Truth Table Task: Working Memory Load, Latencies and Perceived Relevance

Aline Sevenants (Aline.Sevenants@ppw.kuleuven.be)

Kristien Dieussaert (Kristien.Dieussaert@ppw.kuleuven.be)

Walter Schaeken (Walter.Schaeken@ppw.kuleuven.be)

University of Leuven

Keywords: conditionals, truth table task, working memory load, classification times, suppositional theory, mental models accounts

Corresponding author:

Aline Sevenants

Aline.Sevenants@ppw.kuleuven.be

University of Leuven, Laboratory of Experimental Psychology

102 Tiensestraat, B-3000 Leuven, Belgium

Tel: +32 (0)16/32.59.52

Fax: +32 (0)16/32.60.99

Abstract

The aim of the present study is to uncover the relation between cognitive ability and the answer patterns yielded by the truth table task. According to the Classical Mental Models Theory, people with high working memory capacity answer according to two-valued or 'logical' answer patterns. The Suppositional Theory and the Revised Mental Models Theory predict that the answer patterns given by the most intelligent ones are three-valued or 'defective'. Correlations are examined, and in three experiments it is tested with a dual task paradigm whether a differential working memory load alters participants' answer patterns. A positive correlation is observed between cognitive ability and three-valued answer patterns, but no effect of the working memory load manipulation is revealed. With an inspection of the classification times we shed light on the processes underlying truth table judgments. We conclude that the Revised Mental Models Theory provides the best account for our results.

Introduction

The meaning of ‘If’

The word “if” is everywhere in our daily language and in all thinkable contexts: promises, threats, commitments, etc. It enables us to engage in hypothetical thinking, to reflect on the implications of what may come, could have been or will be. We are very used to conditionals in our daily communication and seem to have no problem at all understanding them. When a little boy is told by his mother at the doctor’s examination table: “If you’re quiet now, you’ll get an ice cream when it’s over”, even at young age he is able to grasp the meaning of this. Mothers tend to keep their promises, so being quiet now leads to ice cream later. But what happens when the boy starts screaming? Is there a chance that he might get an ice cream anyway? Since conditionals are extremely widely used in our communication and make up a key factor of our reasoning and decision making, it’s very important to know exactly how people understand and interpret them.

Conditionals are sentences with the following grammatical construction: ‘If A then C’, with A being the antecedent and C the consequent of the clause. E.g.: *If the letter is a T, then the number is a 5*. Combining the truth-value of the antecedent and the consequent, four possibilities or cases arise:

True-antecedent – True-consequent (TT): *a T* and *a 5*

True-antecedent – False-consequent (TF): *a T* and *a 3*

False-antecedent – True-consequent (FT): *a B* and *a 5*

False-antecedent – False-consequent (FF): *a B* and *a 3*

The TT case and the TF case (1 and 2) are referred to as the true-antecedent cases and the FT and FF case (3 and 4) as the false-antecedent cases (FA cases in the remainder of the text).

In the experiments described in the present manuscript we will only make use of so-called basic conditionals, comparable to the one in the example above. Such conditionals concern abstract relationships and are indicative in form, making claims that could apparently be empirically verified and in which content effects and the influence of background knowledge are controlled for.

The Mental Models Theories and the Suppositional Theory

Truth Table Task Judgments

There has been substantial debate in reasoning literature concerning the processes and representations underlying people's understanding of conditional assertions. Two main theories accounting for the mental representation of conditionals are the Classical Mental Models Theory (C-MMT; Johnson-Laird & Byrne, 1991; 2002; Johnson-Laird, Byrne, & Schaeken, 1992) and the Suppositional Theory (ST; Evans & Over, 2004; Evans, Handley, Neilens, & Over, 2007; 2008; Evans, Over & Handley, 2003). The ST and the C-MMT make different predictions about the mental representation of conditionals.

According to the *C-MMT*, people reason with representations resembling the two-valued truth tables from standard logic, judging the TT case, the FT case and the FF case to make the conditional 'true'. The only case that falsifies the conditional is TF. According to the *ST*, people reason with representations matching with three-valued defective truth tables or de Finetti-tables¹ (De Finetti, 1967, 2008; Politzer, Over & Baratgin, 2010), in which the TT case is classified as making the conditional 'true', the TF case as falsifying it and the FA cases FT and FF are judged to be 'irrelevant' with respect to the truth of the rule.

¹ De Finetti (1967, 2008) also departs from standard two-valued logic introducing a third value for the FA cases in conditional events: "null". This truth table, with the FA cases being void, is the so-called 'de Finetti Table' and is comparable to the defective or three-valued truth table (see also Politzer, Over & Baratgin, 2010). In the remainder of this paper the terminology two-valued vs. three-valued will be used.

The most commonly used psychological tasks to study such conditionals are Wason's selection task (1966) and the conditional inference task. These tasks, however, do not enable researchers to examine specifically the representation of the FA cases about which the ST and the C-MMT have different predictions. Hence, we will make use of the truth table task in which participants are asked to indicate for each of the four cases whether the combination makes the given rule true, false, or is irrelevant with respect to the truth of the rule.

According to the *C-MMT* (Barrouillet & Lecas, 1998; Johnson-Laird, 2006; Johnson-Laird & Byrne, 1991, 2002; Johnson-Laird, Byrne, & Schaeken, 1992), people have a natural tendency to think of possibilities. Hence, when they draw a conclusion from a set of premises, they think of the possibilities compatible with the premises: they build the models of one premise, and then update them according to the next premise, and so on. Each mental model represents a possibility in as iconic a way as possible (that is, a mental model has a structure that corresponds to the known structure of what it represents, Johnson-Laird & Byrne, 2002). So when people have to classify the cases in a truth table task, they list all logical possibilities compatible with the conditional rule. In line with the 'principle of truth' ("each mental model of a set of assertions represents a possibility given the truth of the assertions and each mental model represents a clause in these assertions only when it is true in that possibility", p. 652), reasoners construct mental models of the possibilities compatible with the premises, but they initially and by default do not represent what is false. Therefore, their conclusion is based on the initial model, the one with a true antecedent (A) and a true consequent (C):

A C

...

The other possibilities remain implicit at that time and are called 'mental footnotes' about the falsity of clauses (represented by the ellipsis). If the reasoners are able to retain these

footnotes, they can flesh out the implicitly represented information into fully explicit models, representing clauses even when they have false antecedents.

$$\begin{array}{l} A \quad C \\ \neg A \quad C \\ \neg A \quad \neg C \end{array}$$

These three possibilities correspond to the rows of the truth table in which the conditional is deemed ‘true’, including the FA cases FT and FF. This in turn leads to a two-valued truth table pattern.

The *ST* (Evans & Over, 2004; Evans, Handley, Neilens & Over, 2007; 2008;) provides an account of the three-valued truth tables yielded by the truth task at the semantic level. According to the ST, people evaluate conditionals by means of the Ramsey-Test. That is, they “hypothetically add p to their stock of knowledge and evaluate their degree of belief in q given p ” (Ramsey, 1931/1990 p.247). The ST is part of a Bayesian or probabilistic approach to the study of reasoning, as reflected in its dependence on Ramsey and de Finetti (see Gilio & Over, 2012). Probability is then understood as subjective probability, as a degree of belief (Gilio & Over, 2012; Oaksford & Chater, 2007). The Ramsey test states that the probability of a natural language indicative conditional, $P(\text{if } A \text{ then } C)$, is evaluated by judging how likely it is that C will hold given A , making $P(\text{if } A \text{ then } C)$ equal to the conditional probability of C given A , $P(C|A)$. In the de Finetti table, “if A then C ” is classed as ‘true’ when $A \ \& \ C$ (the TT case) holds, ‘false’ when $A \ \& \ \text{not-}C$ (the TF case) holds, and neither true nor false when not- A holds (the FA cases), i.e. not- A is ‘irrelevant’ to the truth or falsity of “if A then C ”. This three-valued table fits perfectly with the Ramsey test, which makes not- A cases ‘irrelevant’ to the assessment of $P(\text{if } A \text{ then } C) = P(C|A)$.

The assertion of an indicative conditional in this Bayesian/probabilistic theory can be related to a conditional bet on “if A then C ”, which is won when $A \ \& \ C$ holds, lost when $A \ \&$

not-C holds, and neither won nor lost when not-A holds. The conditional bet is void when not-A holds and no money changes hands. In other words, not-A is 'irrelevant' to determining the probability that the conditional bet will be won. There is only a (non-void) bet, just as there is only an (non-void) assertion, when A holds, and so the probability of winning the bet is again $P(C|A)$, which also determines the fair betting odds for the bet (Politzer et al., 2010). It has to be noted that in this view, not-A is not 'irrelevant' in some absolute sense. If not-A is the outcome, a bet on if A then C, is neither won nor lost, just as an argument about if A then C is neither won nor lost, and no money or other utility changes hands. That could be significant for the person who made the bet, or the assertion. Relating this account to our example of the mother that promises an ice cream to her son, what is the probability that the mother will keep her promise? It is intuitively the conditional probability that the boy will get the ice cream later given that he is quiet now. The mother keeps her promise when the boy is quiet and gets his ice cream. She breaks her promise when the boy is quiet and does not get his ice cream. But his not keeping quiet is 'irrelevant' to determining whether she keeps her promise or not. In that case, she neither keeps nor breaks her promise, which is hardly an 'irrelevant' outcome in a wider sense.

In the light of the ST and findings of Evans et al. (2007, 2008), Barrouillet (2011; Barrouillet, Gauffroy & Lecas, 2008; Gauffroy & Barrouillet, 2009, 2011) tried to integrate a mental models account within the revised heuristic/analytic model of Evans (2006). This resulted in a *Revised Mental Models Theory* (R-MMT), which states that the initial model represents explicitly those states of affairs that make the sentence psychologically true when they occur (Gauffroy & Barrouillet, 2009). In the case of basic conditionals (those with no semantic, referential or knowledge-based relation between the antecedent and the consequent), this is the situation where p (the antecedent) and q (the consequent) are satisfied. The initial model comes spontaneously to mind through heuristic processes. Fleshing out

leads to an enrichment of this initial model by representing those states of affairs that are compatible with the sentences, but that do not make the sentence true. As a consequence, those who are able to construct the three-model representation would judge the conditional 'true' for the TT case and they would say 'indeterminate' to FF and to FT cases. Fleshed out models are constructed through analytic processes that enrich the representation but are optional in nature. When confronted with these states of affairs, individuals consider the conditional neither true, because the corresponding models did not come spontaneously to mind, nor false because these situations are nonetheless compatible with the conditional and do not contradict it. The TF case does not correspond to any model, and therefore will be considered as incompatible with the sentence and as falsifying it.

This R-MMT is in line with developmental changes that occur in the fleshing out process. Primary school children (about 8 or 9 years of age) are only able to construct the initial model $A \rightarrow C$ but fail to envisage other possibilities, thus adopting a conjunctive interpretation. Children around age 12 are able to flesh out the initial representation by adding a $\neg A \rightarrow C$ model, favoring a biconditional interpretation. Finally, older adolescents are able to construct the complete three-model representation, which results in three-valued answer patterns (Barrouillet et al., 2008; Gauffroy & Barrouillet 2009; 2010).

Correlation with Cognitive Ability and Working Memory Load

There is some debate in the reasoning literature concerning the relation between cognitive ability and the mental representation of conditionals. The three-valued answer patterns -the irrelevance judgments of the FA cases- make up a key element in dividing the C-MMT, the ST and the R-MMT since these theories have opposite predictions about it.

According to the *C-MMT* (Johnson-Laird, 2006; Johnson-Laird & Byrne, 2002), the people with the highest working memory capacity are able to look beyond the initial model ($A \rightarrow C$) and consider all possibilities compatible with the conditional, including the FA cases. This

in turn leads to two-valued answer patterns. The people with lower working memory capacity however do not process beyond the initial model and respond according to a conjunctive or a three-valued pattern.

The *ST* however predicts that the answer patterns given by the most intelligent reasoners are three-valued or ‘defective’. Evans et al. (2007, 2008) show that the participants who are responding according to three-valued patterns are those with the highest score on a general intelligence test: These are the ones that are able to complete both stages of the Ramsey-Test and thus to consider both $A \supset C$ and $A \supset \neg C$. The ones not able to do so consider only the first stage and give a conjunctive answer pattern, regarding TF, FT and FF cases as false. In the latter group of participants, less ‘irrelevance’ judgments are observed compared to the high-ability group since they do not give irrelevance judgments because of a suppositional processing of the conditional: According to Evans et al. (2007), there are only two factors known to cause defective answer patterns in the truth table task: (1) a genuine suppositional processing of the conditional and (2) a mismatch of the lexical content (matching bias). Low-ability participants are not able to complete the Ramsey-Test and to reason suppositionally, so lower rates of defective truth tables are predicted for this group since they are solely caused by matching bias.

Like the *ST* –but for different reasons– the *R-MMT* predicts a positive correlation between cognitive ability and the production of irrelevance responses. According to Barrouillet (2011), in order to produce these irrelevance judgments, one has to flesh out the initial model and construct a three-model representation, which is the result of demanding analytic processes. Hence, one expects more ‘irrelevant’ judgments of the FA cases by the most intelligent ones.

In the present study we are not solely trying to replicate the positive correlation between intelligence and the judgments of FA cases in a truth table task as ‘irrelevant’

(Barrouillet, Gauffroy & Lecas, 2008; Evans et al., 2007, 2008; Sevenants, Dieussaert & Schaeken, 2011). Instead, we want to directly manipulate cognitive ability by means of a dual task paradigm, a manipulation that has never been implemented before in this domain but shown to be successful in combination with other reasoning tasks (see for example De Neys & Schaeken, 2007; De Neys, Schaeken & d'Ydewalle, 2005; De Neys & Verschueren, 2006). Participants will perform a truth table task while their executive resources are burdened with a secondary task. Putting differential load on working memory while participants have to perform a truth table task, we have a clean measure of the impact of working memory capacity on the response patterns yielded by the truth table task. If there is a differential processing depth at play, we expect an effect working memory load, whereas a genuine difference in mental representation will not be altered by the external working memory load.

Classification Times

Next to the manipulation of working memory load, the second innovative part of the present study is the registration of the classification times in the truth table task. In previous studies, the degree to which FA cases are processed in truth table tasks is derived from the resulting answer patterns, which has led to substantial debate between the ST and the C-MMT regarding the underlying mental representation of conditionals. Solely examining participants' truth table task judgments is not the most direct measure of the type and amount of processing that has preceded participants' judgments. Instead of only looking into participant's judgments, in the present experiment we will also inspect the classification times, i.e., the time it takes participants to make their judgments of the truth table cases in order to shed light on the nature of the process that underlies these judgments (see also Evans & Newstead, 1977).

Previous research already showed that longer reading times are indicative of longer mental processing of a clause (Schroyens, Schaeken, Fias & d' Ydewalle, 2000) and that the

time spent making a decision provides a measure of which cards are considered relevant in Wason's selection task, because the more time spent thinking about a particular card, in turn predicts its eventual selection (Ball, Lucas, Miles & Gale 2003; Evans, 1996; Roberts, 1998a; 1998b). The latter observation is referred to as the 'inspection time effect'. The rationale behind this effect is that the heuristic processes draw people's attention towards relevant cards, which would subsequently be considered and then selected (i.e., the relevance-heuristic). Cards not determined as being relevant would be considered for much less time, if at all. In the present experiment we will not use the inspection time paradigm as such, but we will measure the time it takes participants to make their truth table case judgments by registering their classification times, i.e., the time between the appearance of the case and making the judgment by pressing the key.

In the following paragraphs, we will elaborate on predictions of both the ST and both mental models accounts with respect to the classification times on the truth table task. We want to state explicitly, however, that neither the C-MMT nor the ST makes explicit predictions about the classification times in the truth table task and that these predictions are therefore derived from our own reading of the theories. The inferred predictions are the same for the three experiments in the present study and will not be repeated extensively in the introduction of each experiment for the sake of simplicity.

According to the *ST*, participants solve a truth table task by running Ramsey-Tests: Thinking about conditionals in terms of p possibilities, they hypothetically suppose p (the antecedent) and then run a mental simulation in which they evaluate q (the consequent). As a consequence of this Ramsey-Test, participants disregard the FA cases (FT and FF) and only focus on those cases in which the antecedent is 'true'. In his heuristic/analytic model of the conditional truth table task, Evans (2007, p. 58) argues that heuristic processing leads to an 'irrelevant' judgment of the FA cases, whereas the judgment of the true-antecedent cases

requires analytical processing (i.e., the Ramsey-test). This leads to the key prediction of the ST that the judgment of the TT case as ‘true’ has to take more time than the judgment of the FA cases as ‘irrelevant’. The ST thus predicts that the categorization of a FA case as ‘irrelevant’ is a largely automatic, System 1 process (Evans & Over, 2004 p. 24) similar to the ‘relevance-heuristic’ observed by Evans (1996) with Wason’s selection task. Both the TT and the TF case are therefore expected to have a relatively long classification times compared to the FT and the FF case, and this because of the perceived relevance of the former. If it is the case that a relevance-heuristic comes into play, as is the case with the not-chosen cards in Wason’s selection task, participants would not consider the FT and FF case at all if they were not forced to decide on them. Within the task format we use, participants are forced to consider the false antecedent, so we expect that the classification times for the judgments of the FA cases will be very short, and that an irrelevance answer will follow rapidly, without much underlying analytic consideration. This prediction shows an inconsistency in the ST (see Barrouillet, 2011 for this point): The ST argues and it is observed that high-ability reasoners produce more irrelevance answers, which suggest that these answers are based on analytic processes. However, Evans (2007) states that heuristic processes cue true-antecedent cases as relevant and FA cases as ‘irrelevant’.

According to the *C-MMT*, people construct mental models of the possibilities compatible with the premises, but they initially and by default do not represent what is false. When they have disposal of sufficient cognitive ability, participants can flesh out the implicitly represented information into fully explicit models, which represent clauses even when they have false antecedents. If this is the process underlying the representation of conditionals, then we expect short classification times for the TT case since this is the situation that is represented in the initial model that is already constructed at that point. Next, it is to be predicted that the TF situation will rapidly be judged to be ‘false’, since no fleshing

out is required to make this evaluation. At least, this is expected when assuming that the antecedent is represented exhaustively in the initial model, as is the case in the original version of the C-MMT (depicted by square brackets [A] C, see Johnson-Laird et al., 1992). In a later version of the C-MMT (Johnson-Laird et al., 2002), however, these square brackets are omitted, which makes it hard to predict whether fleshing out in fully explicit models is required to judge the TF case to be ‘false’.

Concerning the FA cases, the C-MMT makes a distinction between those participants who flesh out and will eventually judge the FA cases to be ‘true’, and those who fail to flesh out and will judge the FA cases to be ‘irrelevant’. For the participants who are unable to flesh out and base their judgment solely on the initial model, we predict a fairly short classification time for the FA cases as ‘false’ (comparable to the classification time for TT and TF). For the participants who flesh out and give a ‘true’ judgment of the FA cases, which is the ‘normal’ answer pattern according to the mental models theory), classification times for the FA cases are expected to be longer than the classification time on the TT case (i.e., the time it takes to flesh out). The key prediction of the C-MMT therefore is that, for participants responding according to a two-valued answer pattern, a ‘true’ judgment of the FA cases should take longer than a ‘true’ judgment of the TT case.

According to the *R-MMT*, the initial model comes spontaneously to mind through heuristic processes. Therefore, short classification times for the TT case are expected. The ‘false’ judgment of the TF case might be the result of two different processes. Either it results directly from the initial model and a classification time as short as that the classification time for TT as ‘true’ is expected, either it is concluded only after a fleshing out process that the TF case does not match any of the models (Vergauwe, Gauffroy, Morsanyi, Dagry, & Barrouillet, in press).

Fleshed out models of the FA cases are constructed through time consuming analytic processes that enrich the initial representation. When confronted to these states of affairs, high ability individuals consider the conditional neither ‘true’, because the corresponding models did not come spontaneously to mind (i.e., not in the initial model), nor ‘false’ because these situations are nonetheless compatible with the conditional and do not contradict it. Instead, they classify these cases as ‘irrelevant’. Therefore, the key prediction for participants responding according to a three-valued answer pattern is that longer classification times are needed to judge the FA cases as ‘irrelevant’ than to judge the TT case as ‘true’. More specifically, this theory predicts that for participants with a *defective conditional* interpretation, who judge both the FT and the FF case as ‘irrelevant’, it should take longer to classify FT than FF as ‘irrelevant’, because the latter can be based on the representation of two models ($A \supset C$ and $\neg A \supset \neg C$) whereas the judgment of FT as ‘irrelevant’ requires a full three model interpretation ($\neg A \supset C$ in addition to the two other models). Participants with a *defective biconditional* interpretation (TFFI), however, classify the FF case as ‘irrelevant’ and the FT case as ‘false’, both judgments being based on the construction of only two models ($A \supset C$ and $\neg A \supset \neg C$). Classification times of these cases should therefore not differ from each other (Vergauwe, Gauffroy, Morsanyi, Dagry, & Barrouillet, in press).

In Sum

With the present series of experiments we want to shed light on the relation between intelligence and the mental representation of conditionals, assessed by the truth table task paradigm. Doing so, we will test the contrasting predictions of both mental models accounts and the ST. More concretely, the aim of the present study is twofold: Firstly, instead of solely replicating the correlation between cognitive ability and the answer patterns yielded by the truth table task, we directly manipulate working memory load in order to shed light on the observed difference between high and low ability participants regarding their truth table task

judgments. Secondly, since we want to unravel the nature of the processes/representation underlying these truth table judgments, we do not confine ourselves to registering participants' answer patterns but we also inspect the classification time of each judgment.

Experiment 1

Executive working memory resources are the quintessential component of computational cognitive ability (e.g., Evans et al. 2003; Stanovich & West, 2000). Therefore, in the present study we will adopt some ability measures and secondary tasks specifically designed to measure and burden executive working memory resources. In the first experiment, participants will solve a truth table task while differential load (high load – low load – no load) is put on their working memory by means of the dot memory task. Moreover, we will inspect the correlation between the answer patterns yielded by the truth table task and two cognitive ability measures (see Method section)

According to both mental models accounts (Gauffroy & Barouillet, 2009; Girotto & Johnson-Laird, 2004; Johnson-Laird, 2006; Johnson-Laird & Byrne, 2002), participants with low cognitive ability only consider the explicit initial model A C. Therefore, they give either a conjunctive -according to the R-MMT - or a three-valued² answer pattern- according to the C-MMT - on the truth table task. Compared to the no-load conditions, putting *high load* on participants' working memory should therefore result in more conjunctive answer patterns according to the former and in more three-valued answer patterns according to the latter theory. In the *no/low load* conditions, it is expected on the basis of the C-MMT that more two-valued answer patterns will be observed than in the high load condition, since those

² From the current version of the C-MMT, it is hard to predict whether participants will give either a conjunctive or a three-valued answer pattern, since both result from the initial model (depending on whether the initial model contains exhaustivity markers). Based on previous studies, in which far more three-valued than conjunctive patterns were observed, it is to be predicted that participants with a high load will produce more three-valued patterns than conjunctive patterns, but there is need for a more clear account from the C-MMT.

participants have more cognitive resources left to flesh out the implicit information into fully explicit models, containing the FA cases. These are the participants that, according to the R-MMT, should give more three-valued answer patterns under no/low load.

In line with the *ST* (Evans et al, 2007), it is predicted that putting *no/low load* on participants' working memory will result in three-valued answer patterns, since these participants have enough cognitive resources for a deep processing of the conditional. In that case both stages of the Ramsey-Test are completed while the FA cases are ignored, resulting in a three-valued answer pattern. In the *high load* condition, participants have only enough mental resources left to engage in a shallow processing of the conditional. This means that only the first stage of the Ramsey-Test is performed, solely considering A C and leading to a conjunctive answer pattern.

In the present study, all participants' classification times in the truth table task are registered. Due to the nature of the underlying processes (heuristic or analytic), the three theories have different predictions. The *ST* expects 'irrelevant' judgments of the FA cases that are clearly faster than the 'true' judgment of the TT case. According to the *C-MMT*, FA cases are expected to yield slow 'true' judgments whereas the 'true' judgment of TT and the 'false' judgment of TF are quick. Finally, according to the *R-MMT*, fast 'true' judgments of the TT case and slow judgments of both the TF case as 'false' the FA cases as 'irrelevant' are expected.

Method

Ethics statement

All experiments in this study were conducted in accordance with the Declaration of Helsinki and approved by the local ethics committee of the University of Leuven. Written informed consent was obtained from all participants.

Participants

In the present experiment, 176 first-year Psychology undergraduate students (17-24 years of age, $M = 18.3$) at the University of Leuven took part in partial fulfillment of course requirements. None of them had ever taken a logic-course.

Design

Participants were run in groups of 30 and were presented with the computerized tasks described below. The order of the tasks was the same for all participants. First they received the short version of the APM or Raven's Advanced Progressive Matrices (= covariate 1). Next, they were presented with the CRT or Cognitive Reflection Task (= covariate 2). After completing these two tasks measuring their cognitive ability, participants were asked to complete the truth table task. From this truth table task we registered both participants' answer patterns (= dependent variable 1) and their classification times (= dependent variable 2). The truth table task itself was the same for all participants, but it was combined with three different levels of load, imposed by the Dot Memory Task in a dual-task paradigm. This combination of the truth table task with the three load levels of the Dot Memory Task yielded three conditions: one group of the participants received a high-load, the second group a low load and the third group was assigned to the no-load condition (cfr. *Infra*). The assignment of the participants to the load-conditions was completely randomized. So there were three independent variables: working memory load (high, low, no), truth table condition (TT, TF, FT, FF), and case (affirmative antecedent and consequent TT, affirmative antecedent/negative consequent TF, negative antecedent/affirmative consequent FT, and negative antecedent and consequent FF).

All parts of the experiment were presented in Dutch and the study lasted between 30 and 40 minutes. .

Materials and Procedure

In reasoning studies a difference exists in how the concept of cognitive ability is exactly made concrete. According to the C-MMT (Johnson-Laird & Byrne, 2002), cognitive ability is conceived in terms of working memory capacity or the ability to have several possibilities in mind at the same time (Johnson Laird, 2006). In the ST, cognitive ability is more in terms of general intelligence (Evans et al, 2007, 2008). Since this difference in how the concept of ‘intelligence’ is made concrete might have an influence on how individual differences relate to the performance on the truth table task, we will administer two tests of cognitive ability in the present study: a more general intelligence test and a test that measures the ability to inhibit heuristically appealing response alternatives. Both ability measures will be discussed in the following paragraphs.

Raven’s Advanced Progressive Matrices (APM)

The first of the ability measures we are interested in is Raven’s Progressive Matrices (Raven, Raven & Court, 1991). It is a standardized intelligence test that consists of visually presented, geometric-analogy-like problems in which a matrix of geometric figures is presented (3 rows x 3 columns) with one of the nine entries missing. The correct missing entry must be selected from a set of answer choices. During the last decade, this has been one of the most widely used instruments by researchers interested in participants’ inductive or analytic reasoning capacities or fluid intelligence (Cattell, 1963). Raven’s Advanced Progressive Matrices (Raven, Court, & Raven, 1988) or APM is a version of these matrices intended for use with people above average aptitude and designed to reliably differentiate among those in the top 25% of the population (Bors, & Stokes, 1998), as we believe is the case for our first-year university students. The test takes about 40 minutes. In the present study a short version of the APM (Bors, & Stokes, 1998) has been used. It is a selection of 14 items with increasing difficulty drawn from Set 2 of the original APM (item 3, 10, 12, 15, 16,

18, 21, 22, 28, 30, 31 and 34). The APM was administered during the same session than the truth table task, together with the CRT questionnaire (cf. *Infra*). For all 14 items of the APM, participants had to indicate which was the missing segment required to complete the 3x3 matrix, which takes about 20 minutes.

Cognitive Reflection Test (CRT)

The second test of cognitive ability we administer in the present experiments is the Cognitive Reflection Test (CRT; Frederick, 2005). We have three reasons for choosing this test: First of all, it only consists of three questions and is easily administrable for large groups of students. Secondly, it is a measure of general intelligence (Navarrete, De Neys, & Santamaria, 2008; Sevenants, Dieussaert, & Schaeken, 2011) correlating .44 with the Scholastic Achievement Test (SAT; Frederick, 2005) and .35 with a short version of the Raven Progressive Matrices (Sevenants, Dieussaert, Schaeken, & Barrouillet, unpublished manuscript). Finally, the CRT is conceived as a measure for reasoning beyond the quick and automatic response (Frederick, 2005). Since reasoning beyond this first quick and automatic response is what people with high cognitive abilities do, both according to the ST and the mental models accounts, we choose to include the CRT next to the APM as a measure of reasoning beyond the heuristic response. Toplak, West, and Stanovich (2011) and Stanovich (2011) show that although the CRT has a substantial correlation with cognitive ability the CRT is a unique predictor of performance on heuristics-and-biases tasks. They argue that the CRT is a particularly potent measure of the tendency toward miserly processing, predicting rational thinking performance independent not only of intelligence, but also of executive functioning and thinking dispositions. The CRT thus measures properties relevant to rational thinking that go beyond those measured on intelligence tests.

The CRT (Frederick, 2005) was administered during the same session as the truth table tasks. All participants received a sheet on which the instruction “Below are several

problems that vary in difficulty. Try to answer as many as you can.” appeared, followed by the three items of the CRT. The items as well as the instructions were presented in Dutch. The CRT score is the sum of the number of correct items, ranging from 0 to 3.

(1) A bat and a ball cost \$1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?

_____ cents

(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

_____ minutes

(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

_____ days

Truth Table Task

All participants were presented with a three option version of the truth based truth table task. They all received the same instructions, appearing on the first computer screen and explaining that the purpose of the experiment was to examine how people reason with conditionals. The instructions also contained the description of a machine producing cards with a letter on the front side and a number on the back, always doing so following a certain rule, for example *‘If there is an A on the front, there is a 2 on the back of the card’*. Participants could read that in the upcoming task they were going to see 16 cards produced by that machine, and that per card they had to evaluate the compatibility of that specific card with the card-producing rule. Participants were then provided with a concrete example of a conditional rule, as well as with an example of the item in the actual task (no correct answers were provided). The instructions

were followed by two practice trials. In the instructions it was made clear to the participants that they had to judge each time, for one specific card taken randomly from the pack, whether it was ‘true’/‘false’ that this specific card was produced by that machine. This because the scope of the conditional rule cannot be a whole pack of cards (as is the case for general conditionals, which are not made definitely true by only one TT case, see also Schroyens, 2010 on the induction problem) but has to apply only to a single card taken randomly from the pack (Evans et al., 2003, p. 333). The conditional rule in the present truth table tasks is singular and has a definite description, e.g., "If the letter is an A..." and does not begin with an indefinite quantifier characterizing general conditionals, as in "If a card had a T on the front ..." or "If any card has an A on the front ...".

In this study, we employ the truth table task with implicit negations, using the negations paradigm in order to control for the effects of negations and matching bias (see also Evans, Handley, Neilens & Over, 2007). Using the negations paradigm means that conditionals are presented in four forms, with the presence of negations varied in the antecedent and the consequent (Affirmation-Affirmation, Affirmation-Negation, Negation-Affirmation and Negation-Negation; see Table 1 of the present manuscript and also Table 1 in Evans et al, 2007 p. 1775 for an illustration of the different cases). So all participants were successively presented with 16 items, each item consisting of the abstract conditional rule followed by one of the four combinations of occurrence and non-occurrence of A and C. We chose for an evaluation format of the truth table task. This means that for each antecedent-consequent combination (the “card”), participants had to evaluate whether that specific situation made the given rule ‘true’, whether it made the given rule ‘false’, or whether it was ‘irrelevant’ with respect to the truth of the rule. No feedback was provided. For each item, the rule followed by one of the four situations appeared on the screen. Each item appeared on a different screen and presentation order of the items was randomized.

The truth table task was constructed with 'E-prime' software (Psychological Software Tools, Pittsburgh, PA) and presented to the participants on individual standard PCs in a self-paced manner. Responses were given with the arrow-keys on an AZERTY keyboard. The task lasted between 5 and 10 minutes.

Working Memory Load - Dot Memory Task

The dot memory task is a classic spatial storage task (e.g., Bethell-Fox & Shepard, 1988; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) and the version we will use is based on De Neys, d'Ydewalle, Schaeken and Vos (2002). For the present study, a 3 x 3 matrix filled with three or four dots will be presented for 850 ms. Participants memorize the pattern and will be asked to reproduce it afterwards. In the *high-load* condition, the matrix is filled with a complex four-dot pattern. In the *low-load* condition, the pattern consists of three dots on a horizontal line. This pattern should place only a minimal burden on executive resources. In the *no-load* condition finally, no dual task is presented. The load variable is between subjects, implying that each participant will only receive one type of load, either no load, low load or high load. In the low- and high-load conditions, a dot pattern is presented for 850 ms before each of the 16 situations (TT, TF, FT or FF) of the truth table task. Concretely, a trial the dual task combined with the truth table task looks as follows: First, the participant is presented only with the conditional rule. This allows us to disentangle reading times and classification times since participants have already read the conditional rule at the moment they have to evaluate the card. After pressing the space bar (self-paced) the next screen with the dot matrix appears for 850ms (only for the load conditions). The next screen shows the conditional rule together with one of the 16 combinations of antecedent and consequent (cfr. supra). This is the situation the participants have to judge. After participants' evaluation of the situation by means of the key-arrows, an empty matrix is presented on the next screen, and they have to

reproduce the dot pattern. After the reproduction of the dot matrix the next trial starts. Instructions will emphasize that it is crucial for the dot patterns to be reproduced correctly.

Results

Truth Table Task Judgments and Dot Memory Task

With respect to participants' classifications of the truth table cases, the general observation is a judgment of 'true' for the TT case, 'false' for the TF case and 'irrelevant' for the FA cases FT and FF (see Table 1; for the frequencies of the individual answer patterns see the Appendix). As observed by Evans et al. (2007), there is a large amount of variability in the cases other than TT. The main reason for this to happen is matching bias: For all cases, irrelevance judgments are highest for double-mismatch cases (--) and lowest for double-match cases (++), with the single-match cases (+- and - -) intermediate.

“(Table 1 about here)”

Results for the dot memory task indicated that the dual task was properly performed. Regarding the simple dot pattern in the *low-load* group, the mean number of correctly localized dots out of three was 2.62 ($SD = .31$). For the complex dot pattern in the *high-load* condition, the mean number of correctly localized dots out of four was 3.30 ($SD = .36$). Thus, overall, about 83% of a complex dot pattern was still reproduced correctly. Hence, participants were not simply trading off dot recall performance to solve the truth table task.

Correlation with Cognitive Ability and Working Memory Load

We will briefly explain in the paragraphs below the indices that were calculated (by analogy with Evans et al., 2007; 2008; Sevenants et al., 2011; Sevenants, Dieussaert & Schaeken, 2012). Regarding the judgments on the truth table task, we divided the truth tables into two components: the true-antecedent cases (TT and TF) and the FA cases (FT and FF). According to all theories, participants should judge the TT case to make the conditional 'true' and the TF

case to falsify the conditional. We combined these judgments into the logic index (L-I): per participant the TT situations judged as ‘true’ plus the TF cases judged as ‘false’. Regarding the FA cases, we computed the false-antecedent index (FA-I), which is the sum per participant of FA cases (FT and FF) judged as ‘irrelevant’. Both the L-I and the FA-I range from 0 to 8.

In addition to the indices as used by Evans et al. (2007, 2008), we computed three more indexes. According to the ST, participants should judge the TT case as ‘true’, the TF case as ‘false’ and both the FT and the FF case as ‘irrelevant’. We combined these judgments into the *Suppositional Theory index (ST-I)*, ranging from 0 to 16 per participant. Moreover, we calculated the *material implication index (MI-I)*, summing per participant the number of TT cases judged as ‘true’, TF cases as ‘false’, and FT and FF cases as ‘true’ (range 0-16). This is the classification pattern the C-MMT would predict for those participants with [A] C as initial model who are able to flesh out into fully explicit models. Finally we computed a *conjunctivity index (Conj-I)* with a range from 0 to 16. This classification pattern in which only the TT case is judged as ‘true’ and the three other cases as ‘false’ is predicted both by the mental models accounts and the ST for the participants with the least cognitive resources, unable to reason beyond the initial model or to engage in the second stage of the Ramsey-Test.

“(Table 2 about here)”

Considering the relation between participants’ truth table classifications and cognitive ability, we observed a similar pattern of results for all three measures of cognitive ability. First of all, a zero-correlation was revealed between all ability measures and the Conj-Index. Secondly, there was a significant positive correlation between all ability measures and both the FA-Index and the ST-Index (see Table 2). Finally, no significant correlation was observed between the ability indices and the MI-Index.

Next to the inspection of the correlations between cognitive ability and the truth table case judgments, we focus on the effect of the working memory load manipulation. T-tests revealed no effect at all of working memory load on the computed indices (Log-I, FA-I, Conj-I, ST-I and MI-I). There was also no correlation between the performance on the dot memory task (see Table 1, ‘Dual’) and these indices.

Classification Times

For a complete overview of the classification times per truth table case (TT, TF, FT, FF) and per type of conditional (AA, AN, NA and NN), see Appendix B.

In order to test the key predictions of the three different theories discussed in this paper, we compared, per individual participant, the time it took to classify the TT case as ‘true’ on the one hand with the classification time of the FA cases FT and FF as ‘true’ and as ‘irrelevant’ on the other hand. Classification times were averaged over participants and participants not producing both of the judgments under investigation were left out of the analysis. T-tests revealed the following results:

The judgment of TT as ‘true’ ($M = 4997$ ms, $SD = 2674$ ms) was faster than the judgment of FT as ‘true’ ($M = 7076$ ms, $SD = 5143$ ms, $t(138) = -3.00$, $p < .01$). The same was true for TT as ‘true’ ($M = 5182$ ms, $SD = 3153$ ms) versus FF as ‘true’ ($M = 8001$ ms, $SD = 4211$ ms, $t(250) = -6.02$, $p < .01$). Classification times for TT as ‘true’ ($M = 5270$ ms, $SD = 3229$ ms) were shorter than the classification times for FT as ‘irrelevant’ ($M = 6960$ ms, $SD = 4364$ ms, $t(320) = -3.95$, $p < .01$). Similarly, classification times for TT as ‘true’ ($M = 5174$ ms, $SD = 3175$ ms) were shorter than for FF as ‘irrelevant’ ($M = 6175$ ms, $SD = 4054$ ms, $t(312) = -2.45$, $p < .01$). Finally, the judgment of TT as ‘true’ ($M = 5236$ ms, $SD = 3234$ ms) took equally long than the judgment of TF as ‘false’ ($M = 5196$ ms, $SD = 2569$ ms, $t(350) = .13$, $p = .89$).

In the analyses presented above, classification times were averaged over the four types of conditional used in the negations-paradigm (AA, AN, NA and NN), so that matching bias is balanced overall. However, one can say this might obscure the interpretation of the results. Therefore we repeated the analyses, confining ourselves to the AA conditional without negations (i.e., the ‘traditional’ conditional “If A then C”). The pattern of classification times was identical compared to the averaged results described above (see Appendix B): Classification of TT as ‘true’ was faster than all judgments of the FA cases.

Regarding the judgments of the FA cases for the AA conditional, we focused on the classification times of the predominant answer patterns (see Appendix A for the answer patterns and Appendix B for the classification times). For the *defective conditional reasoners* (TFII), the classification of FT as ‘irrelevant’ ($M = 6991$ ms, $SD = 4785$ ms) was longer than the time to judge FF as ‘irrelevant’ ($M = 4430$ ms, $SD = 3636$ ms; $t(154) = 3.76, p < .01$). The judgment of TT as ‘true’ ($M = 3849$ ms, $SD = 2629$ ms) was not significantly shorter than that of FF as ‘irrelevant’ ($t(154) = -1.14, p = .25$). With respect to the *defective biconditional reasoners* (TFFI), classification times for FT as ‘false’ ($M = 4824$ ms, $SD = 3373$ ms) did not differ from those for FF as ‘irrelevant’ ($M = 4716$ ms, $SD = 3520$ ms; $t(129) = .18, p = .86$). The judgment of TT as ‘true’ ($M = 3472$ ms, $SD = 2270$ ms) was faster than both the judgment of FT as ‘false’ ($t(129) = -2.69, p < .01$) and the judgment of FF as ‘irrelevant’ ($t(129) = -2.41, p = .017$).

Discussion

In the present experiment we observe an overall tendency of classification in conformity with a three-valued truth table pattern, i.e., ‘true’ for the TT case, ‘false’ for the TF case and ‘irrelevant’ for the FA cases FT and FF, which is supportive for the ST and the R-MMT.

With respect to the correlation between cognitive ability and the answer patterns, a positive correlation between cognitive ability and the number of irrelevance judgments on

these cases is revealed. This is also in favor of both the ST and the R-MMT, according to which participants with the most cognitive resources judge FA cases to be ‘irrelevant’. This finding is in turn problematic for the C-MMT, which predicts participants low rather than high in cognitive ability to judge FA cases as ‘irrelevant’ instead of as ‘false’. According to this account, participants high in cognitive ability should be able to look beyond the initial model, leading to irrelevance judgments, and consider the three fully explicit models leading to a ‘true’ judgment. In our data however, no correlation was observed between the cognitive ability measures and the MI-Index.

Contrary to the correlational pattern, no effect of the working memory load manipulation was observed: In the no-load, the low-load and the high-load condition, the answer patterns remain the same. This might be due to the fact that the (visuospatial) dual task we used was not able to burden the working memory resources needed to solve the (verbal) truth table task. In order to test this hypothesis, in Experiment 2 we will use a verbal instead in a visuospatial working memory load.

With respect to the classification times, we focus on the key predictions of the three different theories about the classification times per participant, as elaborated on in the general introduction. The *C-MMT* predicts ‘true’ judgments of the TT case to be faster than any other judgment since this is the ‘standard judgment’, based on the initial model and thus almost immediately and automatically available. With respect to the FA cases, the C-MMT predicts slow ‘true’ judgments because these require fleshing out of the initial model: The information about the FA cases that is represented in an implicit way in the initial model, has to be made explicit. And this is both a time and resources consuming process. The C-MMT thus predicts ‘true’ judgments of the FA cases that are considerably slower than the ‘true’ judgments of the TT case. In our data this prediction is confirmed. Participants judging the TT case as well as the FA cases as ‘true’ need significantly more time to give the latter judgment (this goes both

for the FT and the FF case). Problematic for the C-MMT however is that only a small minority of participants respond according to the two-valued answer pattern TFFT.

According to the *Revised Mental Model Theory*, true' judgments of the TT case should be faster than any other judgment since this judgment is based on the initial model and thus automatically available. With respect to the FA cases, the R-MMT predicts slow 'irrelevant' judgments because they require fleshing out of the initial model, which is an analytical and therefore consuming process. Basically, the R-MMT therefore predicts 'irrelevant' judgments of the FA cases which are slower than the 'true' judgment of the TT case. This is what we observe in our data: The participants responding according to a three-valued answer pattern (which is a majority) need more time to judge both FA cases as 'irrelevant' than to give a 'true' judgment for TT. More specifically, the predictions are confirmed that for defective conditional reasoners it should take longer to classify FT as 'irrelevant' (3 models required) than FF (2 models required) as 'irrelevant' and that for participants with a defective biconditional interpretation, classification times of these cases (i.e., FF as 'irrelevant' and FT as 'false') should not differ because both classifications are based on two models. The ST finally predicts fairly long classification times for TT as 'true' and TF as 'false', because these judgments need the completion of both stages of the Ramsey-Test (cfr. Supra) while the FA cases are heuristically determined as 'irrelevant' and this judgment should therefore require almost no time. The ST predicts participants to give a three-valued answer pattern which consists of fairly slow judgments of TT as 'true' and of 'irrelevant' judgments of the FA cases that are significantly faster. This expectation was not confirmed by our data. On the contrary, the judgment of TT as 'true' is faster than the judgment of the FA cases as 'irrelevant'.

Regarding the classification of the TF case as 'false', we observe classification times that are equally long than those to classify TT as 'true'. This finding is in line with the ST,

which predicts 'false' judgments of the TF case that take the time needed to complete the Ramsey-test and thus are comparable to the classification time of the 'true' judgments of the TT case. It is hard to relate this finding to the other theories since both mental models accounts allow for both a fast (based on initial model) and a slow route (judgment after fleshing out) to classify the TF case as 'false'.

In sum, the data of Experiment 1 revealed that a majority of the participants responds according to a three-valued answer pattern. This answer pattern is positively correlated with cognitive ability but we were not able to find an effect of working memory load. Finally, classification times revealed that the predictions of the R-MMT fit our data the best: fast judgments of TT as 'true' and slow irrelevance judgments of the FA cases.

Experiment 2

In Experiment 1 of the present study, no effect of working memory load was observed on the answer patterns yielded by the truth table task. At this point we cannot exclude the hypothesis that a visuospatial dual task is not sufficient to burden the cognitive resources needed to make truth table case judgments, which are verbal in nature. The absence of a correlation between the performance on the load task and the cognitive ability indices also suggests that the verbal load we used might not have been appropriate to burden cognitive resources for this group of university students. In the present experiment, we rule out this hypothesis by using a different type of working memory load.

Some theorists have argued that working memory consists of multiple domain specific subsystems (e.g., Baddeley & Logie, 1999). This implicates that there would only be interference between two concurrent tasks when these tasks both rely on the same subsystem (i.e., both verbal or both spatial). Vergauwe, Barrouillet and Camos (2010) showed that both verbal and visuospatial recall decrease as a function of cognitive load, suggesting that tasks

compete for a common domain-general pool of resources. Nevertheless, they also found that verbal recall performance was poorer when it was combined with verbal processing than when it was combined with visuospatial processing.

In the present experiment we therefore repeat the truth table task with a working memory load manipulation, but instead of using the visuospatial Dot Memory Task we make use of a purely verbal working memory load. Similar to Experiment 1, classification times will be registered.

Method

Participants

Participants were 143 first-year Psychology students at the University of Leuven (17-28 years of age, $M = 19.1$) taking part in this experiment in partial fulfillment of course requirements. None of them had participated in Experiment 1.

Design

The design of Experiment 1 was identical to the design of Experiment 1. The only difference is that the truth table task is now combined with a verbal instead of a visuo-spatial dual task and that the no-load condition was not administered since this condition is exactly the same than the no-load condition of Experiment 1.

Materials and Procedure

Participants were run in groups of 30 students. Regarding the ability measures, both the *CRT* (Frederick, 2005) and a short version of the *APM* (Bors, & Stokes, 1998) were administered, as was the case in Experiment 1. The *truth table task*, with implicit negations and following the negations paradigm, is exactly the same as in Experiment 1: Participants have to evaluate for each of the 16 presented antecedent-consequent combinations whether the combination makes the given rule ‘true’, ‘false’, or is ‘irrelevant’ with respect to the truth of the rule.

Different from the previous experiment is the nature of the load presented in the dual task. In Experiment 1 we made use of the Dot Memory Task (De Neys & Schaeken, 2007; De Neys & Verschueren, 2006), which is a visuospatial load, whereas in the present experiment we use verbal load as a between-subjects variable. In the *high-load* condition, participants are presented successively with four words, in the *low-load* condition with two words. No *no-load* condition was present in this experiment, since this condition is exactly the same as the no-load condition of Experiment 1 and participants are drawn from the same population. Both in the low- and high-load conditions, the conditional rule is presented on the first screen. Pressing the space bar causes the words to appear on the next screen for 800 ms per word. After the two words in the low load condition or the four words in the high load condition are presented, one of the 16 the truth table task situations to be classified appears on the screen. The classification-time, i.e., the time between the moment the situation appears on the screen and the moment the participant presses one of the key-arrows to make his/her judgment, as well as the ‘true’, ‘false’ or ‘irrelevant’ judgment itself is registered per situation. After participants’ evaluation of the situation by means of the key-arrows, they have to write down the words of the dual task in the correct order. Instructions emphasize that it is crucial for the words to be reproduced correctly. After writing down the words, participants have to press the space bar in order to be presented with the next trial of the truth table task, implying that this task is self-paced. Presentation order of the 16 situations is randomized.

Results

Truth Table Task Judgments and Verbal Dual Task

Regarding participants’ truth table task judgments, an overall tendency of classification in conformity with a three-valued truth table pattern is observed, as was the case in Experiment 1 (i.e., ‘true’ for the TT case, ‘false’ for the TF case and ‘irrelevant’ for the FA cases FT and FF, see Appendix A; for the frequencies of the individual answer patterns,). For all cases,

irrelevance judgments are highest for double-mismatch cases (--) and lowest for double-match cases (++), with the single-match cases (+- and --) intermediate (see Table 3).

“(Table 3 about here)”

The scores on the verbal dual task indicated that this dual task was properly performed. Regarding the *low-load* group, the mean number of correctly recalled two-word lists out of 16 was 14.71 ($SD = 1.45$). For the four-word series in the *high-load* condition, the mean number of correctly recalled word-lists out of 16 was 11.46 ($SD = .3.62$). Thus, overall, about 72% of the four-word lists was still reproduced correctly. Hence, participants were not simply trading off word recall performance to solve the truth table task.

Correlation with Cognitive Ability and Working Memory Load

The same indices as in Experiment 1 were computed in order to analyze the judgments of the truth table cases. Regarding the relation between participants' truth table classifications and cognitive ability, we observed the same pattern of results as in the previous experiment: First of all, a null-correlation was observed between both the APM and the CRT on the one hand and the Conj-Index on the other hand. Next, a significant positive correlation is revealed between the APM and the CRT and both the FA-Index (see Table 4) and the ST-Index. Finally, no significant correlation was observed between the ability indices and the MI-Index.

“(Table 4 about here)”

We were not only interested in the correlation between cognitive ability and the truth table judgments, but also in the effect of working memory load on these judgments. So next to the inspection of the correlations between cognitive ability and the truth table classifications, we look into the effects of a verbal working memory load. There was a positive correlation between the score on the dual task (see Table 4, 'Dual') and the computed indices. As was the case in the previous experiment however, no effect of working memory load on the indices (Log-I, FA-I, Conj-I, ST-I and MI-I) was revealed by t-tests. This means that not only

burdening working memory capacity with a visuospatial load has no effect: Putting a verbal working memory load on participants' cognitive resources does not lead to an altered classification of participants' truth table case judgments.

“(Table 4 about here)”

Classification Times

See Appendix B for a complete overview of the classification times per truth table case (TT, TF, FT, FF) and per type of conditional (AA, AN, NA and NN). Similar to Experiment 1, in order to test the key predictions of the three different theories discussed in this paper, classification times of we compared, per individual participant, the time it took to classify the TT case as ‘true’ with the classification time of the FA cases as ‘true’ and as ‘irrelevant’. Again classification times were averaged over participants and participants not producing both of the judgments under investigation were left out of the analysis. T-tests revealed the following results:

The classification time of TT as ‘true’ ($M = 6519$ ms, $SD = 3631$ ms) was shorter than the classification time for FT as ‘true’ ($M = 10536$ ms, $SD = 9078$ ms, $t(102) = -2.96$, $p < .01$). The same was true for the classification time of TT as ‘true’ ($M = 6474$ ms, $SD = 3811$ ms) versus the classification time of FF as ‘true’ ($M = 10237$ ms, $SD = 7199$ ms, $t(236) = -5.03$, $p < .01$). Next, classification times for TT as ‘true’ ($M = 6722$ ms, $SD = 4129$ ms) were shorter than those for FT as ‘irrelevant’ ($M = 8729$ ms, $SD = 6053$ ms, $t(252) = -3.09$, $p < .01$). Likewise, classification times for TT as ‘true’ ($M = 5174$ ms, $SD = 3175$ ms) were shorter than those for FF as ‘irrelevant’ ($M = 6515$ ms, $SD = 4078$ ms, $t(258) = -4.06$, $p < .01$). The classification time of TT as ‘true’ finally ($M = 6592$ ms, $SD = 4089$ ms) was equally long than the one of TF as ‘false’ ($M = 7196$ ms, $SD = 4623$ ms, $t(284) = -1.17$, $p = .24$).

Regarding the AA conditional “If A then C” classification times were identical compared to the averaged results described above (see Appendix B): Classification times of TT as ‘true’ were significantly shorter than all judgments of FT and FF.

With respect to these judgments of the FA cases, we focussed on the classification times of the predominant answer patterns for the AA conditional (see Appendix A for the answer patterns and Appendix B for the classification times). For the *defective conditional reasoners* (TFII), the classification time of FF as ‘irrelevant’ ($M = 8693$ ms, $SD = 7355$ ms) was shorter than that of FT as ‘irrelevant’ ($M = 9784$ ms, $SD = 5807$ ms) but this trend was not significant ($t(80) = .75, p = .45$). The judgment of TT as ‘true’ ($M = 6618$ ms, $SD = 5089$ ms) was not significantly shorter than that of FF as ‘irrelevant’ ($t(41) = -1.49, p = .14$). Regarding the *defective biconditional reasoners* (TFFI), the judgment of FT as ‘false’ ($M = 6901$ ms, $SD = 5519$ ms) took equally long than judging FF as ‘irrelevant’ ($M = 5988$ ms, $SD = 5798$ ms; $t(136) = .95, p = .35$). The judgment of TT as ‘true’ ($M = 4832$ ms, $SD = 3717$ ms) was faster than both the judgment of FT as ‘false’ ($t(136) = -2.58, p < .01$) and the judgment of FF as ‘irrelevant’ ($t(136) = -1.40, p = .16$), but the latter comparison was not significant.

Discussion

First of all, as was observed in Experiment 1, the majority of the observed truth table answer patterns is three-valued in nature, meaning that by a majority of participants the TT case is judged as ‘true’, the TF case as ‘false’ and the FA cases as ‘irrelevant’. This is in favor of the ST as well as the R-MMT.

Regarding the correlation between cognitive ability and the truth table judgments, a positive correlation exists between cognitive ability on the one hand and the amount of irrelevance judgments on the FA cases on the other hand. This correlational pattern is similar to the pattern observed in Experiment 1 and is in line with the ST and the R-MMT, which

both predict a positive correlation between cognitive ability and the number of FA cases classified as ‘irrelevant’. Since no significant correlation between the cognitive ability indices and the MI-Index is observed, it seems not to be the case, as the C-MMT predicts, that participants with high cognitive ability flesh out the initial model into fully explicit models, which in turn should lead to a two-valued pattern of classifications.

Despite the correlation between cognitive ability and the truth table judgments, no effect of the working memory load manipulation itself is revealed. This means that we do not find support for our explanation of the data in Experiment 1 that the lack of an effect of working memory load is due to the differing nature of the load in the truth table task and the dual task does not hold: neither a visuospatial, nor a verbal load has an effect on the truth table case classifications. It still might be the case however, since the truth table task itself is self-paced, that participants rehearse the dual task during the truth table task. This would in turn eliminate the effect of load. In Experiment 3 we will therefore limit participants’ classification time in the truth table task.

The pattern of classification times is exactly the same as the pattern observed in Experiment 1: The majority of participants responds according to a three-valued answer pattern in which the judgment of the TT case as ‘true’ is significantly faster than the judgment of the FA cases as ‘irrelevant’. This finding, along with the confirmation of the specific predictions regarding the classification times of the FT and the FF case for the AA conditional (although not all significant), is supportive for the R-MMT.

Experiment 3

In Experiment 1 and 2 a positive correlation was found between cognitive ability and the tendency to judge FA cases as ‘irrelevant’. No effect was revealed of burdening participants’ working memory capacity on the truth table classifications, neither with a visuospatial load,

nor with a verbal load. However, since the truth table task is self-paced, participants use the time between seeing the case and judging the case not only to make their classification but also to rehearse the words presented in the secondary task (see also Barrouillet, Bernardin & Camos, 2004). This would mask the effect of working memory load, being either verbal or visuospatial in nature. In order to test the rehearsal- hypothesis, in this third experiment we put time pressure on the truth table task. This time pressure is determined so that participants are forced to make their judgments of both the true- and the FA cases within the timeframe that was needed to judge the true-antecedent cases in the previous experiment. That way, we prevent participants from rehearsing the words from the secondary task during the classification time. Evidence for such a rehearsal strategy can be found in the time-based resource-sharing model of Barrouillet et.al. (2004), stating that attention sharing is achieved through a rapid and frequent switching between processing and maintenance that occurs during the completion of the task.

In case of rehearsing the presented words of the dual task during the truth table task, as is stated by the alternative hypothesis, the effect of the load-manipulation would be wiped out. That is why in Experiment 3, next to the verbal working memory load presented in a dual task format, we put time pressure on the classification of the truth table cases. Preventing participants to rehearse the secondary task will give us a clear picture of the effect of working memory load on the truth table task and of the classification times needed to make these truth table task judgments.

Method

Participants

225 first-year Psychology students at the University of Leuven (17-27 years of age, $M = 19.2$) took part in this experiment in partial fulfillment of course requirements. None of them also participated in Experiment 1 or Experiment 2.

Design

The design of Experiment 3 was exactly the same than the design of the two previous experiments. The only difference with Experiment 2 is that now the three load-conditions are run and that the truth table task is timed instead of self-paced.

Materials and Procedure

Participants were run in groups of 30 students. Both the ability measures (*CRT* - Frederick, 2005 and a short version of the *APM* - Bors, & Stokes, 1998) and the truth table task with verbal load were exactly the same as in the previous experiment, except that in the present experiment we also have a no-load (no words) next to the low load (2 words) and the high-load) working memory load conditions. The most important difference with the previous experiment is that we not only make use of a dual task in order to burden participants' cognitive resources performing the truth table task, in the present experiment we also put time pressure on the evaluation of the 16 truth table task items. It was observed in Experiment 2 that it takes participants on average 6.5 seconds to classify a TT case in the truth table task, which was the shorted classification time out of the four cases. In the present experiment, participants in all load conditions are similarly forced to make their classification within that interval needed to classify that true-antecedent cases in order to be sure that participants would not engage in other reasoning processes that lie outside the classification of truth table cases. Judgments that were not made within the 6.5 s time interval were not registered. Following that practice, 131 out of 3584 (i.e., 16 judgments * 224 participants) or 3.6% of the total number of judgments was 'lost'. Instructions were the same as in Experiment 2, but now stated explicitly that they had to make their judgments in the truth table task as fast as possible and that the screen on which they could make their judgment would disappear after a short time.

Results

Truth Table Task Judgments and Verbal Dual Task

With respect to participants' truth table task judgments, as was the case in Experiment 1 and 2, an overall tendency of classification along a two-valued or three-valued truth table pattern is revealed, ('true' for TT, 'false' for TF and 'irrelevant' for FT and FF, see Table 5; for the frequencies of the individual answer patterns see the Appendix). Again for all cases irrelevance judgments are highest for double-mismatch cases (- -) and lowest for double-match cases (++), with the single-match cases (+- and - -) intermediate, a pattern of results which is due to matching bias and previously observed in the two preceding experiments of this study and by Evans et al (2007). Only a limited amount of data was lost due to the time pressure (maximum 11%, see cumulative percentages in Table 5), i.e., classifications not made within the interval of 6.5 seconds.

“(Table 5 about here)”

Results on the verbal dual task indicated that this dual task was not properly performed by the high-load condition under time pressure. Regarding the *low-load* condition, the mean number of correctly recalled two-word lists out of 16 was 13.24 ($SD = 2.38$). Regarding the four-word series in the *high-load* condition, the mean number of correctly recalled lists out of 16 was 7.64 ($SD = .3.62$). Overall, only about 48% of the four-word lists was reproduced correctly.

Correlation with Cognitive Ability and Working Memory Load

The same indices as in Experiment 1 and 2 were computed in order to analyze the judgments of the truth table cases: a *logic index* and a *false-antecedent index* (ranging from 0 to 8) and a *Suppositional Theory index*, a *material implication index* and a *conjunctivity index* (all ranging from 0 to 16). Regarding the relation between participants' judgments on the truth

table task and their cognitive ability a zero-correlation between both the APM and the the Conj-Index was observed, a negative correlation between this Conj-Index and the CRT, and finally a positive correlation between the APM and the CRT and both the FA-Index (see Table 6) and the ST-Index. No significant correlation was observed between both ability indices on the one hand and the MI-Index on the other hand.

As was the cases in the previous experiments we were interested in the effect of working memory load on the truth table classifications. In the present experiment participants were subjected to time pressure on the truth table task next to a secondary task. Similar to the previous experiments however, t-tests revealed that for none of the three working memory load conditions an effect of working memory load on the indices (Log-I, FA-I, Conj-I, ST-I and MI-I) was observed.

“(Table 6 about here)”

Classification Times

See Appendix B for a complete overview of the classification times per truth table case (TT, TF, FT, FF) and per type of conditional (AA, AN, NA and NN). Testing the key predictions of the three different theories, we compared, as was the case in the previous experiments, the time it took an individual participant to classify the TT case as ‘true’ on the one hand with the classification time of the FA cases FT and FF as ‘true’ and as ‘irrelevant’ on the other hand. We averaged the classification times over participants and participants who did not produce both of the judgments under investigation were left out of the analysis. T-tests revealed the following results:

Classification times for TT as ‘true’ ($M = 2509$ ms, $SD = 813$ ms) were shorter than the classification times for FT as ‘true’ ($M = 3379$ ms, $SD = 1261$ ms, $t(220) = -6.11$, $p < .01$). Similarly, classification times for TT as ‘true’ ($M = 2469$ ms, $SD = 854$ ms) were shorter than for FF as ‘true’ ($M = 3284$ ms, $SD = 1251$ ms, $t(306) = -6.68$, $p < .01$). Finally,

the judgment of TT as ‘true’ ($M = 2433$ ms, $SD = 783$ ms) was faster than the judgment of FT as ‘irrelevant’ ($M = 2745$ ms, $SD = 1133$ ms, $t(384) = -3.14$, $p < .01$). The same was true for TT as ‘true’ ($M = 2474$ ms, $SD = 808$ ms) versus FF as ‘irrelevant’ ($M = 2884$ ms, $SD = 1112$ ms, $t(408) = -4.27$, $p < .01$). Finally, judging TT as ‘true’ ($M = 2459$ ms, $SD = 825$ ms) took equally long than judging TF as ‘false’ ($M = 2580$ ms, $SD = 879$ ms, $t(446) = -1.51$, $p = .13$).

Focusing on the AA conditional “If A then C”, classification times were identical compared to the averaged results described above (see Appendix B): Classification times of TT as ‘true’ were significantly shorter than those of the FA cases.

With respect to these judgments of these FA cases, we again focussed on the classification times of the predominant answer patterns for the AA conditional (see Appendix A and Appendix B). For the *defective conditional reasoners* (TFII), judging FF as ‘irrelevant’ ($M = 2218$ ms, $SD = 1012$ ms) was significantly faster than judging FT as ‘irrelevant’ ($M = 2989$ ms, $SD = 1281$ ms; $t(176) = 4.45$, $p < .01$). The classification time of TT as ‘true’ ($M = 2020$ ms, $SD = 895$ ms) was equally long than that of FF as ‘irrelevant’ ($t(176) = -1.38$, $p = .17$). For *defective biconditional reasoners* (TFFI), the judgment of FT as ‘false’ ($M = 2449$ ms, $SD = 922$ ms) took the same time than judging FF as ‘irrelevant’ ($M = 2390$ ms, $SD = 1143$ ms; $t(128) = .33$, $p = .74$). The classification of TT as ‘true’ ($M = 1930$ ms, $SD = 861$ ms) was faster than both the classification of FT as ‘false’ ($t(128) = -3.32$, $p < .01$) and that of FF as ‘irrelevant’ ($t(128) = -2.59$, $p < .01$).

Discussion

The global pattern of results is identical to Experiment 1 and 2. Firstly, the majority of answer patterns is three valued, meaning that the TT case is judged as ‘true’, the TF case as ‘false’ and the FA cases as ‘irrelevant’. This is in line with the ST as well as the R-MMT.

The pattern of correlations is also an exact replication of the pattern observed in the previous experiments of this study and considered to be evidence in support of the ST and the R-MMT, predicting a positive correlation between cognitive ability and the tendency to classify truth table cases along a three-valued or three-valued truth table pattern. The repeated absence of a correlation between cognitive ability the MI-Index is in turn problematic for the C-MMT.

Regarding the effect of our working memory load manipulation, time pressure on the truth table task does not alter the fact that burdening participants' working memory capacity has no effect on their truth table classifications, even though it is observed that in the high-load condition performance on the dual task is worse than in Experiment 2. This implies that we do not find support for our hypothesis that the lack of an effect of working memory load in the two previous experiments is due to rehearsal of the secondary task during the truth table judgments.

Taking the classification times into consideration, we observe once more results that are exactly the same as in Experiment 1 and in Experiment 2. The judgment of the FA cases as 'irrelevant' is significantly slower than the judgment of the TT case as 'true'. This is in line with both mental models accounts. Our data show, however, that only a minority of participants responds according to a two-valued TFFT answer pattern as predicted by the C-MMT. Moreover, the specific predictions of the R-MMT regarding the difference between conditional reasoners and conjunctive reasoners regarding their classification times for FT and FF are again confirmed. This implies that the theory that accounts best for our data is the R-MMT.

General Discussion

The primary aim of the present study is to examine the relation between cognitive ability and the truth table task, this in order to shed light on the mental representation of conditionals. We are fully aware of the restriction that the abstract conditionals we used in the present study are rare in daily life. Nevertheless, it has been shown previously (Evans et al., 2008) that even with everyday contents, similar answer patterns and correlations with intelligence are observed. We are therefore confident that our results make a valuable contribution to the literature examining the mental representation of conditionals.

There are two novel elements to the present experiments. Firstly, we are not only interested in the correlation between cognitive ability and the judgments yielded by the truth table task. Instead we experimentally manipulate working memory load by means of a dual task paradigm. Secondly, instead of solely inspecting participants' answer patterns, as is the case in many previous studies (Barrouillet, Gauffroy & Lecas, 2008; Evans et al., 2007, 2008; Sevenants et al., 2011), we combine this with an analysis of the classification times needed to judge the truth table cases.

The main theories explaining the mental representation of conditionals have opposite predictions regarding the correlation between cognitive ability and the interpretation of FA cases. According to the C-MMT, the people with the highest working memory capacity are able to look beyond the initial model ($A \rightarrow C$) and consider all possibilities compatible with the conditional, which results in two-valued answer patterns. The ST and the R-MMT, however, argue that the participants who are responding according to these three-valued patterns are those with the highest intelligence scores. Both theories have in common that people come to understand 'if A then C' at the highest or oldest stage of development, as being such that $P(\text{if } A \text{ then } C) = P(C|A)$. This in turn results in a three-valued truth table. According to the ST, the intelligent participants are able to complete both stages of the Ramsey-Test, thereby considering both $A \rightarrow C$ and $A \rightarrow \neg C$, and judging the FA cases to be 'irrelevant'

According to the R-MMT, the intelligent participants are able to flesh out the initial model. Because the corresponding models do not pertain to the core meaning of the conditional (i.e., the initial model), but are not false because these situations are compatible with the conditional and do not contradict it, they are judged ‘irrelevant’.

As mentioned above, the present study focuses on the classification times of the truth table task. Since an identical pattern of classification times is observed in three different large scale experiments, we are convinced that what we observed is not just a coincident. The dominant observation is that the majority of the participants responded according to a three-valued answer pattern and that the judgment of the TT case as ‘true’ is faster than the judgment of the FA cases as ‘irrelevant’. The global picture of the answer patterns along with their classification times is supportive for the R-MMT.

The results observed in the present study are problematic for the C-MMT for two reasons. First, the TFFT answer pattern, which is predicted for people able to look behind the initial model, is rarely observed. This is especially problematic since our participants are all drawn from a population of university students which are all considered to have above-average intellectual capacities. Second, a negative or no correlation was observed between intelligence and the classification of FA cases as ‘true’. This is problematic for the C-MMT because this is the judgment expected from the most intelligent reasoners.

The classification times we observe are also problematic for the ST. This account must predict very fast ‘irrelevant’ judgments for the FA cases since these are considered to be heuristic in nature (see Evans, 2007). The fact that the ST predicts ‘irrelevant’ judgments of the FA cases for the most intelligent participants -which is also observed in our data- is also conflicting with the claim that these irrelevance judgments are based on a heuristic and thus non-demanding process (see also Barrouillet (2011) for this contradiction). It has to be noted, however, that the ST makes no predictions itself about classification times in the truth table

task, especially when FA cases are explicitly brought to the attention of the participants. This goes back to our point in the introduction that these cases might not be ‘irrelevant’ in an absolute sense: When an irrelevant matter is sprung on us during a discussion or problem solving, it gives us pause, if only because we are startled or surprised, and we take time to reflect on it. Taking this into account, one has to distinguish (David Over, personal communication) the fundamental ‘*semantic ST*’ from its peripheral heuristic/analytic ‘*processing model*’ (Evans, 2007; see also Barrouillet, 2011, on this point). In this ‘processing model’, when people are evaluating $P(\text{if } A \text{ then } C)$, a heuristic automatically directs their attention to the true-antecedent cases in a Ramsey test, and then an analytic process can take over to assess, sometimes very quickly and sometimes after much reflection, the extent to which C is likely to follow given A . The fact that FA cases might be not ‘irrelevant’ in an absolute sense, should give rise to a *Revised Suppositional Theory* or at least a ‘*revised processing model*’. In this revised version, an ‘irrelevant’ judgment of the FA cases could not necessarily be reached through a heuristic (and thus non-consuming) process but might involve some analytical reasoning. Moreover, it should be allowed that the judgment of the true-antecedent cases (TT and TF) is based on automated (heuristic) rather than analytical processes. Evans (2004, p. 24) already suggests thinking of the Ramsey test as a process that can sometimes largely take place in System 1, grounded in heuristics that are automatic and implicit (Over & Evans, 2003). A processing model closer to this account than to the model described by Evans (2007) would better fit our results.

Next to the examination of the classification times, instead of solely inspecting the correlation between cognitive ability and the responses yielded by a truth table task, we also manipulate working memory capacity by means of a dual task paradigm, which is completely novel in combination with the truth table task. Participants judge the truth table cases while their executive resources are burdened by a secondary task. Making use of a differential load

provides a direct measure of the impact of working memory capacity on the response patterns yielded by the truth table task. In all three experiments we observe a positive correlation between cognitive ability and the judgment of the FA cases as irrelevant, but no effect at all of the working memory load manipulation. Two different hypotheses were tested in order to explain this observation. In Experiment 2 we ruled out the hypothesis that the lack of an effect of the dot memory task is due to the visuospatial nature of this task, which might not burden the resources needed to solve a truth table task more verbal in nature: Making use of a verbal instead of a visuospatial load did not alter the fact that no effect of working-memory load is observed. In Experiment 3 the rehearsal-hypothesis was ruled out, which states that during the self-paced truth table task, participants might have used the time between seeing and classifying the case to rehearse the words presented in the secondary task. In order to test this rehearsal-hypothesis, we put time pressure on the truth table task, preventing participants from rehearsing the words from the secondary task during the classification time. This manipulation however did not have an effect: As was the case in Experiment 1 and 2, a positive correlation between the classification of FA cases as ‘irrelevant’ is observed, but the answer patterns are not altered by the presence of a dual task burdening participant’s working memory resources. It has to be noted, however, that switching between processing and storage as proposed in the Time-Based Resource-Sharing model (Barouillet et al., 2004) is thought to occur on a micro-level (as short as 40ms). That is, during processing, the focus of attention switches very quickly to maintenance and back so to refresh information to be maintained (i.e., the dots or words from the dual task). Therefore, it might be argued that a limit at 6.5 seconds cannot prevent such a micro-level switching.

A third hypothesis might be that the dual working memory load tasks we used was not difficult enough in order to burden the resources of our ‘intelligent’ university students. It has been shown that verbal information can be maintained by two different mechanisms (subvocal

rehearsal and attentional refreshing, Camos, Lagner, & Barrouillet, 2009; Hudjetz & Oberauer, 2007), only one of them being attentionally demanding. A load of four words (or of four dots if participants used verbal labels to recall dot positions) therefore might have been too small to draw on the general purpose resources considered to be involved in reasoning. We do not believe, however that a more difficult load-task will lead to differences regarding the answer patterns because our working memory load manipulation did not have any effect at all: Even comparing the participants with the least cognitive ability in the high-load conditions with the high-ability participants in the no-load conditions, no difference was observed. Anyway, further research is needed in order to examine the lack of an effect of working memory load in the truth table task.

In sum, combining the correlation between cognitive ability and the truth table judgments, the (lack of a) working memory load effect and the classification times, we conclude that the R-MMT provides the best account for the mechanism underlying the mental representation of conditionals.

References

- Baddeley, A. D. & Logie, R. H. (1999). Working Memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 28-61). New York: Cambridge University Press.
- Ball, L. J., Lucas, E. J., Miles, J. N. V, & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, 56, 1053-1077.
- Barrouillet, P. (2011) Dual-process theories of reasoning: The test of development. *Developmental Review*, 31, 151-179.
- Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, 133, 83-100.
- Barrouillet, P., & Lecas, J. F. (1998). How can Mental Models Theory account for content effects in conditional reasoning? A developmental perspective. *Cognition*, 67, 209-253.
- Barrouillet, P., Gauffroy, C., & Lecas, J. F. (2008). Mental models and the suppositional account of conditionals. *Psychological Review*, 115, 760-771.
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 12-23.
- Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, 58, 382-398.
- Camos, V., Lagner, P. & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, 61, 457-469.

- Cattell, J. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1-22.
- de Finetti, B. (1967). Sur quelques conventions qui semblent utiles. [On some conventions that seem useful]. *Revue Roumaine de Mathématiques Pures et Appliquées*, xii, 1227-1234.
- de Finetti, B. (2008). Philosophical Lectures on Probability collected, edited, and annotated by Alberto Mura (Vol. 340). Berlin: Springer.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54, 128-133.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning*, 11, 349-381.
- De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, 42, 177-190.
- De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, 53, 123-131.
- Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, 87, 223-240.
- Evans, J. St. B. T. (2006). The heuristic–analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, 13, 378–395.
- Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. Hove, UK: Psychology Press.
- Evans, J. St. B. T., & Newstead, S. E. (1977). Language and reasoning: A study of temporal factors. *Cognition*, 5, 265-283.

- Evans, J. St. B. T., Over, D., (2004). *If*. Oxford University Press.
- Evans, J. St. B. T., Over, D. E., & Handley, S. H. (2003). A theory of hypothetical thinking. In D. Hardman & L. Maachi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 3-22). Chichester: Wiley.
- Evans, J. St. B. T., Handley, S. H., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition*, 35, 1722-1784.
- Evans, J. St. B. T., Handley, S. H., Neilens, H., & Over, D. E. (2008). Understanding causal conditionals: A study of individual differences. *Quarterly Journal of Experimental Psychology*, 61, 1291-1297.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25-42.
- Gauffroy, C., & Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: An integrative developmental theory. *Developmental Review*, 29, 249-282.
- Gauffroy, C., & Barrouillet, P. (2011). The primacy of thinking about possibilities in the development of reasoning. *Developmental Psychology*, 47, 1000-1011.
- Gilio, A. & Over, D. (2012). The psychology of inferring conditionals from disjunctions: A probabilistic study. *Journal of Mathematical Psychology* 56, 118–131.
- Giroto, V. & Johnson-Laird, P.N. (2004). The probability of conditionals. *Psychologia* (special issue on Reasoning), 47, 207-225.
- Hudjetz, A., & Oberauer, K. (2007). The effects of processing time and processing rate on forgetting in working memory: Testing four models of the complex span paradigm. *Memory & Cognition*, 35, 1675-1684.
- Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646-678.
- Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, 99, 418-439.
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, 130, 621-640.
- Navarrete, G., De Neys, W., & Santamaria, C. (2008, July). Not so natural after all: Intelligence always matters. Paper presented at the XXIX International Congress of Psychology, Berlin, Germany.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Politzer, G., Over, D., & Baratgin, J. (2010). Betting on conditionals. *Thinking and Reasoning*, 3, 172 – 197.
- Raven, J. C., Court, J. H., & Raven, J. (1988). *Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 4)*. London: H. K. Lewis.
- Raven, J. C., Raven, J., & Court, J. H. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 1)*. Oxford, UK: Oxford Psychologists Press.
- Ramsey, F. P. (1990). General propositions and causality (original publication, 1931). In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145-163). Cambridge: Cambridge University Press.
- Roberts, M. J. (1998a). Inspection Times and the selection task: Are they relevant? *Quarterly Journal of Experimental Psychology*, 51, 781-810.

- Roberts, M. J. (1998b). How should relevance be defined? What does inspection time measure? A reply to Evans. *Quarterly Journal of Experimental Psychology*, 51, 815-817.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645–726.
- Schroyens, W. (2010). A meta-analytic review of thinking about what is true, possible, and irrelevant in reasoning from or reasoning about conditional propositions. *European Journal of Cognitive Psychology*, 22, 897-921.
- Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1713-1734.
- Sevenants, A., Dieussaert, K., Schaeken, W. & Barrouillet, P. (2009). Working memory and the truth table task. Unpublished manuscript.
- Sevenants, A., Dieussaert, K., Schaeken, W. (2011). Truth Table Tasks: Irrelevance and Cognitive Ability. *Thinking & Reasoning*, 17, 213 - 246.
- Sevenants, A., Dieussaert, K., & Schaeken, W. (2012). Is the Truth Table Task Mistaken? *Thinking & Reasoning*, 18 (2), 119-132.
- Stanovich, K. E. (2011). Rationality and the reflective mind. New York: Oxford University Press.
- Over, D. E. & Evans, J. St. B. T. (2003). The probability of conditionals: the psychological evidence. *Mind & Language*, 18, 340-358.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39, 1275-1289.
- Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-

general resource? *Psychological Science*, 21, 384-390.

Vergauwe, E., Gauffroy, C., Morsanyi, K., Dagry, I, & Barrouillet, P. (in press).

Chronometric evidence for the dual-process mental model theory of conditional.

Journal of Cognitive Psychology.

Wason, P. C. (1966). *Reasoning*. In B. M. Foss (Ed.), *New horizons in psychology*, Vol. I.

Harmondsworth, UK: Penguin.