposal that the constraints of deduction are not limited to the special case in which premises are assumed to be certain.

The methodology of Experiments 5 to 7 made it possible, not only to assess whether or not conclusion probability judgments are coherent above chance levels, but also to make quantitative comparisons of above-chance coherence between inferences. This was accomplished by a design that equated the chance rate of coherence across inferences and conditions. The experiments revealed that although responses to MT were generally coherent above chance levels, coherence was lower for MT than for MP, in line with the literature using binary paradigm instructions. In the probabilistic approach, the difference in the acceptance of the two inferences can be explained as a result of dynamic reasoning (Oaksford & Chater, 2013). Specifically, MT can sometimes be viewed as an instance of a *reductio ad absurdum* inference. The categorical premise *not-q* negates an element of the conditional premise *if p then q*, with the result that a high degree of belief in both premises is incompatible with a high degree of belief in the element *p* of the conclusion. But the conflict between *if p then q*, *not-q*, and *p* can be resolved in different ways. Depending on background beliefs, one person could see *if p then q* and *not-q* as a reason to disbelieve *p*, in line with MT, while another person could see *not-q* and *p* as a reason to disbelieve *if p then q*, engaging in dynamic reasoning. Logic does not itself tell us which statement should give way, so that without instructions to assume that both premises are certain, the choice can be made on inductive grounds. This

perspective makes it possible to account for the asymmetry between MP and MT in a rational way.

Changing responses to AC and DA

The coherence findings for AC and DA were more equivocal. In Experiment 3, coherence for both inferences was above chance levels in the statements task, but not in the inferences task. In Experiment 4 coherence was above chance levels for both inferences and both task conditions. However, coherence for both inferences was below chance in Experiment 5, and at chance in Experiments 6 and 7. In Experiment 8, coherence was above chance for DA (this experiment did not include AC). Leaving aside the statements tasks of Experiments 3 and 4, this means that coherence for AC was above chance in one out of five experiments, and coherence for DA was above chance in two out of six experiments. As mentioned before, a lack of above-chance coherence for these two inferences is difficult to interpret, because it can be explained away by a biconditional interpretation of the conditional premise – more precisely, by the assumption that the materials used sometimes suggested to participants that the antecedent and consequent of the conditional were positively correlated. In line with this, an analysis of the data in Experiment 3 showed that coherence was above chance levels under the assumption of a biconditional interpretation for the same responses that coherence was at chance levels under the assumption of a conditional interpretation. To be able to interpret the coherence results for these inferences, the two interpretations would therefore have to be disentangled, by controlling explicitly for the correlation between antecedent and consequent. Although such an experiment was beyond the scope of this thesis, the fact that it is needed to interpret the findings on AC and DA is nonetheless a more precise, and less pessimistic, standpoint than the suggestion from previous studies that responses are generally incoherent for these inferences (e.g. Singmann et al., 2014).

Further studies of AC and DA, but also of MP, MT, and further two-premise inferences, could assess to what extent incoherent responses are responses that overestimate or underestimate the probability of the conclusion, given the probabilities assigned to the premises, and to what extent this depends on the risks and benefits suggested by the materials (Oberauer & Wilhelm, 2003). The hypothesis that there can be overconfidence in the conclusion of a valid inference could not even be formulated in the binary paradigm. But this hypothesis can extend the study of suppression effects, making it necessary to distinguish between the suppression of the conclusion of an inference, and the suppression of the inference itself (Over & Cruz, 2018).

Conditionals, or-introduction, and the conjunction fallacy

Experiments 1 and 2 not only extended the coherence results to further inferences, but also provided information about the meanings of the component statements, and of factors involved

in reasoning with them. An analysis of the or-to-if inference showed that people's responses were coherent above chance levels under the assumption that they interpret the conditional in the conclusion as the probability conditional, whereas they were coherent below chance levels under the assumption that they interpret the conditional as the material conditional, providing a novel form of evidence for the interpretation of conditionals in terms of the Equation.

Coherence was reliably above chance levels across four variants of the inference of or-introduction, suggesting that although it may be pragmatically infelicitous, under binary paradigm instructions, to state *p or q* when one could be more informative and precise by stating *p*, people readily treat the inference as valid when asked directly about their degrees of belief in premise and conclusion. This is in accordance with the view that the lower endorsement rates found for the inference under binary paradigm instructions are due to pragmatic effects (Cruz et al., 2017; Orenes & Johnson-Laird, 2012), contrary to the recent proposal in a revision of mental model theory that the inference is in fact invalid (Johnson-Laird et al., 2015).

Further, Experiment 2 showed that although responses were reliably coherent for and-elimination when using neutral materials (see also Politzer & Baratgin, 2016, for converging evidence), coherence for the same inference broke down when using the materials known to cause the conjunction fallacy (Tversky & Kahneman, 1983). This was in spite of the arguably transparent task of inferring the probability of *p* from the probability of *p & q*. The finding underlines the strength, and at the same time the limited scope, of the fallacy.

Comparing above-chance coherence between inferences

It was pointed out in the thesis that it is difficult to make quantitative comparisons of above-chance coherence between inferences when the chance rates of the inferences are not equal. The reason is that the chance rate is subtracted from the observed coherence rate to obtain the measure of above-chance coherence. The larger the chance rate, the lower the probability of detecting above-chance coherence when it is there; that is, the lower the sensitivity of the test for above-chance coherence. The chance rate corresponds to the width of the coherence interval, which in turn depends on the form of the inference, on whether the inference is valid or invalid, and on the premise probabilities. Without holding the chance rate constant, it is possible to compare inferences with regard to whether or not coherence was above chance levels for them (Cruz et al., 2015; Evans et al., 2015; Singmann et al., 2014), but further comparisons seem difficult to interpret.

In Experiments 5 to 7, above-chance coherence was made comparable between inferences and conditions using two methods. In the first, participants were given a set of premise probabilities, each of which was presented with a number of conclusion probabilities. The task was to give a binary response as to whether a given conclusion probability was possible or not, given the premise probabilities. In the second method, participants were again given a set of

premise probabilities, and were asked whether the probability of the conclusion could be higher, and whether it could be lower, than the probability of the premise (in the case of one-premise inferences) or than 50% (in the case of two-premise inferences). These binary response formats rendered the chance rate of a coherent response 50% across inferences and conditions.

Using these methods, it was found that the inference for which above-chance coherence was highest was the contradiction *not de morgan*, followed closely by MP. Except for AC and DA discussed above, coherence for the remaining inferences was lower but still generally above chance. Two oddballs were the inferences from or-to-if and from if-to-or (inferences 3 and 4 in Table 9.1). Coherence was at chance levels for or-to-if in Experiment 5, for if-to-or in Experiment 6, and for both in Experiment 7 (using binary paradigm instructions). Leaving aside the statements task of Experiments 3 and 4, and considering only the inferences task in those experiments, this means that coherence for or-to-if and if-to-or was above chance in four out of six experiments. These inferences contain negations at the start of the premise or of the conclusion, and this could have made them more difficult to process. Responses to both inferences were above chance when measured across positions of the negation in Experiment 1, but further studies would be necessary to establish why coherence was somewhat less reliable for these inferences.

A further step would also be to investigate in more detail what makes the inferences of not de morgan and MP stand out in terms of the high rates of coherent responses to them. For example, one could test whether contradictions in general are detected more easily than other logical relations, or whether it is something specific to the negation of de morgan that is at play.

Generally, the finding that across the inferences investigated, coherence was above chance levels in the great majority of cases, and failures of coherence were the exception rather than the rule, provides strong additional support for the hypothesis that people are sensitive, at some level, to the constraints of coherence over and above the binary constraints of consistency.

The effect of an explicit inference task and working memory

Experiments 3 and 4 extended the results of Evans et al. (2015) on the role of an explicit inference task for coherence. Somewhat surprisingly, coherence was already above chance levels in the large majority of cases in the statements task: when people were given the statements that made up the inferences in random order one at a time on the screen. An explicit inference task, in which the statements were arranged into inferences, and each inference presented one at a time on the screen, tended to increase coherence in the few cases in which it was not already above chance. However, there was an exception in Experiment 3 for AC and DA, where coherence was above chance in the statements task but at chance levels

in the inferences task. It may be that in some cases, pragmatic factors related to the assertability of a statement as a conclusion drawn from other statements, or to the relation between statements that are difficult to integrate due to the presence of negations, can lead to reductions in coherence that would not arise when the statements are considered in isolation. Another possibility is that putting the conditionals in an explicit AC or DA inference task tends to increase the biconditional interpretation, since otherwise the inferences are invalid. But the finding could also simply reflect the fact that different groups were in the statements and inferences task, and some may have interpreted the materials for AC and DA as implying a correlation between antecedent and consequent, and others not. Further replications would be needed to establish the reliability of this finding.

Experiments 3 and 4 also included an inferences task with working memory load. Responses in this condition generally differed little from those in the inferences task without working memory load. But where they differed, the load condition was associated with lower coherence rates, suggesting that the difference between the statements and the inferences task may be due in part to the differing demands they pose on working memory for calibrating beliefs across statements. However, the weak effect of the load condition is in line with the finding that coherence was in most cases already above chance in the statements task, so that the inferences task had only little to add that could be disrupted.

The overall pattern suggests that people may have an implicit, spontaneous tendency to establish coherence between beliefs, but that in situations in which this fails, an explicit inference task, in which all relevant pieces of information are available simultaneously on the screen, and people can focus their attention directly on the relations between them, tends to be helpful. In any case, explicit inference is necessary when establishing relations between novel materials for which no beliefs are yet available.

Certain vs. uncertain premises, probabilistic vs. binary paradigm instructions

Experiments 6 and 7 made it possible to compare above-chance coherence for inferences with certain vs. uncertain premises, and for inferences with probabilistic vs. binary paradigm instructions, using the same inferences, materials, and response format. There was no evidence that coherence is lower when the premises are uncertain, nor that coherence is lower when probabilistic rather than binary paradigm instructions are used. This provides a novel, strong form of evidence that deduction from uncertain premises is possible, and is not restricted to reasoning from certainty. Certain truth and certain falsity did not appear qualitatively different from uncertain degrees of belief, but rather as endpoints on a common scale.

Factors with no systematic effect on above-chance coherence

In addition to the finding that coherence did not differ between certain and uncertain premises, nor between probabilistic and binary paradigm instructions, Experiments 3 and 4 found no

evidence of a systematic difference in people's responses to the reasoning tasks studied in an internet and in a lab setting, making it easier to generalise results between them, as well as between the experiments conducted in this thesis and the earlier lab results from Evans et al. (2015). In addition, Experiment 5 found no evidence for a difference in response coherence as a function of whether people were asked to judge whether a conclusion fell inside or outside the coherence interval. Across experiments, there also seemed to be no systematic difference in response coherence between one- and two-premise inferences. The differences in coherence between inferences rather appeared to be based on more specific factors, such as whether they contained negations or could be interpreted in alternative ways. Finally, across experiments there was no evidence that coherence differed between valid and invalid, i.e. between deductive and inductive, inferences. This result makes sense given that the constraints of coherence hold for both inference types, and deductive inferences merely have stronger constraints on the lower limits of their interval boundaries. The above negative results can help interpret and add precision to the positive findings observed in these experiments.

The precision of people's degrees of belief

Coherence intervals are usually measured using point probabilities, but there was evidence that people's degrees of belief are not that fine grained. Experiment 3 measured above-chance coherence using the exact point intervals, and compared this with above-chance coherence in which the interval boundaries were widened by 5% and by 10%, widening the chance rate of coherence accordingly. This made the measurement scale coarser, without making it necessarily more lenient. Above-chance coherence increased when widening the scale by 5%, i. e. when the number of points on the scale was reduced from 101 to 10, mainly for the equivalence of *de morgan* and the contradiction of *not de morgan*, for which the conclusion coherence interval is a point value. It had only little effect on the other inferences whose coherence intervals were already wider from the beginning. Increasing the coarseness by 10% had no incremental effect. In Experiment 5, the question of the precision of people's degrees of belief was assessed in a different way, comparing response coherence for conclusion probabilities that were clearly inside or outside the interval, with conclusion probabilities that were at the interval edge. Above-chance coherence was higher for conclusion probabilities clearly on one side of the interval, and this effect was not restricted to *de morgan* and *not de morgan* but held more generally across inferences.

It seems to make sense for degrees of belief to be generally coarser than point probabilities, given the uncertain nature of much of the information we receive in everyday situations, and the limits of our working memory for past instances of an event (c.f. Sanborn & Chater, 2016). The present thesis proposed two methods of quantifying this precision, or fuzziness, in people's beliefs. This precision will likely vary across content domains and domain expertise. But the ability to measure it for a given context, using the tools of

probability theory, can be useful for interpreting experimental findings, and seems to disable one of the arguments brought forward by advocates of computational level systems that are themselves coarser than probability theory, like ranking theory or the use of verbal, qualitative probability expressions (Politzer & Baratgin, 2016; Spohn, 2013). Such alternative systems have a built-in, fixed degree of coarseness that is decided a priori, the use of which makes it impossible to measure the actual coarseness of degrees of belief empirically.

The variance of belief distributions

In addition to assessing people's sensitivity to the location of coherence intervals, Experiments 3, 4, 8, and 9 examined people's intuitions about interval width. Experiments 3 and 4 included an assessment of whether the variance of responses was larger when the coherence interval was wide than when it was narrow, using premise probability information to estimate interval width. The hypothesis was that response variance would be higher when the interval was wider, but no relation was found between the two. Experiment 8 assessed whether people's confidence in the correctness of their conclusion probability judgments (Thompson & Johnson, 2014) varied as a function of interval width. If confidence was lower for wider intervals, this might suggest that people are looking for a single optimal response within a distribution, e.g. corresponding to the distribution mean, which is more difficult to find when there are many options. If confidence was higher for wider intervals, this might suggest that people are focussing on the task of rendering their responses coherent, which is easier when the number of coherent response options is larger. But again no relation was found between the two.

Experiment 9 helped interpret the results of Experiment 8, by suggesting that the absence of a relation between response confidence and interval width was not due to a lack of sensitivity for parameters determining distribution variance. Instead, it seems as if people, in the first instance, follow the deductive constraint of coherence, trying to give responses that fall within the interval; but that if the interval is wide enough, then inductive considerations may or may not narrow down the choice of response further. This interpretation was also suggested by an inspection of the distribution of responses for each inference. When the interval was narrow, the distribution of responses was also narrow and seemed to follow the location of the interval closely. When the interval was wide, the distribution of responses was flat in some cases, suggesting that people were mainly trying to be coherent, without narrowing down their responses further in any specific way. But in other cases the distribution of responses was strongly skewed towards one interval edge, or even multimodal, suggesting that additional inductive criteria were playing a strong role in narrowing down people's responses further in various ways. The response distributions computed in Experiment 10 led to similar impressions. Generally, these findings shed further light on the complementary roles of deduction and induction in reasoning from uncertain premises.

P-validity matters over and above coherence

It can be difficult to assess the role of p-validity over and above the role of coherence in reasoning, because the relevant normative constraints are based on coherence in both cases. In this thesis it was proposed to describe p-validity, i.e. probability preservation, as a feature of coherence intervals. P-validity can be used to categorise inferences into two groups (deductive and inductive) according on whether or not their coherence intervals preserve probability from premises to conclusion. With this characterisation, the question is not whether people respect the normative constraints of p-validity in their conclusion probability judgments, because these normative constraints are set by coherence. The question is rather to what extent the distinction marked by p-validity between the two groups of inferences matters to people.

Across experiments, there was no evidence that people distinguish between p-valid (deductive) and p-invalid (inductive) inferences in terms of the effort they invest in drawing them, because above-chance coherence did not differ systematically between p-valid and p-invalid inferences. But Experiment 10 showed that people did distinguish between deductive and inductive inferences in their judgments of inference quality. Deductive inferences that preserved probability were judged more correct than inductive inferences that did not. Further, p-validity was treated as special among the different levels of probability preservation studied, with forms of probability preservation that were stricter than p-validity having only a negligible further impact on quality judgments. This corroborated empirically the special treatment long given to the distinction between deduction and induction in the philosophical literature.

Experiment 10 also drew a distinction, for the inductive inferences, between the following cases. Inferences whose coherence interval is the uninformative unit interval (like the paradoxes of the material conditional); inferences with a coherence interval that is not high probability preserving but is constrained in a different way by the premises (such as AC); and inferences with a conclusion that is the negation of the conclusion of a valid inference, so that the conclusion is impossible when the premises are certain, and the conclusion is very improbable when the premises are very probable. It would be interesting to investigate further to what extent these more fine-grained distinctions play a role in people's evaluations of inference quality.

It would also be worth developing further ways of assessing to what extent, and in which contexts, people treat deductive and inductive inferences differently (c.f. Trippas et al., 2016). In general one can expect the difference to matter in some contexts, but not in others. Probability preservation adds reliability to the conclusion probability of an inference across individual instances. This reliability may be important in situations when, as in some of the experimental materials, much is at stake and careful consideration is called for to avoid jumping to conclusions. But in other contexts it may be more helpful to respond quickly, without hesitating to jump to conclusions, e.g. because only an approximate answer is needed

or possible given the available information, and the reasoner must move on to address the next task. If we relied only on deduction in everyday reasoning, even if it is probabilistic, we might regularly freeze in the absence of sufficient criteria for drawing any conclusion. Moreover, as discussed in relation to Experiments 8 and 9, deduction and induction often seem to work hand in hand. Thus, instead of asking in which contexts deduction is relevant, it may be more useful to ask how the different contributions of deduction and induction can be measured in reasoning contexts in which they both play a role.

Conclusions

The binary deductive concepts of classical logical logic, consistency and validity, can be generalised to cover degrees of belief: consistency can be generalised to coherence, and validity to p-validity. But the fact that this generalisation is possible in formal logic does not imply that people will actually use deduction in a probabilistic way. The research presented in this thesis investigated the role of deduction in reasoning from uncertain premises through ten experiments. It found evidence that coherence and p-validity are not just abstract formalisms, but that people follow the normative constraints set by them in their reasoning. This is evidence for the descriptive adequacy of coherence and p-validity as computational level principles modelling the tasks people set out to accomplish when reasoning. It has implications for the interpretation of past findings in the literature on the roles of deduction and degrees of belief, and it offers a perspective for generating new research hypotheses in the interface between deductive and inductive reasoning.

Implications for belief bias and dual-component theories

The evidence found in the thesis for coherence and p-validity, and with it for deduction from uncertain premises, opens the possibility that belief bias and suppression effects can, in some cases at least, be explained as resulting from rational reasoning processes. To be sure, it is incorrect to reject the conclusion that all cats are bats, when given instructions to assume it certain that all cats are rats, and that all rats are bats. But this does not imply that the rejection of the conclusion results from a non-deductive process. It could be that it instead occurs because people are used to taking into account the probability of the premises when reasoning, and try to give coherent responses given both their degrees of belief in the premises, and the logical structure of the inference. For example, consider the MP inference *if this cat is a rat, then it is a bat. This cat is a rat. Therefore, this cat is a bat.* Given the low probability of the premises, it is coherent to assign to the conclusion a low probability even though the inference

is probability preserving and so deductively valid. When $P(\text{if rat then bat}) = P(\text{rat}) = .1$, the maximum coherent probability one can assign to the conclusion of this MP inference is .11. There is then a sense in which one would be suppressing the inference if one assigned to it a conclusion probability that is too high to be coherent given its logical structure, a situation that could not even be expressed in the binary approach (Over & Cruz, 2018). It may be that belief bias and suppression effects are still observed in a fully probabilistic setting. If so, then this would provide stronger evidence that people sometimes judge the probability of a conclusion taken on its own, regardless of its relation to the premises. But this is still an open question at present.

The extension of deduction to probabilities also renders unnecessary the distinction between binary logical processes on the one side, and probabilistic belief based processes on the other, opening new avenues of research in the context of dual-component theories (De Neys, 2012; Evans & Stanovich, 2013; Oaksford & Chater, 2011; Singmann et al., 2014; Trippas et al., 2017). It appears that the idea of probabilistic deduction can be integrated seamlessly in the dual-process account of Oaksford & Chater (2011), which is already fully probabilistic, as well as in the account of De Neys (2012; see also Trippas et al., 2017), which explicitly states that the term "logical intuitions" refers to intuitions about both binary logical and probabilistic relations. In the dual-source model of Klauer and colleagues (Klauer et al., 2010; Singmann et al., 2014), the two sources of information proposed, logical form and content, can be easily mapped to the logical form of inferences and the probabilities of their premises, respectively, information on both of which is necessary to compute coherence intervals for the conclusions of inferences. This reinterpretation would be in line with the finding in Singmann et al. (2014) that a model based on coherence alone accounted equally well for the data than one based on coherence plus the pattern of results obtained for conditional syllogisms when using binary paradigm instructions.

The dual-process theory of Evans and colleagues (Evans, 2006, 2007; Evans & Stanovich, 2013) has sometimes suggested a distinction between rule based processes of type 2, including binary logic, and probabilistic, heuristic based processes of type 1. But if as more recently suggested, the only feature distinguishing the two types of processes is the involvement of working memory (Evans & Stanovich, 2013), then the distinction between the attribute of being logical on the one side and that of being probabilistic on the other again becomes unnecessary. The theory then seems to diverge from that of De Neys and colleagues mainly in the algorithmic level question of how the two types of processes interact.

The accounts of Verschueren et al. (2005) and of Markovits et al. (Markovits et al., 2013) draw a distinction between inferences based on the probability of the conclusion, given the premises, and inferences based on the number of different types of counterexamples to the conclusion, given the premises. They attribute the use of conclusion probability information to type 1 processes, and the use of counterexample information to type 2 processes, and find an

effect of both on reasoning (see also Geiger & Oberauer, 2007, on the relation between the two kinds of information). The distinction between conclusion probability and counterexample information seems highly relevant for our understanding of how people arrive at a degree of belief about the premises and the conclusion of inferences. However, this appears to be an algorithmic level question. At the computational level, both conclusion probability and counterexample information can be used in a binary or probabilistic way, depending on instructions. The case made in this thesis is that when participants use it in a probabilistic way, their reasoning can still be deductive if it respects the deductive constraints of coherence (Over & Cruz, 2018).

It may be that at the algorithmic level, people use different mechanisms to judge certainty preservation and probability preservation, and that this is in part driving effects of instruction. It may also be that the way people judge whether an inference is deductive or inductive is different from the way people judge the probability of a conclusion when it is considered on a case-by-case basis, regardless of its logical status. And it could be that all these mental operations occur both in an intuitive and in a reflective way, or that some are more often carried out intuitively and others reflectively. Overall, the generalisation of deduction to inferences from uncertain premises allows the formulation of more specific hypotheses about the components involved in dual-component theories.

Limits of deduction and dynamic reasoning

Even when people are sensitive to the constraints of coherence, and to the distinction between deductive inferences that preserve probability and inductive inferences that do not, the role of deduction in everyday reasoning may still be limited. An inherent limitation on the scope of deduction is that its constraints are static, restricted to a specific point in time, unless further assumptions are made. Given the probability of the premises, the probability of the conclusion must lie within a certain interval on the probability range to be coherent. But if the probabilities of the premises dynamically change over time, then that interval changes with them (Baratgin & Politzer, 2010; Douven, 2012; Hadjichristidis et al., 2014; Hartmann & Rafiee-Rad, 2014; Oaksford & Chater, 2013; Zhao & Osherson, 2010). If the probabilities of the premises do not change over time, then they are said to be *invariant* (Oaksford & Chater, 2013), and the coherence interval computed for an inference remains in force. But the problem is that it is not usually p-valid to infer that invariance holds. It often takes an inductive inference to establish invariance. For short intervals of time, it can be inductively reasonable to infer that there are minimal changes in the circumstances relevant to the probabilities of the premises we are using in reasoning (Hartmann & Rafiee-Rad, 2014; Lewis, 1976; Stalnaker, 1968). For example, when students in London today go out of their flats to use the

Underground, they can safely infer that their flats will still be there on their return. But this was an unreliable inference when the Underground was used as a bomb shelter during the London Blitz (see also Wittgenstein, 1991, on questioning information previously considered certain). The criteria people use for the dynamic revision of their beliefs during reasoning, and the place of deduction in this reasoning, is a remaining challenge for the new paradigm.

Where next?

Dynamic reasoning

Moving forward from the research presented here, it will be important to investigate further how deduction and induction work together in the reasoning process. In a broader sense, this would also mean studying different types of inductive reasoning, and assessing what these types can tell us about reasoning as a whole.

Following up from the previous section, one aspect of reasoning that would be worth studying further is dynamic reasoning. This could involve questions like the following. Under which conditions do participants' judgments conform to Jeffrey conditionalisation (Hadjichristidis et al., 2014; Stern, 2017; Zhao & Osherson, 2010; 2014) and its generalisation in the Kulback-Leibler divergence (Hartmann & Rafiee-Rad, 2014)? Are there some contexts in which people's belief revision seems to conform more closely to imaging than to conditionalisation (Bartgin & Politzer, 2010; Lewis, 1976; Zhao & Osherson, 2014)? Are there situations in which people do not seem to minimize the change in their beliefs upon receiving new information (Hartmann & Rafiee-Rad, 2014; Lewis, 1976; Stalnaker, 1968), but revise more than this minimum? A related question is whether updating more than the minimum is invariably a signal of prejudice and delusions (Dudley & Over, 2003), or whether there are situations in which it is normatively justified.

To address questions of dynamic reasoning, it may be useful to complement the method of coherence computation (Capotorti et al., 2003; Coletti & Scozzafava, 2002) with the method of constructing and manipulating Bayesian networks (Pearl, 1988; Sloman, 2005). The computation of coherence has the advantage that it is straightforward to carry out in situations of incomplete information. Conversely, the use of Bayesian networks has the advantage that it is straightforward to assess changes in probabilities (all else being equal) after receiving a piece of new information, including situations in which the coherence interval before and after receiving the information is the unit interval (McGee, 1985; Stern, 2017, June).

Counterfactuals, generals, and universals

One way in which people's beliefs change over time as they receive new information is when a conditional, e.g. "If it rains today, the road will be muddy" becomes void as an indicative in the sense that there is no actual state of affairs that could confirm or refute it, and is replaced with a counterfactual, "If it had rained today, the road would have been muddy". The work in this thesis focussed on singular indicative conditionals, which were modelled by the Equation, $P(\text{if } p \text{ then } q) = P(q|p)$. But it would be important to assess to what extent the Equation can be generalised to counterfactual conditionals, and to indicative conditionals that refer not to singular events, but to law-like (causal or conceptual) relations, like "If it rains on dirt tracks, they get muddy" (Cruz & Oberauer, 2014; Oaksford & Chater, 2010). There is, moreover, a close relation between counterfactuals and law-like generalisations, in that one way of characterising the latter is that they support the former (Chater & Oaksford, 2013; Edgington, 2011; Oaksford & Chater, 2010a; Over, 2017; Pearl, 2000). A generalisation, expressed as a general indicative conditional, is counterfactual supporting when there is a positive correspondence between the probability of the indicative and the probability of the corresponding counterfactual. Some authors have proposed that the meaning of counterfactuals is similar to that of indicatives, and that both are assessed by the Ramsey test, but with different accounts of what that test would be (Edgington, 2008; Over, 2017; see also Pearl, 2013, on interventions and the Ramsey test). On the other hand, if law-like generalisations are conditionals for which there is a positive correspondence between the probabilities of the indicative and the counterfactual, then this suggests that such general conditionals are determined not just by $P(q|p)$, but also by $P(\text{not-}q|\text{not-}p)$, yielding a positive covariation between p and q (Cruz et al., 2016; Oaksford & Chater, 2017; Oberauer, Weidenfeld, et al., 2007; Skovgaard-Olsen et al., 2016).

In any event, the relations between indicatives and counterfactuals, and between counterfactuals and generals, are an important area of future research. The distinction between singular and general conditionals also connects to the study of quantified statements, e.g. to the "singular" "all dirt tracks [in the surroundings] got muddy when it rained yesterday", vs. the "general" "all dirt tracks get muddy when it rains" (see also Popper's 2002 distinction between numerical and specific generality, and Ramsey's 1929 distinction between conjunctions and variable hypotheticals).

Within Bayesian approaches, the quantifier "all" is often still represented using the material conditional (Adams, 1998; Howson & Urbach, 2006). But this appears to come at the cost of the paradoxes of the material conditional, which Bayesian approaches have so clearly argued against in the context of conditional statements. One might see this in the example of Raven's paradox (Hempel, 1945; Howson & Urbach, 2006). The paradox is often described using the hypothesis "All ravens are black". It is called a paradox because it was originally

described as a situation in which a seemingly counterintuitive conclusion followed from seemingly unproblematic and straightforward assumptions. These assumptions are:

- (1) Hypotheses of the form *All R are B* are confirmed by evidence of the form *R & B*.
- (2) Hypotheses that are logically equivalent are confirmed by the same evidence.

The argument is that by (1), evidence of the form *not-R & not-B* confirms the hypothesis that *All not-B are not-R*. By contraposition, *All not-B are not-R* is supposed to be equivalent to *All R are B*, so that by (2) evidence of the form *not-R & not-B* also confirms the hypothesis that *All R are B*. It follows that for example the finding of a white sock, or of a red herring, provides corroborating evidence for the hypothesis that all ravens are black, which seems counterintuitive.

Howson & Urbach (2006) argue that in the Bayesian approach, this is not really a paradox once one realises that confirming evidence comes in degrees. Both *R & B* and *not-R & not-B* confirm the hypothesis *All Rs are B*, but *not-R & not-B* does so to a far smaller extent. This is because the prior probability of observing something that is not black and not a raven is very high, and therefore less informative about the hypothesis in question (c.f. the rarity assumption, Oaksford, 2002).

However, in the above analysis the application of (2) depends on the assumption that *All R are B* is logically equivalent to *All not-B are not-A*. This is the case when "all" is expressed using the material conditional, but not when it is expressed using the probability conditional because contraposition is invalid for the probability conditional: $P(q|p)$ is not equivalent to $P(not-p|not-q)$ (c.f. Jeffrey, 1964; Stalnaker, 1968). As mentioned before, when the conditional probability for the Equation is computed via the Ramsey test rather than via the ratio formula, contraposition fails even in the case in which the probability of the premise is 1 (Gilio, 1990; Jeffrey, 1991). This seems to imply that in a fully Bayesian analysis, which replaces the material conditional with the probability conditional not only in the case of explicit conditionals statements, *if p then q*, but also in the case of universals, *all p are q*, the counterintuitive nature of the paradox is justified. The paradox then resolves not by explaining why the seemingly counterintuitive result is normative, but by pointing out that the counterintuitive nature of the result is based on assumptions about the meaning of conditionals that are themselves highly counterintuitive. As an example of how the paradoxes of the material conditional create the same problems for universals as they do for natural language conditionals, consider the hypothesis *All ravens on the moon are black*. If this hypothesis is represented using a material conditional, then it must be confirmed with certainty, and also the alternative hypothesis *All ravens on the moon are not-black* must be confirmed with certainty, by the trivial fact that there are no ravens on the moon.

A representation of universals that is in line with the Equation is the proposal $P(\text{all } p \text{ are } q) = 1$ if and only if $P(q|p) = 1$, so that the statement is true when the conditional probability is 1, and false otherwise (Chater & Oaksford, 1999; Cruz & Oberauer, 2014; Pfeifer, 2006). This

relation may express that a binary property holds for every single member of a set, for example, when the ten people in a room are wearing glasses, and not just 9 of the 10. An alternative would be to replace the classical logical expression "For all x (x is not a raven or x is black)" with the probabilistic expression "For all x (there is a specific conditional probability that x is black, given that x is a raven)", so that the statement is true when the given conditional probability applies to every single element of a set, and false if it applies only to a subset of its elements. In contrast to the probability of a conditional, the probability of the universal would then correspond not to the proportion of a set for which a property holds, but to the probability that the property holds for every single member of the set. When the property is "the coin lands heads", then this probability decreases exponentially with the number of tosses (Cruz & Oberauer, 2014).

These are just initial thoughts for a probabilistic representation of universals (see Chater & Oaksford, 1999; Oaksford & Chater, 2007; Pfeifer & Sanfilippo, 2017, for more developed accounts; and Johnson-Laird et al., 2015, as well as Baratgin et al., 2015, for a comparison with the assumptions of mental model theory). An ongoing challenge for the new paradigm is to provide a formal specification of these statements in a probabilistic predicate logic that would allow the formulation of coherence intervals for them.

Coherence and rationality

The finding that overall people tend to reason coherently from the premises to the conclusion of inferences, even if not in all cases, has wider practical implications that can only briefly be touched upon here. If people are coherent not only in their beliefs, but also between what they believe, what they want, and what they do (assuming external circumstances allow for this), then they can use this coherence to obtain clarity about what they can do to get what they want, given what they believe (Adams, 1998; Evans, Over, & Manktelow, 1993; Ramsey, 1926).

A central tool to achieve such clarity about decision making is Subjective Expected Utility (SEU, Chater & Oaksford, 2012; Jeffrey, 1990). On this account, the utility a person can expect of an action is calculated by assessing how desirable and probable the consequences of the action are, multiplying the desirability of each consequence with its probability, and summing up the resulting terms. The same can be done for an alternative action with its respective consequences, and one can then see which action has the highest subjective utility.

Adams (1998) points out that this tool can only be useful when people's degrees of belief are good approximations of the relevant facts. If people's degrees of belief are distant from what is actually the case, then their actions based on those beliefs are less likely to help them get what they want. This creates a practical incentive for having beliefs that correspond to the facts, beyond the purely logical, and "internal", considerations of coherence and p-validity.

As discussed above, our beliefs about the facts are often uncertain. They depend, not only on our concrete experiences, but also on the theories we construct to understand and navigate these experiences. Does this imply that what counts as rational is always relative to our theories about the world at a particular point in time? Coherence between our beliefs, desires and actions is sufficient to apply the SEU formula. But is it sufficient for rationality, or does rationality also involve an ability to be receptive to our environment, so that we can learn from experience and revise our theories about what is the case to take new information into account? And are there ways of comparing personal theories about the world as more or less justified, given a body of shared knowledge at a particular point in time, as is done with scientific theories (Howson & Urbach, 2006)? These are difficult questions (Elqayam & Over, 2016), but relevant to them are the increasing studies of the legal, moral, social, and environmental contexts of subjective probability and utility judgments (Hahn, Harris, & Corner, 2016; Lagnado & Gerstenberg, 2017; Misyak & Chater, 2014; Stanovich, West, & Toplak, 2013). These studies should broaden our understanding of human thought and action, and contribute to further integration of research on reasoning, learning, and decision making.

References

- Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht, NL: Reidel Publishing Company.
- Adams, E. (1996). Four probability-preserving properties of inferences. *Journal of Philosophical Logic*, 25(1), 1-24.
- Adams, E. (1998). *A primer of probability logic*. Stanford, US: CLSI publications.
- Allan, L. G. (1980). A note on the measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychometric Society*, 15, 147–149.
- Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, 119, 403-418.
- Appiah, A. (1984). Jackson on the material conditional. *Australasian journal of Philosophy*, 62, 77-81.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Baratgin, J., Douven, I., Evans, J. St. B. T., Oaksford, M., Over, D. E., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, 19(10), 547-548.
- Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and the de Finetti tables. *Thinking & Reasoning*, 19, 308-328.
- Baratgin, J., & Politzer, G. (2010). Updating: A psychologically basic situation of probability revision. *Thinking & Reasoning*, 16, 245-287.

- Baratgin J., & Politzer, G. (2016). Logic, probability, and inference: A methodology for a new paradigm. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive Unconscious and Human Rationality*. Cambridge, US: MIT Press.
- Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, *65*(6), 1119-1131.
- Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.
- Barrouillet, P., & Gauffroy, C. (2015). Probability in reasoning: A developmental test on conditionals. *Cognition*, *137*, 22-39.
- Barrouillet, P., Gauffroy, C., & Lecas, J-F. (2008). Mental models and the suppositional account of conditionals. *Psychological Review*, *115*(3), 760-772.
- Barrouillet, P., Grosset, N., & Lecas, J-F. (2000). Conditional reasoning by mental models: Chronometric and developmental evidence. *Cognition*, *75*, 237-266.
- Barrouillet, P., & Lecas, J-F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning*, *5*(4), 289-302.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv:1506.04967v1 [stat.ME]. Retrieved from <http://arxiv.org/abs/1506.04967>.
- Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford, UK: Oxford University Press.
- Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind and Language*, *19*(2), 199-210.
- Bonnefon, J-F. (2013). New ambitions for a new paradigm: Putting the psychology of reasoning at the service of humanity. *Thinking & Reasoning*, *19*(3), 381-398.
- Bradley, R. (2012). Multidimensional possible-world semantics for conditionals. *Philosophical Review*, *121*(4), 539-571.
- Braine, M. D. S., & O'Brien, D. P. (Eds.) (1998). *Mental logic*. Mahwah, US: Lawrence Erlbaum.
- Braine, M. D. S., Reiser, B. J., & Rumin, B. (1984). Some empirical justification for a theory of natural propositional logic. *Psychology of Learning and Motivation*, *18*, 313-337.
- Braithwaite, R. B. (1953). *Scientific Explanation*. Cambridge, UK: Cambridge University Press.
- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301-338.

- Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23(3), 247-303.
- Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31, 61-83.
- Byrne, R. M. J. (2016). Counterfactual thought. *Annual Review of Psychology*, 67, 135-157.
- Byrne, R. M. J., & Johnson-Laird, P. N. (2009). 'If' and the problems of conditional reasoning. *Trends in Cognitive Sciences*, 13(7), 282-287.
- Byrne, R. M. J., & Johnson-Laird, P. N. (2010). Conditionals and possibilities. In M. Oaksford, & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thought* (pp. 55-68). Oxford, UK: Oxford University Press.
- Capotorti, A., Galli, L., & Vantaggi, B. (2003). How to use locally strong coherence in an inferential process based on upper-lower probabilities. *Soft Computing*, 7(5), 280-287.
- Cariani, F., & Rips, L. J. (2017). Conditionals, context, and the suppression effect. *Cognitive Science*, 41(3), 540-589.
- Carnap, R. (1950/1962). *The logical Foundations of Probability* (2nd. Ed.). Chicago, US: University of Chicago Press.
- Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 7, 335-344.
- Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38, 191-258.
- Chater, N., & Oaksford, M. (2012). Normative systems: Logic, probability, and rational choice. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 11-21). New York: Oxford University Press.
- Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, 37, 1171-1191.
- Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6), 811-823.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367-405.
- Coletti, G., & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Dordrecht, NL: Kluwer.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1), 1-73.
- Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3), 463-480.
- Costello, F., & Watts, P. (2016a). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, 89, 106-133.

- Costello, F., & Watts, P. (2016b). A test of two models of probability judgment: Quantum versus noisy probability. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1104-1109). Philadelphia, USA: Cognitive Science Society.
- Costello, F., & Watts, P. (2017). Mathematical invariants in people's probabilistic reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *The 39th Annual Meeting of the Cognitive Science Society*. London, UK: Cognitive Science Society.
- Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*(2), 182-199.
- Crupi, V., Tentori, K., & Lombardi, L. (2009). Pseudodiagnosticity revisited. *Psychological Review*, *116*(4), 971-985.
- Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology*, *6*, 192.
- Cruz, N., & Oberauer, K. (2014). Comparing the meanings of "if" and "all". *Memory & Cognition*, *42*, 1345-1356.
- Cruz, N., Over, D., & Oaksford, M. (2017). The elusive oddness of or-introduction. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *The 39th Annual Meeting of the Cognitive Science Society* (pp. 259-264). London, UK: Cognitive Science Society.
- Cruz, N., Over, D., Oaksford, M., & Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1104–1109). Austin, US: Cognitive Science Society.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180.
- Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646-658.
- Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*(3), 274-282.
- De Finetti, B. (1936/1995): The logic of probability. *Philosophical Studies*, *77*, 181–190.
- De Finetti, B. (1937/1980). Foresight: its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokier (Eds.), *Studies in subjective probability* (55-118). New York, US: Wiley.
- De Finetti, B. (1970/1974). *Theory of Probability: A critical introductory treatment* (Vols. 1 & 2). Bristol, UK: John Wiley & Sons.
- De Finetti, B. (1972). *Probability, induction, and statistics*. London, UK: Wiley.
- De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*(5), 428-433.

- De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, 7(1), 28-38.
- De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, 20(2), 169-187.
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, 6(1), Art. e15954.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual-process theories of thinking. *Cognition*, 106, 1248-1299.
- De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 208-216.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, 31(4), 581-595.
- De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, 19(5), 483-489.
- Dieussaert, K., Schaeken, W., & d'Ydewalle, G. (2002). The relative contribution of content and context factors on the interpretation of conditionals. *Experimental Psychology*, 49(3), 181-195.
- Douven, I. (2011). Abduction. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. URL = <<http://plato.stanford.edu/archives/spr2011/entries/abduction/>>.
- Douven, I. (2012). Learning conditional information. *Mind & Language*, 27(3), 239-263.
- Douven, I. (2015a). On de Finetti on iterated conditionals. Tech. rep. Paris, France: CNRS.
- Douven, I. (2015b). How to account for the oddness of missing-link conditionals. *Synthese*. DOI: 10.1007/s11229-015-0756-7.
- Douven, I. (2016). *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*. Cambridge, UK: Cambridge University Press.
- Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, 117(3), 831-863.
- Dudley, R. E. J., & Over, D. E. (2003). People with delusions jump to conclusions: A theoretical account of research findings on the reasoning of people with delusions. *Clinical Psychology and Psychotherapy*, 10, 263-274.
- Edgington, D. (1995). On conditionals. *Mind*, 104, 235-329.
- Edgington, D. (2008). Counterfactuals. *The Presidential Address, Proceedings of the Aristotelian Society*, CVIII(1), 1-21.
- Edgington, D. (2011). Causation first: Why causation is prior to counterfactuals. Birkbeck, University of London manuscript.

- Edgington, D. (2014, Winter). Indicative conditionals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. URL = [<https://plato.stanford.edu/archives/win2014/entries/conditionals/>](https://plato.stanford.edu/archives/win2014/entries/conditionals/).
- Elqayam, S. (2016). Scams and rationality: Dutch book arguments are not all they're cracked up to be. In N. Galbraith, D. E. Over, and E. J. Lucas (Eds.), *The thinking mind: A Festschrift for Ken Manktelow*. Hove, UK: Psychology Press.
- Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by S. Elqayam, J.F. Bonnefon, & D. E. Over. *Thinking & Reasoning*, *19*, 249-265.
- Elqayam, S., & Over, D. (2016). Editorial: From is to ought: The place of normative models in the study of human thought. *Frontiers in Psychology*, *7*, Art. 628.
- Espino, O., & Byrne, R. M. J. (2013). The compatibility heuristic in non-categorical hypothetical reasoning: Inferences between conditionals and disjunctions. *Cognitive Psychology*, *67*, 98-129.
- Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology*, *24*, 193-199.
- Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454-459.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378-395.
- Evans, J. St. B. T. (2007). On the resolution of conflict in dual-process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321-339.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.
- Evans, J. St. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*, 313-326.
- Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *2012*, *18*(1), 5-31.
- Evans, J. St. B. T. (2017). Dual process theory: Perspectives and problems. In W. De Neys (Ed.), *Dual Process Theory 2.0*. London, UK: Routledge.
- Evans, J. St. B. T., Clibbens, J., & Rood, B. (1995). Bias in conditional inference: Implications for mental models and mental logic. *The Quarterly Journal of Experimental Psychology*, *48A*(3), 644-670.
- Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *The Quarterly Journal of Experimental Psychology*, *52A*(3), 739-769.
- Evans, J. St. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure: A study of causal conditional inference. *Experimental Psychology*, *2009*, *56*(2), 77-83.

- Evans, J. St. B. T., Handley, S. J., Hadjichristidis, C., Thompson, V., Over, D., & Bennett, S. (2007). On the basis of belief in causal and diagnostic conditionals. *The Quarterly Journal of Experimental Psychology*, *60*(5), 635-643.
- Evans, J. St. B. T., Handley, S., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition*, *35*, 1772-1784.
- Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology*, *63*(5), 892-909.
- Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 321-335.
- Evans, J. St. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*, 197-213.
- Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, UK: Oxford University Press.
- Evans, J. St. B. T., & Over, D. E. (2013). Reasoning to and from belief: Deduction and induction are still distinct. *Thinking & Reasoning*, *19*, 268-283.
- Evans, J. St. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review*, *112*(4), 1040-1052.
- Evans, J. St. B. T., Over, D. E., & Manktelow, K. I. (1993). Reasoning, decision making, and rationality. *Cognition*, *49*, 165-187.
- Evans, J. St. B. T., Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.
- Evans, J. St. B. T., Thompson, V., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology*, *6*, 398.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329-336.
- Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*, 119-130.
- Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh, UK: Oliver and Boyd.
- Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105-128.
- Fugard, A. J. B., Pfeifer, N., & Mayerhofer, B. (2011). Probabilistic theories of reasoning need pragmatics too: Modulating relevance in uncertain conditionals. *Journal of Pragmatics*, *43*, 2034-2042.

- Fugard, A. J. B., Pfeifer, N., Mayerhofer, B., & Kleiter, G. D. (2011). How people interpret conditionals: Shifts toward the conditional event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 635-648.
- Geiger, S. M., & Oberauer, K. (2007). Reasoning with conditionals: Does every counterexample count? It's frequency that counts. *Memory & Cognition*, *35*(8), 2060-2074.
- Geiger, S. M., & Oberauer, K. (2010). Towards a reconciliation of mental model theory and probabilistic theories of conditionals. In M. Oaksford, & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (289-308). Oxford, UK: Oxford University Press.
- George, C. (1997). Reasoning from uncertain premises. *Thinking and Reasoning*, *3*, 161-190.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684-704.
- Gilio, A. (1990). Criterio di penalizzazione e condizioni di coerenza nella valutazione soggettiva della probabilità. *Bollettino dell'Unione Matematica Italiana*, *7a*, 4-B(3), 645-660.
- Gilio, A. (2002). Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence*, *34*, 5-34.
- Gilio, A., & Over, D.E. (2012). The psychology of inferring conditionals from disjunctions: A probabilistic study. *Journal of Mathematical Psychology*, *56*, 118-131.
- Gilio, A., Over, D. E., Pfeifer, N., & Sanfilippo, G. (2016). Centering and compound conditionals under coherence. In M. B. Ferraro, P. Giordani, B. Vantaggi, M. Gagolewski, M. A. Gil, P. Grzegorzewski, & O. Hryniewicz (Eds.), *Soft methods for data science* (pp. 253-260). Springer.
- Gilio, A., & Sanfilippo, G. (2014). Conditional random quantities and compounds of conditionals. *Studia Logica*, *102*(4), 709-729.
- Giroto, V., & Johnson-Laird, P. N. (2004). The probability of conditionals. *Psychologia*, *47*, 207-225.
- Giroto, V., & Johnson-Laird, P. N. (2010). Conditionals and probability. In M. Oaksford, & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (103-116). Oxford, UK: Oxford University Press.
- Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435-441.
- Goel, V. (2009). Cognitive neuroscience of thinking. In G. Berntson, & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences, I* (pp. 417-430). New York, US: John Wiley & Sons Inc.
- Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General*, *143*, 1214 -1241.

- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, Massachusetts: Harvard University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661-716.
- Hadjichristidis, C., Sloman, S. A., & Over, D. E. (2014). Categorical induction from uncertain premises: Jeffrey's doesn't completely rule. *Thinking & Reasoning*, *20*(4), 405-431.
- Hahn, U. (2014) The Bayesian boom: good thing or bad? *Frontiers in Psychology* *5*, Art. 765.
- Hahn, U., Harris, A. J. L., & Corner, A. (2016). Public perception of climate science: Coherence, reliability, and independence. *Topics in Cognitive Science*, *8*, 180-195.
- Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*, 646-678.
- Hájek, A. (2012, Winter). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. URL = <https://plato.stanford.edu/archives/win2012/entries/probability-interpret/>.
- Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28-43.
- Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. *Psychology of Learning and Motivation*, *62*, 33-58.
- Hartmann, S., & Rafiee-Rad, S. (2014). Learning conditionals. University of Munich manuscript.
- Hattori, M. (2016). Probabilistic representation in syllogistic reasoning: A theory to integrate mental models and heuristics. *Cognition*, *157*, 296-320.
- Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind*, *54*, 1-26, 97-121.
- Hintikka (1965). *On a Combined System of Inductive Logic*. *Studia Logico-Mathematica et Philosophica in Honorem Rolf Nevanlinna, Acta Philosophica Fennica*, *18*, 21-30.
- Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd Ed.). Illinois, US: Open Court.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. New York, US: Routledge.
- Hsu, A. S., Chater, N., & Vitányi, P. M. B. (2011). The probabilistic analysis of language acquisition: theoretical, computational, and experimental analysis. *Cognition*, *120*, 380-390.
- Jackson, F. (1987). *Conditionals*. Oxford, UK: Blackwell.
- Jeffrey, R. C. (1964). If (abstract). *Journal of Philosophy*, *61*, 702-703.
- Jeffrey, R. C. (1990). *The Logic of Decision* (2nd Ed.). Chicago, US: University of Chicago Press.

- Jeffrey, R. C. (1991). Matter of fact conditionals. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 65, 161-183.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, US: Erlbaum.
- Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, 50, 189-209.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109(4), 646-678.
- Johnson-Laird, P. N., Khemlani, S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19(4), 201-214.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62-88.
- Johnson-Laird, P. N., & Ragni, M. (2017, July). What are the possibilities? In M. Oaksford, V. Thompson, & N. Adams (Chairs), *The 10th London Reasoning Workshop*. London, UK: Birkbeck, University of London.
- Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, 71, 191-229.
- Johnson-Laird, P. N., & Tagart, J. (1969). How implication is understood. *The American Journal of Psychology*, 82(3), 367-373.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54-69.
- Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, 116, 856-874.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-292.
- Kant, I. (1781/1998). *Kritik der Reinen Vernunft*. Hamburg, DE: Felix Meiner Verlag.
- Kellen, D., Singmann, H., & Batchelder, W. H. (2017). Classic-probability accounts of mirrored (Quantum-like) order effects in human judgments. *Decision*. DOI: <<http://dx.doi.org/10.1037/dec0000080>>.
- Kemeny, J. (1955). Fair bets and inductive probabilities. *Journal of Symbolic Logic*, 20(3), 263-273.
- Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13, 150-158.
- Keynes, J. M. (1921). *A treatise on Probability*. London, UK: Macmillan.
- Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8, Art. 849.

- Khemlani, S. S., Hinterecker, T., & Johnson-Laird, P. N. (2017). The provenance of modal inference. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *The 39th Annual Meeting of the Cognitive Science Society*. London, UK: Cognitive Science Society.
- Khemlani, S. S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: Model-based estimates of unique events. *Cognitive Science*, 39(6), 1216-1258.
- Klauer, K. C., Beller, S., & Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2), 298-323.
- Klayman, J., & Ha, Young-Won (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211-228.
- Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford, UK: Oxford University Press
- Kolmogorov, A. N. (1933/1950). *Foundations of the theory of probability*. New York, US: Chelsea Publishing Company.
- Kripke, S. (1963). Semantical analysis of modal logics, I. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9, 67-96.
- Krzyżanowska, K., Wenmackers, S., and Douven, I. (2013). Inferential conditionals and evidentiality. *Journal of Logic, Language and Information*, 22(3), 315-334.
- Kyburg, H. E., & Smokler, H. E. (1980). Introduction. In H. E. Kyburg, & H. E. Smokler, *Studies in Subjective Probability* (pp. 1-22). New York, US: Robert E. Krieger Publishing Company.
- Lagnado, D. A. & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In Michael Waldmann (Ed.), *Oxford Handbook of Causal Reasoning* (pp. 565-601). Oxford University Press.
- Lehman, R. S. (1955). On confirmation and rational betting. *Journal of Symbolic Logic*, 20(3), 251-262.
- Leiner, D. J. (2014). SoSci Survey (Version 2.5.00-i) [Computer software]. Available at <https://www.soscisurvey.de>
- Lewis, D. (1973). *Counterfactuals*. Oxford, UK: Basil Blackwell.
- Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297-315.
- Liu, I., Lo, K., & Wu, J. (1996). A probabilistic interpretation of "if-then". *Quarterly Journal of Experimental Psychology*, 49A, 828-844.
- Malinas, G., & Bigelow, J. (2016, Fall). Simpson's paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. URL = <https://plato.stanford.edu/archives/fall2016/entries/paradox-simpson/>.

- Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition*, *39*, 85-105.
- Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review*, *22*, 5-36.
- Markovits, H., Brisson, J., & Chantal, P.-L. (2015). Additional evidence for a dual-strategy model of reasoning: Probabilistic reasoning is more invariant than reasoning about logical validity. *Memory & Cognition*, *43*, 1208-1215.
- Markovits, H., Brunet, M.-L., Thompson, V., & Brisson, J. (2013). Direct evidence for a dual-process model of deductive inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1213-1222.
- Markovits, H., Lortie-Forgues, H., & Brunet, M.-L. (2010). Conditional reasoning, frequency of counterexamples, and the effect of response modality. *Memory & Cognition*, *38*(4), 485-492.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, US: Freeman.
- Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203-220.
- McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, *82*(9), 462-471.
- Misyak, J. B., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical transactions of the Royal Society B*, *369*, 1-9.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 for generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133-142
- Neyman, J. (1952). *Lectures and Conferences on mathematical Statistics and Probability* (2nd Ed.). Washington, US: Department of Agriculture.
- Oaksford, M. (2002). Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking and Reasoning*, *8*(2), 135-151.
- Oaksford, M. (2013). Quantum probability, intuition, and human rationality. *Behavioral and Brain Sciences*, *36*, 303.
- Oaksford, M. (2015). Imaging deductive reasoning and the new paradigm. *Frontiers in Human Neuroscience*, *9*, Art. 101.
- Oaksford, M., & Chater, N. (1996). Rational explanation of the Selection task. *Psychological Review*, *103*(2), 381-391.
- Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349-357.
- Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, *10*(2), 289-318.

- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.
- Oaksford, M., & Chater, N. (2010a). Causation and conditionals in the cognitive science of human reasoning. *The Open Psychology Journal*, 3, 105-118.
- Oaksford, M., & Chater, N. (2010b). Conditionals and constraint satisfaction: Reconciling mental models and the probabilistic approach? In M. Oaksford, & N. Chater (Eds.), *Cognition and Conditionals: Probability and logic in human thought* (pp. 309-334). Oxford, UK: Oxford University Press.
- Oaksford, M., & Chater, N. (2011). Dual systems and dual processes but a single function. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The Science of Reason: A festschrift for Jonathan St. B. T. Evans* (pp. 339-352). Hove, UK: Psychology Press.
- Oaksford, M., & Chater, N. (2012). Dual processes, probabilities, and cognitive architecture. *Mind & Society*, 11, 15-26.
- Oaksford M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, 19, 346-379.
- Oaksford, M., & Chater, N. (2016). Probabilities, causation, and logic programming in conditional reasoning: Reply to Stenning and van Lambalgen (2016). *Thinking & Reasoning*, 22(3), 355-368.
- Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. Oxford, UK: Oxford University Press.
- Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(4), 883-899.
- Oaksford, M., & Hall, S. (2016). On the source of human irrationality. *Trends in Cognitive Sciences*, 20(5), 336-344.
- Oaksford, M., Over, D., & Cruz, N. (2018). Paradigms, possibilities, and probabilities: Comment on Hinterecker et al. (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4), 835-854.
- Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, 53, 238-283.
- Oberauer, K. (2013). The focus of attention in working memory – from metaphors to mechanisms. *Frontiers in Human Neuroscience*, 7, 673.

- Oberauer, K., Geiger, S., & Fischer, K. (2010). Conditionals and disjunctions. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The Science of Reason: A festschrift for Jonathan St. B. T. Evans* (pp. 93-118). Hove, UK: Psychology Press.
- Oberauer, K., Geiger, S. M., & Fischer, K. (2011). Conditionals and disjunctions. In K. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The Science of Reason: A Festschrift for Jonathan St. B. T. Evans* (pp. 93-118). Hove, UK: Psychology Press.
- Oberauer, K., Geiger, S. M., Fischer, K., & Weidenfeld, A. (2007). Two meanings of “if”? Individual differences in the interpretation of conditionals. *The Quarterly Journal of Experimental Psychology*, *60*(6), 790–819.
- Oberauer, K., & Oaksford, M. (2008). What must a psychological theory of reasoning explain? Comment on Barrouillet, Gauffroy, and Lecas (2008). *Psychological Review*, *115*(3), 773-778.
- Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. *Thinking & Reasoning*, *13*(4), 340-369.
- Oberauer, K., & Wilhelm, O. (2000). Effects of directionality in deductive reasoning: I. The comprehension of single relational premises. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1702-1712.
- Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 680-693.
- Orenes, I., & Johnson-Laird, P. N. (2012). Logic, models, and paradoxical inferences. *Mind & Language*, *27*(4), 357-377.
- Over, D. E. (1993). Deduction and degrees of belief. *Behavioral and Brain Sciences*, *16*, 361-362.
- Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*, 431-438.
- Over, D. E. (2016). The paradigm shift in the psychology of reasoning: The debate. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive unconscious and human rationality* (pp. 79-97). Cambridge US: MIT Press.
- Over, D. E. (2017). Causation and the probability of causal conditionals. In M. Waldmann (Ed.), *The Oxford handbook of causal reasoning*. Oxford, UK: Oxford University Press.
- Over, D. E., & Baratgin, J. (2017). The “defective” truth table: Its past, present, and future. In N. Galbraith, D. E. Over, & E. Lucas, (Eds.), *The thinking mind: The use of thinking in everyday life* (pp. 15-28). Hove, UK: Psychology Press.
- Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In L. J. Ball, & V. A. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 434-450). Hove, UK: Psychology Press.

- Over, D., & Cruz, N. (under revision). Philosophy and the psychology of conditional reasoning. In A. Aberdein, & M. Inglis (Eds.), *Advances in Experimental Philosophy of Logic and Mathematics*. London, UK: Bloomsbury Publishing Plc.
- Over, D. E., Evans, J. St. B. T., & Elqayam, S. (2010). Conditionals and non-constructive reasoning. In M. Oaksford, & N. Chater (Eds.), *Cognition and Conditionals: Probability and Logic in Human Thinking* (pp. 135-151). Oxford, UK: Oxford University Press.
- Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*, 62-97.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, US: Morgan-Kaufmann.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
- Pearl, J. (2013). Structural counterfactuals: a brief introduction. *Cognitive Science*, *37*, 977-985.
- Peer, E., Brandimarte, L., Samat, S., & Aquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.
- Pfeifer, N. (2006). Contemporary syllogistics: Comparative and quantitative syllogisms. In G. Kreuzbauer & G. J. W. Dorn (Eds.), *Argumentation in Theorie und Praxis: Philosophie und Didaktik des Argumentierens* (pp. 57–71). Vienna, Austria: LIT.
- Pfeifer, N., & Kleiter, G. D. (2005). Coherence and nonmonotonicity in human reasoning. *Synthese* *146*(1-2), 93–109.
- Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206-217.
- Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford, & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 153-173). Oxford, UK: Oxford University Press.
- Pfeifer, N., & Sanfilippo, G. (2017). Probabilistic squares and hexagons of opposition under coherence. *International Journal of Approximate Reasoning*, *88*, 282-294.
- Pfeifer, N., & Stöckle-Schobel, R. (2015, September). Uncertain conditionals and counterfactuals in (non-)causal settings. In G. Arienti, B. G. Bara, & G. Sandini (Eds.), *Proceedings of the EuroAsianpacific Joint Conference on Cognitive Science* (pp. 651-656). Torino, Italy.
- Politzer, G., & Baratgin, J. (2016). Deductive schemas with uncertain premises using qualitative probability expressions. *Thinking & Reasoning*, *22*, 78-98.
- Politzer, G., & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research*, *20*(2), 83-103.

- Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking & Reasoning*, *16*, 172-197.
- Popper, K. (2002). *The logic of scientific discovery*. New York, US: Routledge.
- Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*, 255-327).
- Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (2017). The rational status of Quantum Cognition. *Journal of Experimental Psychology: General*, *146*(7), 968-987.
- Prado, J., Chadha, A., & Booth, J. R. (2011). The brain network for deductive reasoning: A quantitative meta-analysis of 28 neuroimaging studies. *Journal of Cognitive Neuroscience*, *23*(11), 3483-3497.
- Priest, G. (2008). *An introduction to non-classical logic*. New York, US: Cambridge University Press.
- Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561-588.
- Ramsey, F. P. (1926/1990). Truth and probability. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 52-94). Cambridge, UK: Cambridge University Press.
- Ramsey, F. P. (1929/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145-163). Cambridge, UK: Cambridge University Press.
- Reichenbach (1949). *The theory of probability*.
- Rips, L. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*, 38-71.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, US: MIT Press.
- Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science*, *12*(2), 129-134.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351-357.
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883-893.
- Schroyens, W., Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2000). Conditional reasoning with negations: Implicit and explicit affirmation or denial and the role of contrast classes. *Thinking and Reasoning*, *6*(3), 221-251.
- Schwan, B., & Stern, R. (2017). A causal understanding of when and when not to Jeffrey conditionalize. *Philosopher's imprint*, *17*(8), 1-21.
- Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*, 191-198.
- Singmann, H., & Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. *Thinking & Reasoning*, *17*(3), 247-281.
- Singmann, H., Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, *5*, 316.

- Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition, 150*, 26-36.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22.
- Sloman, S. (2005). *How people think about the world and its alternatives*. New York, US: Oxford University Press.
- Sloman, S. A. (2014). Two systems of reasoning, an update. In J. Sherman, B. Gawronski, & Y. Trope, (Eds.). *Dual process theories of the social mind*. New York, US: Guilford Press.
- Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5-39.
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*, 296-309.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.
- Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science, 37*(6), 1074-1106.
- Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98-112). Oxford, UK: Blackwell.
- Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science, 37*, 64-80.
- Stalnaker, K., & Jeffrey, R. (1994). Conditionals as random variables. In E. Eels, & B. Skyrms (Eds.), *Probability and conditionals: Belief revision and rational decision* (pp. 31-46). Cambridge, UK: Cambridge University Press.
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Psychological Science, 22*(4), 259-264.
- Stern, R. (2017). A causal understanding of when and when not to Jeffrey conditionalize. *Philosopher's Imprint, 17*(8), 1-21.
- Stern, R. (2017, June). Talk on dynamic reasoning with embedded conditionals. In J. Baratgin, S. Hartmann, & G. Sanfilippo (Chairs), *Third Villa Vigoni trilateral workshop on Human rationality: Probabilistic perspectives*. Lovenno di Menaggio, Italy.
- Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Journal of Experimental Psychology, 48A*(3), 613-643.
- Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1-26.
- Tabachnick, B., & Fidell, L. S. (2007). *Using multivariate statistics: International edition*. London, UK: Pearson.

- Te Grotenhuis, M., Eisinga, R., & Subramanian, S. V. (2011). Robinson's *Ecological Correlations and the Behavior of Individuals: Methodological corrections*. *International Journal of Epidemiology*, *40*(4), 1123-1125.
- Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: probability vs inductive confirmation. *Journal of Experimental Psychology: General*, *142*, 235-255.
- Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*, 467 – 477.
- Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition*, *22*, 742-758.
- Thompson, V. A., & Byrne, R. M. J. (2002). Reasoning counterfactually: Making inferences about things that didn't happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1154-1170.
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215-244.
- Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1448-1457.
- Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, *45*(4), 539-552.
- Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.
- Tversky, A., & Köhler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567.
- Van Heijenoort, J. (Ed.) (1967). *From Frege to Goedel: A source book in mathematical logic, 1879-1931*. Cambridge, US: Harvard University Press.
- Van Wijnbergen-Huitink, J., Elqayam, S., & Over, D. E. (2015). The probability of iterated conditionals. *Cognitive Science*, *39*(4), 788-803.
- Venn, J. (1886/1963). *The Logic of Chance*. London, UK: Macmillan (reprinted 1963 in New York, US: Chelsea).
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, *11*(3), 239-278.
- Vidal, M., & Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, *29*(6), 769-781.
- Vineberg, S. (2016, Spring). Dutch book arguments. In E. Z. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. URL = <https://plato.stanford.edu/archives/spr2016/entries/dutch-book/>.

- Von Mises, R. (1951). *Probability, Statistics, and Truth*. New York, US: Macmillan.
- Von Sydow, M. (2017). Rational and semi-rational explanations of the conjunction fallacy: A polycasual approach. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *The 39th Annual Meeting of the Cognitive Science Society*. London, UK: Cognitive Science Society.
- Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3), 273-281.
- Westfall, J., Judd, C. M., & Kenny, D. A. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020-2045.
- Wittgenstein, L. (1991). *On certainty*. G.E. M. anscombe, & G. H. von Wright (Eds.), Denis Paul (Transl.). Oxford, UK: Blackwell.
- Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, 16, 288-307.
- Zhao, J., & Osherson, D. (2014). Category-based updating. *Thinking & Reasoning*, 20(1), 1-15.

Appendix A

Jeffrey tables for the probability biconditional.

Table A2 shows the Jeffrey table for the probability biconditional, *if p then q & if q then p*, which adds the converse of the original conditional. Table A3 shows the Jeffrey table for another probability biconditional *if p then q & if not-p then not-q*, which adds the inverse of the original conditional. Table A1 shows the classical truth table for the material biconditional for comparison. The Jeffrey tables were derived from Gilio et al. (2016).

Table A1. Truth table for the material biconditional.

<i>p, q</i>	1
<i>p, not-q</i>	0
<i>not-p, q</i>	0
<i>not-p, not-q</i>	1

Table A2. The Jeffrey table for the probability biconditional *if p then q & if q then p*.

	<i>q p</i>	<i>p q</i>	<i>(q p)&(p q)</i>
<i>p, q</i>	1	1	1
<i>p, not-q</i>	0	$P(p q)$	0
<i>not-p, q</i>	$P(q p)$	0	0
<i>not-p, not-q</i>	$P(q p)$	$P(p q)$	$P[(q p)&(p q)] = P(p \ \& \ q/p \ \text{or} \ q)$

Table A3. Jeffrey table for the probability biconditional *if p then q & if not-p then not-q*.

	<i>q p</i>	<i>not-q not-p</i>	<i>(q p)&(not-q not-p)</i>
<i>p, q</i>	1	$P(\text{not-}q \text{not-}p)$	$P(\text{not-}q \text{not-}p)$
<i>p, not-q</i>	0	$P(\text{not-}q \text{not-}p)$	0
<i>not-p, q</i>	$P(q p)$	0	0
<i>not-p, not-q</i>	$P(q p)$	1	$P(q p)$

Appendix B

The materials used in Experiment 1.

Table B1. The scenarios used in the experiment.

Linda	Linda is single, outspoken, and intelligent. She majored in Philosophy at university, was concerned with social justice, and was anti-nuclear.
Nigel	Nigel comes from a wealthy family and got a first class degree in Classics at Oxford, where he was president of the wine society. He campaigned for the ban on fox hunting to be repealed.

Table B2. The 24 sentences used in the statements task, consisting of all premises and conclusions occurring in the inferences task.

Booklet	Statement type	#	Content
Linda	<i>p or q</i>	1	Linda votes for the Labour Party or the Green Party.
		2	Linda votes for the Labour Party or the Conservative Party.
		3	Linda is a social worker or a youth worker.
		4	Linda is a social worker or a bank teller.
	<i>if not-p then q</i>	5	If Linda does not vote for the Labour Party, then she votes for the Green Party.
		6	If Linda does not vote for the Labour Party, then she votes for the Conservative Party.
		7	If Linda is not a social worker then she is a youth worker.
		8	If Linda is not a social worker then she is a bank teller.
	<i>p</i>	9	Linda is a social worker.
		10	Linda is a youth worker.
		11	Linda is a bank teller.
		12	Linda votes for the Labour Party.
		13	Linda votes for the Green Party.
		14	Linda votes for the Conservative Party.
	<i>not-p or q</i>	15	Linda does not vote or she votes for the Labour Party.
		16	Linda does not vote or she votes for the Conservative Party.
		17	Linda does not work in the service sector or she is a social worker.
		18	Linda does not work in the service sector or she is a bank teller.
	<i>if p then q</i>	19	If Linda does votes, then she votes for the Labour Party.
		20	If Linda votes, then she votes for the Conservative Party.
		21	If Linda works in the service sector, then

			she is a social worker.
		22	If Linda works in the service sector then she is a bank teller.
	<i>not-p</i>	23	Linda does not vote.
		24	Linda does not work in the service sector.
Nigel	<i>p or q</i>	1	Nigel drives a BMW or a Porsche.
		2	Nigel drives a BMW or a Ford Fiesta.
		3	Nigel is a corporate lawyer or a merchant banker.
		4	Nigel is a corporate lawyer or an organic food salesman.
	<i>if not-p then q</i>	5	If Nigel does not drive a BMW, then he drives a Porsche.
		6	If Nigel does not drive a BMW, then he drives a Ford Fiesta.
		7	If Nigel is not a corporate lawyer, then he is a merchant banker.
		8	If Nigel is not a corporate lawyer, then he is an organic food salesman.
	<i>p</i>	9	Nigel drives a BMW.
		10	Nigel drives a Porsche.
		11	Nigel drives a Ford Fiesta.
		12	Nigel is a corporate lawyer.
		13	Nigel is a merchant banker.
		14	Nigel is an organic food salesman.
	<i>not-p or q</i>	15	Nigel does not drive a car or he drives a Porsche.
		16	Nigel does not drive a car or he drives a Ford Fiesta.
		17	Nigel does not work in the service sector or he is a corporate lawyer.
		18	Nigel does not work in the service sector or he is an organic food salesman.
	<i>if p then q</i>	19	If Nigel drives a car, then he drives a Porsche.
		20	If Nigel drives a car, then he drives a Ford Fiesta.
		21	If Nigel works in the service sector, then he is a corporate lawyer.
		22	If Nigel works in the service sector, then he is an organic food salesman.
	<i>not-p</i>	23	Nigel does not drive a car.
		24	Nigel does not work in the service sector.

Table B3. The 16 inferences for the inferences task, booklet 1.

Booklet	Inference	#	Content
Linda	<i>p or q ∴ if not-p</i>	1h	Linda votes for the Labour Party or the Green Party.

A1	<i>then q</i>		Therefore, if Linda does not vote for the Labour Party, then she votes for the Green Party.
		1l	Linda votes for the Labour Party or the Conservative Party Therefore, if Linda does not vote for the Labour Party, then she votes for the Conservative Party.
	<i>not-p or q ∴ if p then q</i>	2h	Linda does not work in the service sector or she is a social worker. Therefore, if Linda works in the service sector, then she is a social worker.
		2l	Linda does not work in the service sector or she is a bank teller. Therefore, if Linda works in the service sector then she is a bank teller.
	<i>if p then q ∴ not-p or q</i>	3h	If Linda votes, then she votes for the Labour Party. Therefore, Linda does not vote or she votes for the Labour Party.
		3l	If Linda votes, then she votes for the Conservative Party. Therefore, Linda does not vote or she votes for the Conservative Party.
	<i>if not-p then q ∴ p or q</i>	4h	If Linda is not a social worker then she is a youth worker. Therefore, Linda is a social worker or a youth worker.
		4l	If Linda is not a social worker then she is a bank teller. Therefore, Linda is a social worker or a bank teller.
	<i>p ∴ p or q</i>	5h	Linda votes for the Labour Party. Therefore, Linda votes the Labour Party or the Green Party.
		5l	Linda votes for the Labour Party. Therefore, Linda votes for the Labour Party or the Conservative Party.
	<i>not-p ∴ not-p or q</i>	6h	Linda does not work in the service sector. Therefore, Linda does not work in the service sector or she is a social worker.
		6l	Linda does not work in the service sector. Therefore, Linda does not work in the service sector or she is a bank teller.
	<i>q ∴ p or q</i>	7h	Linda votes for the Green Party. Therefore, Linda votes for the Labour Party or the Green Party.
		7l	Linda votes for the Conservative Party. Therefore, Linda votes the Labour Party or the Conservative Party.
	<i>q ∴ not-p or q</i>	8h	Linda is a social worker. Therefore, Linda does not work in the service sector or she is a social worker.
		8l	Linda is a bank teller. Therefore, Linda does not work in the service sector or

she is a bank teller.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B4. The 16 inferences for the inferences task, booklet 2.

Booklet	Inference	#	Content
Linda A2	$p \text{ or } q$	1h	Linda votes for the Labour Party or the Green Party.
		1l	Therefore, if Linda does not vote for the Labour Party, then she votes for the Green Party.
	$\therefore \text{if not-} p \text{ then } q$	2h	Linda votes for the Labour Party or the Conservative Party. Therefore, if Linda does not vote for the Labour Party, then she votes for the Conservative Party.
		2l	Linda does not work in the service sector or she is a social worker. Therefore, if Linda works in the service sector, then she is a social worker.
	$\text{not-}p \text{ or } q \therefore \text{if } p \text{ then } q$	3h	Linda does not work in the service sector or she is a bank teller. Therefore, if Linda works in the service sector then she is a bank teller.
		3l	If Linda works in the service sector, then she is a social worker. Therefore, Linda does not work in the service sector or she is a social worker.
	$\text{if } p \text{ then } q \therefore \text{not-}p \text{ or } q$	4h	If Linda works in the service sector then she is a bank teller. Therefore, Linda does not work in the service sector or she is a bank teller.
		4l	If Linda does not vote for the Labour Party, then she votes for the Green Party. Therefore, Linda votes for the Labour Party or the Green Party.
	$\text{if not-}p \text{ then } q \therefore p \text{ or } q$	5h	If Linda does not vote for the Labour Party, then she votes for the Conservative Party. Therefore, Linda votes for the Labour Party or the Conservative Party.
		5l	Linda votes for the Labour Party. Therefore, Linda votes for the Labour Party or the Green Party.
	$p \therefore p \text{ or } q$	6h	Linda votes for the Labour Party. Therefore, Linda votes for the Labour Party or the Green Party.
		6l	Linda votes for the Labour Party. Therefore, Linda votes for the Labour Party or the Conservative Party.
	$\text{not-}p \therefore \text{not-}p \text{ or } q$	7h	Linda does not work in the service sector. Therefore, Linda does not work in the service sector or she is a social worker.
		7l	Linda does not work in the service sector. Therefore, Linda does not work in the service sector or she is a bank teller.
	$q \therefore p \text{ or } q$	7h	Linda votes for the Green Party.

			Therefore, Linda votes for the Labour Party or the Green Party.
	7l		Linda votes for the Conservative Party. Therefore, Linda votes for the Labour Party or the Conservative Party.
$q \therefore \text{not-}p$ $\text{or } q$	8h		Linda is a social worker. Therefore, Linda does not work in the service sector or she is a social worker.
	8l		Linda is a bank teller. Therefore, Linda does not work in the service sector or she is a bank teller.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B5. The 16 inferences for the inferences task, booklet 3.

Booklet	Inference	#	Content
Linda B1	$p \text{ or } q \therefore \text{if not-}p$ $\text{then } q$	1h	Linda is a social worker or a youth worker. Therefore, if Linda is not a social worker then she is a youth worker.
		1l	Linda is a social worker or a bank teller. Therefore, if Linda is not a social worker then she is a bank teller.
	$\text{not-}p \text{ or } q \therefore \text{if } p$ $\text{then } q$	2h	Linda does not vote or she votes for the Labour Party. Therefore, if Linda votes, then she votes for the Labour Party.
		2l	Linda does not vote or she votes for the Conservative Party. Therefore, if Linda votes, then she votes for the Conservative Party.
	$\text{if } p \text{ then } q \therefore \text{not-}$ $p \text{ or } q$	3h	If Linda works in the service sector, then she is a social worker. Therefore, Linda does not work in the service sector or she is a social worker.
		3l	If Linda works in the service sector then she is a bank teller. Therefore, Linda does not work in the service sector or she is a bank teller.
	$\text{if not-}p \text{ then } q \therefore$ $p \text{ or } q$	4h	If Linda does not vote for the Labour Party, then she votes for the Green Party. Therefore, Linda votes for the Labour Party or the Green Party.
		4l	If Linda does not vote for the Labour Party, then she votes for the Conservative Party. Therefore, Linda votes for the Labour Party or the Conservative Party.
	$p \therefore p \text{ or } q$	5h	Linda is a social worker. Therefore, Linda is a social worker or a youth worker.
		5l	Linda is a social worker Therefore, Linda is a social worker or a bank teller

$not-p \therefore not-p \text{ or } q$	6h	Linda does not vote. Therefore, Linda does not vote or she votes for the Labour Party.
	6l	Linda does not vote. Therefore, Linda does not vote or she votes for the Conservative Party.
$q \therefore p \text{ or } q$	7h	Linda is a youth worker. Therefore, Linda is a social worker or a youth worker.
	7l	Linda is a bank teller. Therefore, Linda is a social worker or a bank teller.
$q \therefore not-p \text{ or } q$	8h	Linda votes for the Labour Party. Therefore, Linda does not vote or she votes for the Labour Party.
	8l	Linda votes for the Conservative Party. Therefore, Linda does not vote or she votes for the Conservative Party.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B6. The 16 inferences for the inferences task, booklet 4.

Booklet	Inference	#	Content
Linda B2	$p \text{ or } q \therefore \text{if } not-p \text{ then } q$	1h	Linda is a social worker or a youth worker. Therefore, if Linda is not a social worker then she is a youth worker.
		1l	Linda is a social worker or a bank teller. Therefore, if Linda is not a social worker then she is a bank teller.
	$not-p \text{ or } q \therefore \text{if } p \text{ then } q$	2h	Linda does not vote or she votes for the Labour Party. Therefore, if Linda votes, then she votes for the Labour Party.
		2l	Linda does not vote or she votes for the Conservative Party. Therefore, if Linda votes, then she votes for the Conservative Party.
	$\text{if } p \text{ then } q \therefore not-p \text{ or } q$	3h	If Linda votes, then she votes for the Labour Party. Therefore, Linda does not vote or she votes for the Labour Party.
		3l	If Linda votes, then she votes for the Conservative Party. Therefore, Linda does not vote or she votes for the Conservative Party.
	$\text{if } not-p \text{ then } q \therefore p \text{ or } q$	4h	If Linda is not a social worker then she is a youth worker. Therefore, Linda is a social worker or a youth worker.
		4l	If Linda is not a social worker then she is a bank teller. Therefore, Linda is a social worker or a bank teller.

$p \therefore p \text{ or } q$	5h	Linda is a social worker. Therefore, Linda is a social worker or a youth worker.
	5l	Linda is a social worker. Therefore, Linda is a social worker or a bank teller.
$\text{not-}p \therefore \text{not-}p \text{ or } q$	6h	Linda does not vote. Therefore, Linda does not vote or she votes for the Labour Party.
	6l	Linda does not vote. Therefore, Linda does not vote or she votes for the Conservative Party.
$q \therefore p \text{ or } q$	7h	Linda is a youth worker. Therefore, Linda is a social worker or a youth worker.
	7l	Linda is a bank teller. Therefore, Linda is a social worker or a bank teller.
$q \therefore \text{not-}p \text{ or } q$	8h	Linda votes for the Labour Party. Therefore, Linda does not vote or she votes for the Labour Party.
	8l	Linda votes for the Conservative Party. Therefore, Linda does not vote or she votes for the Conservative Party.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B7. The 16 inferences for the inferences task, booklet 5.

Booklet	Inference	#	Content
Nigel C1	$p \text{ or } q \therefore \text{if not-}p \text{ then } q$	1h	Nigel drives a BMW or a Porsche. Therefore, if Nigel does not drive a BMW, then he drives a Porsche.
		1l	Nigel drives a BMW or a Ford Fiesta. Therefore, if Nigel does not drive a BMW, then he drives a Ford Fiesta.
	$\text{not-}p \text{ or } q \therefore \text{if } p \text{ then } q$	2h	Nigel does not work in the service sector or he is a corporate lawyer. Therefore, if Nigel works in the service sector, then he is a corporate lawyer.
		2l	Nigel does not work in the service sector or he is an organic food salesman. Therefore, if Nigel works in the service sector, then he is an organic food salesman.
	$\text{if } p \text{ then } q \therefore \text{not-}p \text{ or } q$	3h	If Nigel drives a car, then he drives a Porsche. Therefore, Nigel does not drive a car or he drives a Porsche.
		3l	If Nigel drives a car, then he drives a Ford Fiesta. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.
	$\text{if not-}p \text{ then } q \therefore$	4h	If Nigel is not a corporate lawyer, then he is a merchant banker.

$p \text{ or } q$			Therefore, Nigel is a corporate lawyer or a merchant banker.
	4l	If Nigel is not a corporate lawyer, then he is an organic food salesman. Therefore, Nigel is a corporate lawyer or an organic food salesman.	
$p \therefore p \text{ or } q$	5h	Nigel drives a BMW. Therefore, Nigel drives a BMW or a Porsche.	
	5l	Nigel drives a BMW. Therefore, Nigel drives a BMW or a Ford Fiesta.	
$\text{not-}p \therefore \text{not-}p \text{ or } q$	6h	Nigel does not work in the service sector. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.	
	6l	Nigel does not work in the service sector. Therefore, Nigel does not work in the service sector or he is an organic food salesman.	
$q \therefore p \text{ or } q$	7h	Nigel drives a Porsche. Therefore, Nigel drives a BMW or a Porsche.	
	7l	Nigel drives a Ford Fiesta. Therefore, Nigel drives a BMW or a Ford Fiesta.	
$q \therefore \text{not-}p \text{ or } q$	8h	Nigel is a corporate lawyer. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.	
	8l	Nigel is an organic food salesman. Therefore, Nigel does not work in the service sector or he is an organic food salesman.	

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B8. The 16 inferences for the inferences task, booklet 6.

Booklet	Inference	#	Content
Nigel C2	$p \text{ or } q \therefore \text{if not-}p \text{ then } q$	1h	Nigel drives a BMW or a Porsche. Therefore, if Nigel does not drive a BMW, then he drives a Porsche.
		1l	Nigel drives a BMW or a Ford Fiesta. Therefore, if Nigel does not drive a BMW, then he drives a Ford Fiesta.
	$\text{not-}p \text{ or } q \therefore \text{if } p \text{ then } q$	2h	Nigel does not work in the service sector or he is a corporate lawyer. Therefore, if Nigel works in the service sector, then he is a corporate lawyer.
		2l	Nigel does not work in the service sector or he is an organic food salesman. Therefore, if Nigel works in the service sector, then he is an organic food salesman.
	$\text{if } p \text{ then } q \therefore \text{not-}p \text{ or } q$	3h	If Nigel works in the service sector, then he is a corporate lawyer. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.

	3l	If Nigel works in the service sector, then he is an organic food salesman. Therefore, Nigel does not work in the service sector or he is an organic food salesman.
<i>if not-p then q ∴ p or q</i>	4h	If Nigel does not drive a BMW, then he drives a Porsche. Therefore, Nigel drives a BMW or a Porsche.
	4l	If Nigel does not drive a BMW, then he drives a Ford Fiesta. Therefore, Nigel drives a BMW or a Ford Fiesta.
<i>p ∴ p or q</i>	5h	Nigel drives a BMW. Therefore, Nigel drives a BMW or a Porsche.
	5l	Nigel drives a BMW. Therefore, Nigel drives a BMW or a Ford Fiesta.
<i>not-p ∴ not-p or q</i>	6h	Nigel does not work in the service sector. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.
	6l	Nigel does not work in the service sector. Therefore, Nigel does not work in the service sector or he is an organic food salesman.
<i>q ∴ p or q</i>	7h	Nigel drives a Porsche. Therefore, Nigel drives a BMW or a Porsche.
	7l	Nigel drives a Ford Fiesta. Therefore, Nigel drives a BMW or a Ford Fiesta.
<i>q ∴ not-p or q</i>	8h	Nigel is a corporate lawyer. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.
	8l	Nigel is an organic food salesman. Therefore, Nigel does not work in the service sector or he is an organic food salesman.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B9. The 16 inferences for the inferences task, booklet 7.

Booklet	Inference	#	Content
Nigel D1	<i>p or q ∴ if not-p then q</i>	1h	Nigel is a corporate lawyer or a merchant banker. Therefore, if Nigel is not a corporate lawyer, then he is a merchant banker.
		1l	Nigel is a corporate lawyer or an organic food salesman. Therefore, if Nigel is not a corporate lawyer, then he is an organic food salesman.
	<i>not-p or q ∴ if p then q</i>	2h	Nigel does not drive a car or he drives a Porsche. Therefore, if Nigel drives a car, then he drives a Porsche.
		2l	Nigel does not drive a car or he drives a Ford Fiesta. Therefore, if Nigel drives a car, then he drives a Ford Fiesta.

<i>if p then q . : not- p or q</i>	3h	If Nigel works in the service sector, then he is a corporate lawyer. Therefore, Nigel does not work in the service sector or he is a corporate lawyer.
	3l	If Nigel works in the service sector, then he is an organic food salesman. Therefore, Nigel does not work in the service sector or he is an organic food salesman.
<i>if not-p then q . : p or q</i>	4h	If Nigel does not drive a BMW, then he drives a Porsche. Therefore, Nigel drives a BMW or a Porsche.
	4l	If Nigel does not drive a BMW, then he drives a Ford Fiesta. Therefore, Nigel drives a BMW or a Ford Fiesta.
<i>p . : p or q</i>	5h	Nigel is a corporate lawyer. Therefore, Nigel is a corporate lawyer or a merchant banker.
	5l	Nigel is a corporate lawyer. Therefore, Nigel is a corporate lawyer or an organic food salesman.
<i>not-p . : not-p or q</i>	6h	Nigel does not drive a car. Therefore, Nigel does not drive a car or he drives a Porsche.
	6l	Nigel does not drive a car. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.
<i>q . : p or q</i>	7h	Nigel is a merchant banker. Therefore, Nigel is a corporate lawyer or a merchant banker.
	7l	Nigel is an organic food salesman. Therefore, Nigel is a corporate lawyer or an organic food salesman.
<i>q . : not-p or q</i>	8h	Nigel drives a Porsche. Therefore, Nigel does not drive a car or he drives a Porsche.
	8l	Nigel drives a Ford Fiesta. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Table B10. The 16 inferences for the inferences task, booklet 8.

Booklet	Inference	#	Content
Nigel D2	<i>p or q . : if not-p then q</i>	1h	Nigel is a corporate lawyer or a merchant banker. Therefore, if Nigel is not a corporate lawyer, then he is a merchant banker.
		1l	Nigel is a corporate lawyer or an organic food salesman. Therefore, if Nigel is not a corporate lawyer, then he

		is an organic food salesman.
$not-p \text{ or } q \therefore \text{if } p \text{ then } q$	2h	Nigel does not drive a car or he drives a Porsche. Therefore, if Nigel drives a car, then he drives a Porsche.
	2l	Nigel does not drive a car or he drives a Ford Fiesta. Therefore, if Nigel drives a car, then he drives a Ford Fiesta.
$\text{if } p \text{ then } q \therefore not-p \text{ or } q$	3h	If Nigel drives a car, then he drives a Porsche. Therefore, Nigel does not drive a car or he drives a Porsche.
	3l	If Nigel drives a car, then he drives a Ford Fiesta. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.
$\text{if } not-p \text{ then } q \therefore p \text{ or } q$	4h	If Nigel is not a corporate lawyer, then he is a merchant banker. Therefore, Nigel is a corporate lawyer or a merchant banker.
	4l	If Nigel is not a corporate lawyer, then he is an organic food salesman. Therefore, Nigel is a corporate lawyer or an organic food salesman.
$p \therefore p \text{ or } q$	5h	Nigel is a corporate lawyer. Therefore, Nigel is a corporate lawyer or a merchant banker.
	5l	Nigel is a corporate lawyer. Therefore, Nigel is a corporate lawyer or an organic food salesman.
$not-p \therefore not-p \text{ or } q$	6h	Nigel does not drive a car. Therefore, Nigel does not drive a car or he drives a Porsche.
	6l	Nigel does not drive a car. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.
$q \therefore p \text{ or } q$	7h	Nigel is a merchant banker. Therefore, Nigel is a corporate lawyer or a merchant banker.
	7l	Nigel is an organic food salesman. Therefore, Nigel is a corporate lawyer or an organic food salesman.
$q \therefore not-p \text{ or } q$	8h	Nigel drives a Porsche. Therefore, Nigel does not drive a car or he drives a Porsche.
	8l	Nigel drives a Ford Fiesta. Therefore, Nigel does not drive a car or he drives a Ford Fiesta.

Note. h = high plausibility for scenario, l = low plausibility for scenario.

Appendix C

The materials used in Experiments 3 and 4.

Table C1. Materials for inference 1: DM.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog does not have both three legs and blue fur The dog does not have three legs or the dog does not have blue fur	Mary walks through a random city and passes by a dog. How likely is it that The dog does not have both teeth and fur The dog does not have teeth or the dog does not have fur
Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant does not have both an ostrich farm and a butchery The restaurant does not have an ostrich farm or the restaurant does not have a butchery	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant does not have both chairs and tables The restaurant does not have chairs or the restaurant does not have tables
Lisa walks through a random city and passes by a family house. How likely is it that The house is not both made of plastic and built on a tree The house is not made of plastic or the house is not built on a tree	Lisa walks through a random city and passes by a family house. How likely is it that The house does not have both a roof and walls The house does not have a roof or the house does not have walls
Ben walks through a random city park, and passes by a tree. How likely is it that The tree does not both grow sideways and have bells in its branches The tree does not grow sideways or the tree does not have bells in its branches	Ben walks through a random city park, and passes by a tree. How likely is it that The tree does not have both a bark and branches The tree does not have a bark or the tree does not have branches
Emil walks through a random city and passes by a dove. How likely is it that The dove does not have both one leg and a green beak The dove does not have one leg or the dove does not have a green beak	Emil walks through a random city and passes by a dove. How likely is it that The dove does not have both feathers and a beak The dove does not have feathers or the dove does not have a beak
Berta walks through a random city and goes into a bar. How likely is it that The bar does not have both a book store and a laundry The bar does not have a book store or the bar does not have a laundry	Berta walks through a random city and goes into a bar. How likely is it that The bar does not have both beer and wine The bar does not have beer or the bar does not have wine
Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood does not have both a waterfall and a silver mine The neighbourhood does not have a waterfall or the neighbourhood does not have a silver	Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood does not have both people and cars The neighbourhood does not have people or the neighbourhood does not have cars

mine	
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus does not carry both horses and stones The bus does not carry horses or the bus does not carry stones	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus does not have both seats and windows The bus does not have seats or the bus does not have windows
Milo walks through a random city and passes by a cat. How likely is it that The cat does not both have three legs and have mud in its fur The cat does not have three legs or the cat does not have mud in its fur	Milo walks through a random city and passes by a cat. How likely is it that The cat does not have both claws and eyes The cat does not have claws or the cat does not have eyes
Laura walks through a random city and into a park. How likely is it that The park does not have both a cement floor and artificial plants The park does not have a cement floor or it does not have artificial plants	Laura walks through a random city and into a park. How likely is it that The park does not have both trees and grass The park does not have trees or the park does not have grass
Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle is not both made of bread and covered in sugar The bicycle is not made of bread or the bicycle is not covered in sugar	Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle does not have both wheels and pedals The bicycle does not have wheels or the bicycle does not have pedals
Ines walks through a random city and passes by a train station. How likely is it that The train station does not have both a theatre and a cinema The train station does not have a theatre or the train station does not have a cinema	Ines walks through a random city and passes by a train station. How likely is it that The train station does not have both seats and a time table The train station does not have seats or the train station does not have a time table

Table C2. Materials for inference 2: nDM.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog has teeth and it has fur The dog does not have teeth or the dog does not have fur	Mary walks through a random city and passes by a dog. How likely is it that The dog has three legs and it has blue fur The dog does not have three legs or the dog does not have blue fur
Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has chairs and it has tables The restaurant does not have chairs or the restaurant does not have tables	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has an ostrich farm and it has a butchery The restaurant does not have an ostrich farm or the restaurant does not have a butchery
Lisa walks through a random city and passes	Lisa walks through a random city and passes

<p>by a family house. How likely is it that The house has a roof and it has walls The house does not have a roof or the house does not have walls</p>	<p>by a family house. How likely is it that The house is made of plastic and it is built on a tree The house is not made of plastic or the house is not built on a tree</p>
<p>Ben walks through a random city park, and passes by a tree. How likely is it that The tree has a bark and it has branches The tree does not have a bark or the tree does not have branches</p>	<p>Ben walks through a random city park, and passes by a tree. How likely is it that The tree does grows sideways and it has bells in its branches The tree does not grow sideways or the tree does not have bells in its branches</p>
<p>Emil walks through a random city and passes by a dove. How likely is it that The dove has feathers and it has a beak The dove does not have feathers or the dove does not have a beak</p>	<p>Emil walks through a random city and passes by a dove. How likely is it that The dove has one leg and it has a green beak The dove does not have one leg or the dove does not have a green beak</p>
<p>Berta walks through a random city and goes into a bar. How likely is it that The bar has beer and it has wine The bar does not have beer or the bar does not have wine</p>	<p>Berta walks through a random city and goes into a bar. How likely is it that The bar has a book store and it has a laundry The bar does not have a book store or the bar does not have a laundry</p>
<p>Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood has people and it has cars The neighbourhood does not have people or the neighbourhood does not have cars</p>	<p>Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood does has a waterfall and it has a silver mine The neighbourhood does not have a waterfall or the neighbourhood does not have a silver mine</p>
<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus has seats and it has windows The bus does not have seats or the bus does not have windows</p>	<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus carries horses and it carries stones The bus does not carry horses or the bus does not carry stones</p>
<p>Milo walks through a random city and passes by a cat. How likely is it that The cat has claws and it has eyes The cat does not have claws or the cat does not have eyes</p>	<p>Milo walks through a random city and passes by a cat. How likely is it that The cat has three legs and it has mud in its fur The cat does not have three legs or the cat does not have mud in its fur</p>
<p>Laura walks through a random city and into a park. How likely is it that The park has trees and it has grass The park does not have trees or the park does not have grass</p>	<p>Laura walks through a random city and into a park. How likely is it that The park has a cement floor and it has artificial plants The park does not have a cement floor or it does not have artificial plants</p>
<p>Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle has wheels and it has pedals</p>	<p>Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle is made of bread and it is covered</p>

The bicycle does not have wheels or the bicycle does not have pedals	in sugar The bicycle is not made of bread or the bicycle is not covered in sugar
Ines walks through a random city and passes by a train station. How likely is it that The train station has seats and it has a time table The train station does not have seats or the train station does not have a time table	Ines walks through a random city and passes by a train station. How likely is it that The train station has a theatre and it has a cinema The train station does not have a theatre or the train station does not have a cinema

Table C3. Materials for inference 3: &E.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog has teeth and it has fur The dog has teeth	Mary walks through a random city and passes by a dog. How likely is it that The dog has three legs and it has blue fur The dog has three legs
Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has chairs and it has tables The restaurant has chairs	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has an ostrich farm and a butchery The restaurant has an ostrich farm
Lisa walks through a random city and passes by a family house. How likely is it that The house has a roof and it has walls The house has a roof	Lisa walks through a random city and passes by a family house. How likely is it that The house is made of plastic and it is built on a tree The house is made of plastic
Ben walks through a random city park, and passes by a tree. How likely is it that The tree has a bark and it has branches The tree has a bark	Ben walks through a random city park, and passes by a tree. How likely is it that The tree grows sideways and it has bells in its branches The tree grows sideways
Emil walks through a random city and passes by a dove. How likely is it that The dove has feathers and it has a beak The dove has feathers	Emil walks through a random city and passes by a dove. How likely is it that The dove has one leg and it has a green beak The dove has one leg
Berta walks through a random city and goes into a bar. How likely is it that the bar The bar has beer and it has wine The bar has beer	Berta walks through a random city and goes into a bar. How likely is it that the bar The bar has a book store and it has a laundry The bar has a book store
Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood has people and it has cars The neighbourhood has people	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood has a waterfall and it has a silver mine The neighbourhood has a waterfall
Anne walks through a random city, and	Anne walks through a random city, and

passes by a bus from the local public transport system. How likely is it that The bus has seats and it has windows The bus has seats	passes by a bus from the local public transport system. How likely is it that The bus carries horses and it carries stones The bus carries horses
Milo walks through a random city and passes by a cat. How likely is it that The cat has claws and it has eyes The cat has claws	Milo walks through a random city and passes by a cat. How likely is it that the cat The cat has three legs and it has mud in its fur The cat has three legs
Laura walks through a random city and into a park. How likely is it that The park has trees and it has grass The park has trees	Laura walks through a random city and into a park. How likely is it that The park has a cement floor and it has artificial plants The park has a cement floor
Michael walks through a random city and passes by a bicycle. How likely is it that the bicycle The bicycle has wheels and it has pedals The bicycle has wheels	Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle is made of bread and it is covered in sugar The bicycle is made of bread
Ines walks through a random city and passes by a train station. How likely is it that The train station has seats and it has a time table The train station has seats	Ines walks through a random city and passes by a train station. How likely is it that The train station has a theatre and it has a cinema The train station has a theatre

Table C4. Materials for inference 4: &I.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog has teeth The dog has teeth and it has fur	Mary walks through a random city and passes by a dog. How likely is it that The dog has three legs The dog has three legs and it has blue fur
Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has chairs The restaurant has chairs and it has tables	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has an ostrich farm The restaurant has an ostrich farm and it has a butchery
Lisa walks through a random city and passes by a family house. How likely is it that The house has a roof The house has a roof and it has walls	Lisa walks through a random city and passes by a family house. How likely is it that The house is made of plastic The house is made of plastic and it is built on a tree.
Ben walks through a random city park, and passes by a tree. How likely is it that The tree has a bark The tree has a bark and it has branches	Ben walks through a random city park, and passes by a tree. How likely is it that The tree grows sideways The tree grows sideways and it has bells in its branches
Emil walks through a random city and passes	Emil walks through a random city and passes

by a dove. How likely is it that The dove has feathers The dove has feathers and it has a beak	by a dove. How likely is it that The dove has one leg The dove has one leg and it has a green beak
Berta walks through a random city and goes into a bar. How likely is it that the bar The bar has beer The bar has beer and it has wine	Berta walks through a random city and goes into a bar. How likely is it that the bar The bar has a book store The bar has a book store and it has a laundry
Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood has people The neighbourhood has people and it has cars	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood has a waterfall The neighbourhood has a waterfall and it has a silver mine
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus has seats The bus has seats and it has windows	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus carries horses The bus carries horses and it carries stones
Milo walks through a random city and passes by a cat. How likely is it that The cat has claws The cat has claws and it has eyes	Milo walks through a random city and passes by a cat. How likely is it that the cat The cat has three legs The cat has three legs and it has mud in its fur
Laura walks through a random city and into a park. How likely is it that The park has trees The park has trees and it has grass	Laura walks through a random city and into a park. How likely is it that The park has a cement floor The park has a cement floor and it has artificial plants
Michael walks through a random city and passes by a bicycle. How likely is it that the bicycle The bicycle has wheels The bicycle has wheels and it has pedals	Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle is made of bread The bicycle is made of bread and it is covered in sugar
Ines walks through a random city and passes by a train station. How likely is it that The train station has seats The train station has seats and it has a time table	Ines walks through a random city and passes by a train station. How likely is it that The train station has a theatre The train station has a theatre and it has a cinema

Table C5. Materials for inference 5: &Or.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog has teeth and it has fur The dog has teeth or it has fur	Mary walks through a random city and passes by a dog. How likely is it that The dog has three legs and it has blue fur The dog has three legs or it has blue fur

Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has chairs and it has tables The restaurant has chairs or it has tables	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has an ostrich farm and it has a butchery The restaurant has an ostrich farm or it has a butchery
Lisa walks through a random city and passes by a family house. How likely is it that The house has a roof and it has walls The house has a roof or it has walls	Lisa walks through a random city and passes by a family house. How likely is it that The house is made of plastic and it is built on a tree The house is made of plastic or it is built on a tree
Ben walks through a random city park, and passes by a tree. How likely is it that The tree has a bark and it has branches The tree has a bark or it has branches	Ben walks through a random city park, and passes by a tree. How likely is it that The tree grows sideways and it has bells in its branches The tree grows sideways or it has bells in its branches
Emil walks through a random city and passes by a dove. How likely is it that The dove has feathers and it has a beak The dove has feathers or it has a beak	Emil walks through a random city and passes by a dove. How likely is it that The dove has one leg and it has a green beak The dove has one leg or it has a green beak
Berta walks through a random city and goes into a bar. How likely is it that The bar has beer and it has wine The bar has beer or it has wine	Berta walks through a random city and goes into a bar. How likely is it that the bar The bar has a book store and it has a laundry The bar has a book store or it has a laundry
Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood has people and it has cars The neighbourhood has people or it has cars	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood has a waterfall and it has a silver mine The neighbourhood has a waterfall or it has a silver mine
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus has seats and it has windows The bus has seats or it has windows	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus carries horses and it carries stones The bus carries horses or it carries stones
Milo walks through a random city and passes by a cat. How likely is it that The cat has claws and it has eyes The cat has claws or it has eyes	Milo walks through a random city and passes by a cat. How likely is it that the cat The cat has three legs and it has mud in its fur The cat has three legs or it has mud in its fur
Laura walks through a random city and into a park. How likely is it that The park has trees and it has grass The park has trees or it has grass	Laura walks through a random city and into a park. How likely is it that The park has a cement floor and it has artificial plants The park has a cement floor or it has artificial plants
Michael walks through a random city and	Michael walks through a random city and

<p>passes by a bicycle. How likely is it that</p> <p>The bicycle has wheels and it has pedals</p> <p>The bicycle has wheels or it has pedals</p>	<p>passes by a bicycle. How likely is it that</p> <p>The bicycle is made of bread and it is covered in sugar</p> <p>The bicycle is made of bread or it is covered in sugar</p>
<p>Ines walks through a random city and passes by a train station. How likely is it that</p> <p>The train station has seats and it has a time table</p> <p>The train station has seats or it has a time table</p>	<p>Ines walks through a random city and passes by a train station. How likely is it that</p> <p>The train station has a theatre and it has a cinema</p> <p>The train station has a theatre or it has a cinema</p>

Table C6. Materials for inference 6: Or&.

High condition	Low condition
<p>Mary walks through a random city and passes by a dog. How likely is it that</p> <p>The dog has teeth or it has fur</p> <p>The dog has teeth and it has fur</p>	<p>Mary walks through a random city and passes by a dog. How likely is it that</p> <p>The dog has three legs or it has blue fur</p> <p>The dog has three legs and it has blue fur</p>
<p>Daniel walks through a random city and enters a restaurant. How likely is it that</p> <p>The restaurant has chairs or it has tables</p> <p>The restaurant has chairs and it has tables</p>	<p>Daniel walks through a random city and enters a restaurant. How likely is it that</p> <p>The restaurant has an ostrich farm or it has a butchery</p> <p>The restaurant has an ostrich farm and it has a butchery</p>
<p>Lisa walks through a random city and passes by a family house. How likely is it that</p> <p>The house has a roof or it has walls</p> <p>The house has a roof and it has walls</p>	<p>Lisa walks through a random city and passes by a family house. How likely is it that</p> <p>The house is made of plastic or it is built on a tree</p> <p>The house is made of plastic and it is built on a tree</p>
<p>Ben walks through a random city park, and passes by a tree. How likely is it that</p> <p>The tree has a bark or it has branches</p> <p>The tree has a bark and it has branches</p>	<p>Ben walks through a random city park, and passes by a tree. How likely is it that</p> <p>The tree grows sideways or it has bells in its branches</p> <p>The tree grows sideways and it has bells in its branches</p>
<p>Emil walks through a random city and passes by a dove. How likely is it that</p> <p>The dove has feathers or it has a beak</p> <p>The dove has feathers and it has a beak</p>	<p>Emil walks through a random city and passes by a dove. How likely is it that</p> <p>The dove has one leg or it has a green beak</p> <p>The dove has one leg and it has a green beak</p>
<p>Berta walks through a random city and goes into a bar. How likely is it that</p> <p>The bar has beer or it has wine</p> <p>The bar has beer and it has wine</p>	<p>Berta walks through a random city and goes into a bar. How likely is it that</p> <p>The bar has a book store or it has a laundry</p> <p>The bar has a book store and it has a laundry</p>
<p>Brian walks through a random city and enters one of its neighbourhoods.</p>	<p>Brian walks through a random city and enters one of its neighbourhoods. How likely is it</p>

The neighbourhood has people or it has cars	that
The neighbourhood has people and it has cars	The neighbourhood has a waterfall or it has a silver mine
	The neighbourhood has a waterfall and it has a silver mine
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that
The bus has seats or it has windows	The bus carries horses or it carries stones
The bus has seats and it has windows	The bus carries horses and it carries stones
Milo walks through a random city and passes by a cat. How likely is it that	Milo walks through a random city and passes by a cat. How likely is it that the cat
The cat has claws or it has eyes	The cat has three legs or it has mud in its fur
The cat has claws and it has eyes	The cat has three legs and it has mud in its fur
Laura walks through a random city and into a park. How likely is it that	Laura walks through a random city and into a park. How likely is it that
The park has trees or it has grass	The park has a cement floor or it has artificial plants
The park has trees and it has grass	The park has a cement floor and it has artificial plants
Michael walks through a random city and passes by a bicycle. How likely is it that	Michael walks through a random city and passes by a bicycle. How likely is it that
The bicycle has wheels or it has pedals	The bicycle is made of bread or it is covered in sugar
The bicycle has wheels and it has pedals	The bicycle is made of bread and it is covered in sugar
Ines walks through a random city and passes by a train station. How likely is it that	Ines walks through a random city and passes by a train station. How likely is it that
The train station has seats or it has a time table	The train station has a theatre or it has a cinema
The train station has seats and it has a time table	The train station has a theatre and it has a cinema

Table C7. Materials for inference 7: IfOr.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that	Mary walks through a random city and passes by a dog. How likely is it that
If the dog does not have teeth, then it has fur	If the dog does not have three legs, then it has blue fur
The dog has teeth or it has fur	The dog has three legs or it has blue fur
Daniel walks through a random city and enters a restaurant. How likely is it that	Daniel walks through a random city and enters a restaurant. How likely is it that
If the restaurant does not have chairs, then it has tables	If the restaurant does not have an ostrich farm, then it has a butchery
The restaurant has chairs or it has tables	The restaurant has an ostrich farm or it has a butchery
Lisa walks through a random city and passes	Lisa walks through a random city and passes

by a family house. How likely is it that If the house does not have a roof, then it has walls The house has a roof or it has walls	by a family house. How likely is it that If the house is not made of plastic, then it is built on a tree The house is made of plastic or it is built on a tree
Ben walks through a random city park, and passes by a tree. How likely is it that If the tree does not have a bark, then it has branches The tree has a bark or it has branches	Ben walks through a random city park, and passes by a tree. How likely is it that If the tree does not grow sideways, then it has bells in its branches The tree grows sideways or it has bells in its branches
Emil walks through a random city and passes by a dove. How likely is it that If the dove does not have feathers, then it has a beak The dove has feathers or it has a beak	Emil walks through a random city and passes by a dove. How likely is it that If the dove does not have one leg, then it has a green beak The dove has one leg or it has a green beak
Berta walks through a random city and goes into a bar. How likely is it that If the bar does not have beer, then it has wine The bar has beer or it has wine	Berta walks through a random city and goes into a bar. How likely is it that If the bar does not have a book store, then it has a laundry The bar has a book store or it has a laundry
Brian walks through a random city and enters one of its neighbourhoods. If the neighbourhood does not have people, then it has cars The neighbourhood has people or it has cars	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that If the neighbourhood does not have a waterfall, then it has a silver mine The neighbourhood has a waterfall or it has a silver mine
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus does not have seats, then it has windows The bus has seats or it has windows	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus does not carry horses, then it carries stones The bus carries horses or it carries stones
Milo walks through a random city and passes by a cat. How likely is it that If the cat does not have claws, then it has eyes The cat has claws or it has eyes	Milo walks through a random city and passes by a cat. How likely is it that the cat If the cat does not have three legs, then it has mud in its fur The cat has three legs or it has mud in its fur
Laura walks through a random city and into a park. How likely is it that If the park does not have trees, then it has grass The park has trees or it has grass	Laura walks through a random city and into a park. How likely is it that If the park does not have a cement floor, then it has artificial plants The park has a cement floor or it has artificial plants
Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle does not have wheels, then it has pedals	Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle is not made of bread, then it is covered in sugar

The bicycle has wheels or it has pedals	The bicycle is made of bread or it is covered in sugar
Ines walks through a random city and passes by a train station. How likely is it that If the train station does not have seats, then it has a time table The train station has seats or it has a time table	Ines walks through a random city and passes by a train station. How likely is it that If the train station does not have a theatre, then it has a cinema The train station has a theatre or it has a cinema

Table C8. Materials for inference 8: OrIf.

High condition	Low condition
Mary walks through a random city and passes by a dog. How likely is it that The dog has teeth or it has fur If the dog does not have teeth, then it has fur	Mary walks through a random city and passes by a dog. How likely is it that The dog has three legs or it has blue fur If the dog does not have three legs, then it has blue fur
Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has chairs or it has tables If the restaurant does not have chairs, then it has tables	Daniel walks through a random city and enters a restaurant. How likely is it that The restaurant has an ostrich farm or it has a butchery if the restaurant does not have an ostrich farm, then it has a butchery
Lisa walks through a random city and passes by a random house. How likely is it that The house has a roof or it has walls If the house does not have a roof, then it has walls	Lisa walks through a random city and passes by a random house. How likely is it that The house is made of plastic or it is built on a tree If the house is not made of plastic, then it is built on a tree
Ben walks through a random city park, and passes by a tree. How likely is it that The tree has a bark or it has branches If the tree does not have a bark, then it has branches	Ben walks through a random city park, and passes by a tree. How likely is it that The tree grows sideways or it has bells in its branches If the tree does not grow sideways, then it has bells in its branches
Emil walks through a random city and passes by a dove. How likely is it that The dove has feathers or it has a beak If the dove does not have feathers, then it has a beak	Emil walks through a random city and passes by a dove. How likely is it that The dove has one leg or it has a green beak If the dove does not have one leg, then it has a green beak
Berta walks through a random city and goes into a bar. How likely is it that The bar has beer or it has wine If the bar does not have beer, then it has wine	Berta walks through a random city and goes into a bar. How likely is it that The bar has a book store or it has a laundry If the bar does not have a book store, then it has a laundry
Brian walks through a random city and enters one of its neighbourhoods. The neighbourhood has people or it has cars If the neighbourhood does not have people,	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that The neighbourhood has a waterfall or it has a

then it has cars	silver mine If the neighbourhood does not have a waterfall, then it has a silver mine
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus has seats or it has windows If the bus does not have seats, then it has windows	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that The bus carries horses or it carries stones If the bus does not carry horses, then it carries stones
Milo walks through a random city and passes by a cat. How likely is it that The cat has claws or it has eyes If the cat does not have claws, then it has eyes	Milo walks through a random city and passes by a cat. How likely is it that The cat has three legs or it has mud in its fur If the cat does not have three legs, then it has mud in its fur
Laura walks through a random city and into a park. How likely is it that The park has trees or it has grass If the park does not have trees, then it has grass	Laura walks through a random city and into a park. How likely is it that The park has a cement floor or it has artificial plants If the park does not have a cement floor, then it has artificial plants
Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle has wheels or it has pedals If the bicycle does not have wheels, then it has pedals	Michael walks through a random city and passes by a bicycle. How likely is it that The bicycle is made of bread or it is covered in sugar If the bicycle is not made of bread, then it is covered in sugar
Ines walks through a random city and passes by a train station. How likely is it that The train station has seats or it has a time table If the train station does not have seats, then it has a time table	Ines walks through a random city and passes by a train station. How likely is it that The train station has a theatre or it has a cinema If the train station does not have a theatre, then it has a cinema

Table C9. Materials for inference 9: MP.

High-high condition	High-low condition
Mary walks through a random city and passes by a dog. How likely is it that If the dog has teeth, then it can bite The dog has teeth The dog can bite	Mary walks through a random city and passes by a dog. How likely is it that If the dog falls into a river, then it gets wet The dog falls into a river The dog gets wet
Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has chairs, then it has tables The restaurant has chairs The restaurant has tables	Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has a butchery, then it serves meat. The restaurant has a butchery The restaurant serves meat
Lisa walks through a random city and passes by a family house. How likely is it that	Lisa walks through a random city and passes by a family house. How likely is it that

<p>If the house has a roof then it has walls The house has a roof The house has walls</p>	<p>If the house has a water mill, then it is next to a river. The house has a water mill The house is next to a river</p>
<p>Ben walks through a random city park, and passes by a tree. How likely is it that If the tree has a bark then it has branches The tree has a bark The tree has branches</p>	<p>Ben walks through a random city park, and passes by a tree. How likely is it that If the tree falls down, then it has broken branches The tree falls down The tree has broken branches</p>
<p>Emil walks through a random city and passes by a dove. How likely is it that If the dove can fly then it has feathers The dove can fly The dove has feathers</p>	<p>Emil walks through a random city and passes by a dove. How likely is it that If the dove flies into a bucket of brown paint, then it is brown The dove flies into a bucket of brown paint The dove is brown</p>
<p>Berta walks through a random city and goes into a bar. How likely is it that the bar If the bar has beer then it has wine The bar has beer The bar has wine</p>	<p>Berta walks through a random city and goes into a bar. How likely is it that the bar If the bar has a book store, then it has a place to read The bar has a book store The bar has a place to read</p>
<p>Brian walks through a random city and enters one of its neighbourhoods. If the neighbourhood has cars then it has people The neighbourhood has cars The neighbourhood has people</p>	<p>Brian walks through a random city and enters one of its neighbourhoods. How likely is it that If the neighbourhood has a waterfall, then it has a river The neighbourhood has a waterfall The neighbourhood has a river</p>
<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus has seats, then it has windows The bus has seats The bus has windows</p>	<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus carries a refrigerator, then it has a heavy load The bus carries a refrigerator The bus has a heavy load</p>
<p>Milo walks through a random city and passes by a cat. How likely is it that If the cat has claws, then it can scratch The cat has claws The cat can scratch</p>	<p>Milo walks through a random city and passes by a cat. How likely is it that the cat If the cat falls into the mud, then it is dirty The cat falls into the mud The cat is dirty ("the" was removed in statements task)</p>
<p>Laura walks through a random city and into a park. How likely is it that If the park has trees then it has grass The park has trees The park has grass</p>	<p>Laura walks through a random city and into a park. How likely is it that If the park has dolphins then it has a pond The park has dolphins The park has a pond</p>
<p>Michael walks through a random city and passes by a bicycle. How likely is it that the</p>	<p>Michael walks through a random city and passes by a bicycle. How likely is it that</p>

bicycle	If the bicycle is made of silver, then it is expensive
If the bicycle has wheels then it has pedals	The bicycle is made of silver
The bicycle has wheels	The bicycle is expensive
The bicycle has pedals	
Ines walks through a random city and passes by a train station. How likely is it that	Ines walks through a random city and passes by a train station. How likely is it that
If the train station has seats then it has a time table	If the train station is flooded, then it is closed
The train station has seats	The train station is flooded
The train station has a time table	The train station is closed

Table C10. Materials for inference 10: MT.

High-high condition	High-low condition
Mary walks through a random city and passes by a dog. How likely is it that	Mary walks through a random city and passes by a dog. How likely is it that
If the dog falls into a sack of flour, then it has flour on its fur	If the dog has teeth, then it can bite
The dog does not have flour on its fur	The dog cannot bite
The dog does not fall into a sack of flour	The dog does not have teeth
Daniel walks through a random city and enters a restaurant. How likely is it that	Daniel walks through a random city and enters a restaurant. How likely is it that
If the restaurant has an ostrich farm, then it serves ostrich.	If the restaurant has chairs, then it has tables
The restaurant does not serve ostrich	The restaurant does not have tables
The restaurant does not have an ostrich farm	The restaurant does not have chairs
Lisa walks through a random city and passes by a family house. How likely is it that	Lisa walks through a random city and passes by a family house. How likely is it that
If the house is made of glass, then it is transparent	If the house has a roof then it has walls
The house is not transparent	The house does not have walls
The house is not made of glass	The house does not have a roof
Ben walks through a random city park, and passes by a tree. How likely is it that	Ben walks through a random city park, and passes by a tree. How likely is it that
If the tree falls on an electricity cable, then the electricity cable is damaged.	If the tree has a bark, then it has branches
The electricity cable is not damaged.	The tree does not have branches
The tree does not fall on the electricity cable.	The tree does not have a bark
Emil walks through a random city and passes by a dove. How likely is it that	Emil walks through a random city and passes by a dove. How likely is it that
If the dove flies into a bucket of green paint, then the dove is green.	If the dove can fly, then it has feathers
The dove is not green.	The dove does not have feathers
The dove does not fly into a bucket of green paint.	The dove cannot fly
Berta walks through a random city and goes into a bar. How likely is it that	Berta walks through a random city and goes into a bar. How likely is it that
If the bar has a library, then it has a silent area.	If the bar has beer then it has wine
	The bar does not have wine

The bar does not have a silent area. The bar does not have a library	The bar does not have beer
Brian walks through a random city and enters one of its neighbourhoods. If the neighbourhood is flooded, then the basements of the houses are under water. The basements of the houses are not under water. The neighbourhood is not flooded.	Brian walks through a random city and enters one of its neighbourhoods. How likely is it that If the neighbourhood has cars then it has people The neighbourhood does not have people The neighbourhood does not have cars
Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus is made of chocolate, then it tastes sweet The bus does not taste sweet The bus is not made of chocolate	Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus has seats, then it has windows The bus does not have windows The bus does not have seats
Milo walks through a random city and passes by a cat. How likely is it that If the cat bites a bear then it has bear hairs in its mouth The cat does not have bear hairs in its mouth The cat does not bite a bear	Milo walks through a random city and passes by a cat. How likely is it that If the cat has claws, then it can scratch The cat cannot scratch The cat does not have claws
Laura walks through a random city and into a park. How likely is it that If the park has zebras, then it has zebra footprints The park does not have zebra footprints The park does not have zebras	Laura walks through a random city and into a park. How likely is it that If the park has trees then it has grass The park does not have grass The park does not have trees
Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle is made of bread, then it can be eaten The bicycle cannot be eaten The bicycle is not made of bread	Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle has wheels then it has pedals The bicycle does not have pedals The bicycle does not have wheels
Ines walks through a random city and passes by a train station. How likely is it that If the train station has a boat factory, then it has pieces of boats. The train station does not have pieces of boats. The train station does not have a boat factory.	Ines walks through a random city and passes by a train station. How likely is it that If the train station has seats then it has a time table The train station does not have a time table The train station does not have seats

Table C11. Materials for inference 11: AC.

High-high condition	High-low condition
Mary walks through a random city and passes by a dog. How likely is it that	Mary walks through a random city and passes by a dog. How likely is it that

<p>If the dog has teeth, then it can bite The dog can bite The dog has teeth</p>	<p>If the dog falls into a sack of flour, then it has flour on its fur The dog has flour on its fur The dog falls into a sack of flour</p>
<p>Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has chairs, then it has tables The restaurant has tables The restaurant has chairs</p>	<p>Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has an ostrich farm, then it serves ostrich The restaurant serves ostrich The restaurant has an ostrich farm</p>
<p>Lisa walks through a random city and passes by a family house. How likely is it that If the house has a roof then it has walls The house has walls The house has a roof</p>	<p>Lisa walks through a random city and passes by a family house. How likely is it that If the house is made of glass, then it is transparent The house is transparent The house is made of glass</p>
<p>Ben walks through a random city park, and passes by a tree. How likely is it that If the tree has a bark, then it has branches The tree has branches The tree has a bark</p>	<p>Ben walks through a random city park, and passes by a tree. How likely is it that If the tree falls on an electricity cable, then the electricity cable is damaged The electricity cable is damaged The tree falls on the electricity cable</p>
<p>Emil walks through a random city and passes by a dove. How likely is it that If the dove can fly then it has feathers The dove has feathers The dove can fly</p>	<p>Emil walks through a random city and passes by a dove. How likely is it that If a dove flies into a bucket of green paint, then the dove is green The dove is green The dove flies into a bucket of green paint</p>
<p>Berta walks through a random city and goes into a bar. How likely is it that If the bar has beer then it has wine The bar has wine The bar has beer</p>	<p>Berta walks through a random city and goes into a bar. How likely is it that If the bar has a library, then it has a silent area The bar has a silent area The bar has a library</p>
<p>Brian walks through a random city and enters one of its neighbourhoods. If the neighbourhood has cars then it has people The neighbourhood has people The neighbourhood has cars</p>	<p>Brian walks through a random city and enters one of its neighbourhoods. How likely is it that If the neighbourhood is flooded, then the basements of the houses are under water The basements of the houses are under water The neighbourhood is flooded</p>
<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus has seats, then it has windows The bus has windows The bus has seats</p>	<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus is made of chocolate, then it tastes sweet The bus tastes sweet The bus is made of chocolate</p>
<p>Milo walks through a random city and passes by a cat. How likely is it that</p>	<p>Milo walks through a random city and passes by a cat. How likely is it that</p>

If the cat has claws, then it can scratch The cat can scratch The cat has claws	If the cat bites a bear then it has bear hairs in its mouth The cat has bear hairs in its mouth The cat bites a bear
Laura walks through a random city and into a park. How likely is it that If the park has trees then it has grass The park has grass The park has trees	Laura walks through a random city and into a park. How likely is it that If the park has zebras, then it has zebra footprints The park has zebra footprints The park has zebras
Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle has wheels then it has pedals The bicycle has pedals The bicycle has wheels	Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle is made of bread, then it can be eaten The bicycle can be eaten The bicycle is made of bread
Ines walks through a random city and passes by a train station. How likely is it that If the train station has seats then it has a time table The train station has a time table The train station has seats	Ines walks through a random city and passes by a train station. How likely is it that If the train station has a boat factory, then it has pieces of boats The train station has pieces of boats The train station has a boat factory

Table C12. Materials for inference 12: DA.

High-high condition	High-low condition
Mary walks through a random city and passes by a dog. How likely is it that If the dog falls into a river, then it gets wet The dog does not fall into a river The dog does not get wet	Mary walks through a random city and passes by a dog. How likely is it that If the dog has teeth, then it can bite The dog does not have teeth The dog cannot bite
Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has a butchery, then it serves meat. The restaurant does not have a butchery The restaurant does not serve meat	Daniel walks through a random city and enters a restaurant. How likely is it that If the restaurant has chairs, then it has tables The restaurant does not have chairs The restaurant does not have tables
Lisa walks through a random city and passes by a family house. How likely is it that If the house has a water mill, then it is next to a river. The house does not have a water mill The house is not next to a river	Lisa walks through a random city and passes by a family house. How likely is it that If the house has a roof, then it has walls The house does not have a roof The house does not have walls
Ben walks through a random city park, and passes by a tree. How likely is it that If the tree falls down, then it has broken branches The tree does not fall down The tree does not have broken branches	Ben walks through a random city park, and passes by a tree. How likely is it that If the tree has a bark then it has branches The tree does not have a bark The tree does not have branches

<p>Emil walks through a random city and passes by a dove. How likely is it that If the dove flies into a bucket of brown paint, then it is brown The dove does not fly into a bucket of brown paint The dove is not brown</p>	<p>Emil walks through a random city and passes by a dove. How likely is it that If the dove can fly then it has feathers The dove cannot fly The dove does not have feathers</p>
<p>Berta walks through a random city and goes into a bar. How likely is it that the bar If the bar has a book store, then it has a place to read The bar does not have a book store The bar does not have a place to read</p>	<p>Berta walks through a random city and goes into a bar. How likely is it that the bar If the bar has beer then it has wine The bar does not have beer The bar does not have wine</p>
<p>Brian walks through a random city and enters one of its neighbourhoods. If the neighbourhood has a waterfall, then it has a river The neighbourhood does not have a waterfall The neighbourhood does not have a river</p>	<p>Brian walks through a random city and enters one of its neighbourhoods. How likely is it that If the neighbourhood has cars then it has people The neighbourhood does not have cars The neighbourhood does not have people</p>
<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus carries a refrigerator, then it has a heavy load The bus does not carry a refrigerator The bus does not have a heavy load</p>	<p>Anne walks through a random city, and passes by a bus from the local public transport system. How likely is it that If the bus has seats, then it has windows The bus does not have seats The bus does not have windows</p>
<p>Milo walks through a random city and passes by a cat. How likely is it that If the cat falls into the mud, then it is dirty The cat does not fall into the mud The cat is not dirty</p>	<p>Milo walks through a random city and passes by a cat. How likely is it that the cat If the cat has claws, then it can scratch. The cat does not have claws The cat cannot scratch</p>
<p>Laura walks through a random city and into a park. How likely is it that If the park has dolphins then it has a pond The park does not have dolphins The park does not have a pond</p>	<p>Laura walks through a random city and into a park. How likely is it that If the park has trees then it has grass The park does not have trees The park does not have grass</p>
<p>Michael walks through a random city and passes by a bicycle. How likely is it that the bicycle If the bicycle is made of silver, then it is expensive The bicycle is not made of silver The bicycle is not expensive</p>	<p>Michael walks through a random city and passes by a bicycle. How likely is it that If the bicycle has wheels then it has pedals The bicycle does not have wheels The bicycle does not have pedals</p>
<p>Ines walks through a random city and passes by a train station. How likely is it that If the train station is flooded, then it is closed The train station is not flooded The train station is not closed</p>	<p>Ines walks through a random city and passes by a train station. How likely is it that If the train station has seats then it has a time table The train station does not have seats</p>

The train station does not have a time table

Appendix D

The materials used in Experiment 5.

Table D1. The six context stories used in Experiment 5, together with a sample inference for each.

Context story	Sample inference
<p>1 Imagine you are part of a team of researchers investigating the bird species on the island of Liaku. You are preparing a report on your current research, and the research quality of the report will determine whether you will be able to have enough funding to continue your work. It is important that you analyse the report in detail so that you can be sure that no claims are made that are not warranted by the data. You are reviewing the report with Lisa, another researcher in the team.</p>	<p>Premise: The next Dayo bird she sees on Liaku will eat yam fish and auk fish.</p> <p>Conclusion: The next Dayo bird she sees on Liaku will eat yam fish.</p>
<p>2 Imagine you are part of a team of epidemiologists that is working to find the source of a current cholera epidemic in the Taoli region. You are sampling the water from a number of wells to see whether they carry the virus or are safe for drinking. You need to analyse the water samples very thoroughly in order to give to the residents the information they need to regain control over the situation. You are reviewing the latest data with Paul, another epidemiologist in the team.</p>	<p>Premise: Well A and well B have drinkable water.</p> <p>Conclusion: Well A has drinkable water.</p>
<p>3 Imagine you are part of a team of astronomers that is coordinating a mission to Mars. You sent a computer robot to the planet, and are controlling its movements from earth, to gather important information. You need to be very careful in how you move it because if it falls over, it will not be able to send further data to earth. You are discussing its next movements with Emma, another astronomer in the team.</p>	<p>Premise: It's safe for the robot to move forward and to turn left.</p> <p>Conclusion: It's safe for the robot to move forward.</p>
<p>4 Imagine you are part of a team of engineers that is in charge of the security control of the huge water dam of Arom. One night during your shift, the security alarm suddenly goes on, indicating that there are cracks in the dam wall. This can be potentially very dangerous and you need to be careful in your evaluation of the situation. You go out with your colleague Brian to try to find where the cracks are located.</p>	<p>Premise: The upper north area and the lower north area of the dam are sealed.</p> <p>Conclusion: The upper north area of the dam is sealed.</p>
<p>5 Imagine you are part of a team of doctors who are working in an emergency hospital. Several patients are brought in by the ambulance with a variety of severe injuries. It is important that you act carefully on these cases because a wrong diagnosis could be fatal. You are reviewing their files with Miriam, another doctor in the team.</p>	<p>Premise: Patient H. D. has a liver injury and a kidney injury.</p> <p>Conclusion: Patient H. D. has a liver injury.</p>
<p>6 Imagine you are part of a group of detectives investigating a murder on a member of parliament. You are preparing a press</p>	<p>Premise: The post officer and the driver</p>

conference on the latest stage of your findings and analyses. It is important that you avoid criminalising any individual before you have gathered enough evidence to do so. You are reviewing the points you will make at the conference with Simon, another detective in the team.

are implicated in the crime.

Conclusion: The post officer is implicated in the crime.

Appendix E

The materials used in Experiment 6.

Table E1. The 12 context stories used in Experiment 6, together with a sample inference for each.

Context story	Sample inference
<p>1 Imagine you are part of a team of researchers investigating the bird species on the island of Liaku. You are preparing a report on your current research, and the research quality of the report will determine whether you will be able to have enough funding to continue your work. It is important that you analyse the report in detail so that you can be sure that no claims are made that are not warranted by the data. You are reviewing the report with the team.</p>	<p>Premise: The next Dayo bird you see on Liaku will not eat both yam fish and auk fish.</p> <p>Conclusion: The next Dayo bird you see on Liaku will not eat yam fish or will not eat auk fish.</p>
<p>2 Imagine you are part of a group of detectives investigating a murder on a member of parliament. You are preparing a press conference on the latest stage of your findings and analyses. Your statements in this conference will have a strong impact on the perception that the court and the public will have of the case, and it is important that you avoid criminalising any individual before you have gathered enough evidence to do so. You are reviewing the points you will make at the conference with the team.</p>	<p>Premise: The post officer and the driver are not both implicated in the crime.</p> <p>Conclusion: The post officer is not implicated in the crime or the driver is not implicated in the crime.</p>
<p>3 Imagine you are part of a team of doctors who are working in an emergency hospital. Several patients are brought in by the ambulance with a variety of severe injuries. It is important that you act carefully on these cases because a wrong diagnosis could be fatal. You are reviewing their files with the team.</p>	<p>Premise: Patient H. D. does not have both a liver injury and a kidney injury.</p> <p>Conclusion: Patient H. D. does not have a liver injury or he does not have a kidney injury.</p>
<p>4 Imagine you are part of a team of engineers that is in charge of the security control of the huge water dam of Arom. One night during your shift, the security alarm suddenly goes on, indicating that there are cracks in the dam wall. This can be potentially very dangerous and you need to be careful in your evaluation of the situation. You go out with the team to try to find where the cracks are located.</p>	<p>Premise: The upper north area and the lower north area of the dam are not both sealed.</p> <p>Conclusion: The upper north area of the dam is not sealed or the lower north area of the dam is not sealed.</p>
<p>5 Imagine you are part of a team of astronomers that is coordinating a mission to Mars. You sent a computer robot to the planet, and are controlling its movements from earth, to</p>	<p>Premise: It's not safe for the robot to both move forward and to turn left.</p>

	gather important information. You need to be very careful in how you move it because if it falls over, it will not be able to send further data to earth. You are discussing its next movements with the team.	Conclusion: It's not safe for the robot to move forward or it's not safe for the robot to turn left.
6	Imagine you are part of a team of epidemiologists that is working to find the source of a current cholera epidemic in the Taoli region. You are sampling the water from a number of wells to see whether they carry the virus or are safe for drinking. You need to analyse the water samples very thoroughly in order to give to the residents the information they need to regain control over the situation. You are reviewing the latest data with the team.	Premise: Well A and well B do not both have drinkable water. Conclusion: Well A does not have drinkable water or well B does not have drinkable water.
7	Imagine you are part of a team of engineers who have been called to the city of Barku after a severe earthquake. Your task is to go through the buildings that are still standing, to check which of them are stable enough for the people to go back in, and which are too dangerous for this. You have to examine each building carefully to minimise the risks for the population. You are reviewing a row of houses with the team.	Premise: The green house and the blue house are not both stable. Conclusion: The green house is not stable or the blue house is not stable.
8	Imagine you are part of a team of aid workers who are removing the mines from the Dunlar fields, where a war took place recently. You have to act very thoroughly in order to make sure the area is cleared and safe again for the residents. You are reviewing the latest data on the fields with the team.	Premise: The oat field and the barley field are not both cleared. Conclusion: The oat field is not cleared or the barley field is not cleared.
9	Imagine you are part of a team of computer scientists working at a communications company. You recently discovered a dangerous virus in your system, that threatens to steal the data of thousands of users. You need to analyse carefully what areas are affected to avoid this happening. You are reviewing the situation with the team.	Premise: Drive A and drive B are not both affected. Conclusion: Drive A is not affected or drive B is not affected.
10	Imagine you are part of a team of police officers who is trying to capture a group of armed robbers. The robbers entered an old factory building, but you are not sure where exactly they are. You need to analyse the information you have carefully to minimise the risks when entering the building. You are reviewing the situation with the team.	Premise: The first robber did not both go up the stairs and enter the storage room. Conclusion: The first robber did not go up the stairs or the first robber did not enter the storage room.
11	Imagine you are part of a team of cooks at a prestigious restaurant. Some of the clients just got sent to hospital with food poisoning. You have to find out which of the dishes are	Premise: The potatoes and the rice are not both toxic.

<p>toxic and which safe, in order to identify the cause and prevent further intoxications. You are reviewing the situation with the team.</p>	<p>Conclusion: The potatoes are not toxic or the rice is not toxic.</p>
<p>12 Imagine you are part of a team of technicians surveying the underground system of the city of Limro. A severe flood has started to affect part of the underground. You need to analyse carefully which areas are flooded so that you can evacuate the passengers and avoid the flood spreading further. You are reviewing the situation with the team.</p>	<p>Premise: The red line and the yellow line are not both flooded.</p> <p>Conclusion: The red line is not flooded or the yellow line is not flooded.</p>

Appendix F.

The materials used in Experiment 9.

Table F1. The scenarios used in Experiment 9, shown for the sample condition, a set size of 10, and a proportion of .9.

Topic	Content
1 National park	<p>The administrators of Leodi national park admit different numbers of people each day based on conservation criteria, and all those who wish to get in must register their interest beforehand.</p> <p>A botanist took random samples of the number of applicants on a series of days, and recorded how many of them were admitted.</p> <p>On day one the botanist took a sample of 10 applicants, and observed that 9 of them were admitted.</p>
2 Universities	<p>The universities in Irmau state determine the number of persons accepted each year based on the staff available for teaching.</p> <p>A researcher took random samples of the number of people who applied for mathematics, and how many of them were accepted, at some of the universities in the state.</p> <p>At university one the researcher took a sample of 10 applications for mathematics, and observed that 9 of them were accepted.</p>
3 Tree disease	<p>A tree disease has spread to the orange plantations of the farmers of Orisau.</p> <p>An agronomist went to some of the fields and took random samples of trees on each field, to record the number of affected trees among them.</p> <p>On Field %one% the agronomist took a sample of 10 trees, and observed that 9 of them were affected.</p>
4 Concert	<p>An event manager of a concert hall wants to know how often the people who go to the concerts drink beer.</p> <p>To this end, the event manager took random samples of people attending a series of concerts, and recorded how many of them drank beer.</p> <p>At Concert one the event manager took a sample of 10 people, and observed that 9 of them drank beer.</p>
5 Apple bags	<p>A cooperative of apple farmers offers a certain number of its apples for sale each season at the local market.</p> <p>The coordinator of the cooperative took random samples of the apple bags offered for sale at the local market by the farmers, and recorded how many of these apple bags were sold.</p> <p>From farmer one the coordinator took a sample of 10 bags, and observed that 9 of them were sold.</p>
6 River crossing	<p>In order to get to Lindau, travellers must cross the Duni River. But the number of travellers that can cross on any one day depends on the number of boats available.</p> <p>One of the boat operators took a random sample of the people who wanted to cross on a series of days, and recorded how many of these were able to do so.</p> <p>On day one the boat operator took a sample of 10 travellers who wanted to cross the river, and observed that 9 of them were able to do so.</p>
7 Sun screen	<p>A dermatologist wants to find out how often the people on a beach use sun</p>

	<p>screen.</p> <p>To this end the dermatologist took random samples of the people on the beach on a series of sunny days, and recorded how many of them used sun screen.</p> <p>On day one the dermatologist took a sample of 10 people, and observed that 9 of them used sun screen.</p>
8 Train luggage	<p>A train designer is investigating how many of the passengers travelling on the train between Arbei and Lindel carry luggage.</p> <p>To this end, the designer took random samples of the train passengers travelling between Arbei and Lindel on a series of days, and recorded how many of them carried luggage.</p> <p>On day one the designer took a sample of 10 passengers, and observed that 9 of them carried luggage.</p>
9 Coffee with sugar	<p>The waiter of a cafe wants to find out how often the clients put sugar in their coffee.</p> <p>To this end the waiter took random samples of the clients ordering a coffee, and recorded the number of times they put sugar in their coffee, on a series of days.</p> <p>On day one the waiter took a sample of 10 clients who ordered a coffee, and observed that 9 of them put sugar in it.</p>

Appendix G

The materials used in Experiment 10.

Table F1. The scenarios in Experiment 10 together with a sample premise for each.

Name	Scenario	Sample premise
Bird	Imagine you are part of a team of researchers investigating the bird species on the island of Liaku. You have gathered a range of data on the birds and are now starting to analyse it. You need to be careful to make sure that any conclusions you draw are warranted by the data. You are reviewing the findings with the team.	If the next Dayo bird you see on Liaku eats yam seeds, then it will eat auk seeds.
Murder	Imagine you are part of a group of detectives investigating a murder on a member of parliament. You have brought together the information you have on the case and are trying to obtain new insights from it. You need to be careful to make sure you only draw conclusions that are justified by the evidence. You are discussing the situation with the team.	If the post officer is implicated in the crime, then the driver is implicated in the crime.
Injury	Imagine you are part of a team of doctors who are working in an emergency hospital. Several patients are brought in by the ambulance with a variety of severe injuries. It is important that you act carefully on these cases because a wrong diagnosis could be fatal. You are reviewing their files with the team.	If patient H. D. has a liver injury, then he has a kidney injury.
Tube	Imagine you are part of a team of technicians surveying the underground system of the city of Limro. A severe flood has started to affect part of the underground. You need to analyse carefully which areas are flooded so that you can evacuate the passengers and avoid the flood spreading further. You are reviewing the situation with the team.	If the red line is flooded, then the green line is flooded.
Robot	Imagine you are part of a team of astronomers that is coordinating a mission to Mars. You sent a computer robot to the planet, and are controlling its movements from earth, to gather important information. You need to be very careful in how you move it because if it falls over, it will not be able to send further data to earth. You are discussing its next movements with the team.	If it's safe for the robot to move forward, then it's safe for the robot to rise up.
Water	Imagine you are part of a team of epidemiologists that is working to find the source of a current cholera epidemic in the Taoli region. You are sampling the water from a number of wells in this region to check whether they carry the virus or are safe for drinking. You need to analyse the water samples very thoroughly in order to give to the residents the information they need to regain control over the situation. You are reviewing the latest data with the team.	If the well in Laka district has drinkable water, then the well in Yerlo district has drinkable water.
Quake	Imagine you are part of a team of engineers who have been called to the city of Barku after a severe earthquake. Your task is to go through the buildings that are still standing, to check which of them are stable enough for the people to go	If the green house is stable, then the blue house is stable.

	back in, and which are too dangerous for this. You have to examine each building carefully to minimise the risk for the population. You are reviewing a row of houses with the team.	
Field	Imagine you are part of a team of aid workers who are removing the mines from the Dunlar fields, where a war took place recently. You have to act very thoroughly in order to make sure the area is cleared and safe again for the residents. You are reviewing the latest data on the fields with the team.	If the oat field is cleared, then the barley field is cleared.
System	Imagine you are part of a team of computer scientists working at a communications company. You recently discovered a dangerous virus in your system, which threatens to steal the data of thousands of users. You need to analyse carefully what areas are affected to avoid this happening. You are reviewing the situation with the team.	If drive A is affected, then drive B is affected.
Armed	Imagine you are part of a team of police officers who is trying to capture an armed robber. The robber entered an old factory building. You need to analyse carefully the information you have on its location to minimise the risk when entering the building. You are reviewing the situation with the team.	If the robber went up the stairs, then he entered the storage room.