Ψ **Psychology Press**
Taylor & Francis Group

Defective truth tables and falsifying cards: Two measurement models yield no evidence of an underlying fleshing-out propensity

Jean-François Bonnefon
CNRS and Université de Toulouse, France

Stéphane Vautier
Université de Toulouse, France

Using a latent variable modelling strategy we study individual differences in patterns of answers to the selection task and to the truth table task. Specifically we investigate the prediction of mental model theory according to which the individual tendency to select the false consequent card (in the selection task) is negatively correlated with the tendency to judge the false antecedent cases as irrelevant (in the truth table task). We fit a psychometric model to two large samples ($N = 486$, twice), and find no evidence for this negative correlation. We examine which of the assumptions of the model theory must be amended to accommodate our findings.

Keywords: Conditional reasoning; Confirmatory factor analysis; Individual differences; Mental models.

Conditional reasoning (i.e., the cognitive manipulation of "if *p*, then *q*" relations) is the cornerstone of hypothetical thinking (Evans, 2007), and has generated a massive experimental literature over the last four decades. Traditionally this literature has focused on a small number of core tasks, and on explaining the *modal* answer to these tasks. In recent years, however, a number of authors have emphasised the existence of individual differences in the way people represent conditional statements and reason from them

(e.g., Bonnefon, Eid, Vautier, & Jmel, 2008; Bonnefon, Vautier, & Eid, 2007; Evans, Handley, Neilens, & Over, 2007, in press; Newstead, Handley, Harley, Wright, & Farelly, 2004; Oberauer, Geiger, Fischer, & Weidenfeld, 2007).

As a consequence a new literature has emerged, which is concerned with explaining *variation* in the responses to conditional reasoning tasks, as the result of quantitative and qualitative differences between reasoners. A general strategy within that line of research is to identify individual patterns of responses across different tasks, and to compare these patterns to the predictions of current theories of conditional reasoning. One concern with that strategy, though, is that modest correlations across different tasks might reflect the poor psychometric qualities of the scores derived from the tasks (Bonnefon et al., 2007; Stanovich, 1999). This concerns calls for the use of psychometric modelling techniques that can correct correlation estimation for measurement error.

In this article we target a specific prediction of the mental model theory of conditional reasoning (Johnson-Laird & Byrne, 2002). Mental model theory predicts that variations in responses to conditional reasoning tasks partly depend on whether individuals "flesh out" their initial representation of the conditional. We argue that mental model theory has to commit to the following theses: (a) individuals differ in their propensity to flesh out their initial models of a conditional; (b) this propensity positively impacts their tendency to choose the falsifying card $\neg q$ in the indicative selection task; and (c) this propensity negatively impacts their tendency to give a defective truth table of the conditional. Therefore, mental model theory must predict a negative correlation between the individual tendency to give defective truth tables and the individual tendency to choose falsifying $\neg q$ cards.

We apply two psychometric models implementing these theses—one model designed for a categorical response format, and the other designed for a scale response format. The two models are fitted to responses given by two large samples of reasoners, allowing us to estimate the relevant correlations, corrected for measurement error.

THE FLESHING-OUT PROCESS

According to the mental model theory of conditionals (Johnson-Laird & Byrne, 2002), a statement "if p, then q" is represented as the set of possibilities that make it true. The more economic representation of "if p then q" consists of one explicit model $<pq>$ and one implicit model $< \ldots >$, which serves as a mental footnote pointing to possibilities where p is false. This initial representation may be fully *fleshed out* into the three explicit models $<pq>$ $<\neg pq>$ $<\neg p \neg q>$, where $\neg$ is an operator for negation.

Even though the initial representation of the conditional includes a mental footnote pointing to the models $<\neg pq>$ and $<\neg p \neg q>$, individuals

are liable to disregard these possibilities, and indeed to forget about the implicit model altogether, when they mentally manipulate the conditional. Importantly for our present purpose, this tendency to disregard the implicit model is (a) linked to individual differences in working memory capacity, (b) predictive of responses to the selection task, and (c) predictive of responses to the truth table task. We now address in turn these three aspects of the fleshing-out process.

Individual variations

According to mental model theory, individuals start with the $<pq>$ $<...>$ representation because the $<pq>$ $<\neg pq>$ $<\neg p\neg q>$ representation places too high a burden on working memory (Johnson-Laird & Byrne, 2002; Johnson-Laird, Byrne, & Schaeken, 1992). Now, greater working memory capacity should alleviate this burden: Therefore, all other things being equal, individuals with greater working memory capacity should be comparatively more likely to flesh out their initial models of a conditional into a fully explicit, unpacked representation (Barouillet & Lecas, 1999).

> Thesis 1 (Individual propensity). *Individuals vary in their propensity to flesh out the initial model of a conditional (in positive relation to, e.g., their working memory capacity).*

Importantly, when they engage in a reasoning task individuals should reach different conclusions as a function of whether they worked from a fully fleshed-out representation of the conditional. Although this is true of all standard reasoning tasks, we now consider in detail two specific tasks.

The selection task

In the original form of the selection task (Wason, 1966), participants are presented with the pictures of four cards, and told that the cards all have a letter on one side and a number on the other side. Two of the cards display their lettered side (e.g., *A* and *B*), and the two other cards display their numbered side (e.g., 2 and 5). Participants are instructed to choose those cards and only those cards that need to be turned over in order to decide whether the following rule is true: "If there is a vowel on one side, then there is an even number on the other side." The *A* and 5 cards are the only ones that can yield a vowel–odd number falsifying combination, but most reasoners fail to select the 5 card. This robust finding has generated a large experimental and theoretical literature (Klauer, Stahl, & Erdfelder, 2007), aimed at explaining why and when reasoners select the critical $\neg q$ card. Importantly for our present purpose, mental model theory predicts that reasoners who flesh out their initial models of the conditional are in a better

position to select the $\neg q$ card (Johnson-Laird & Byrne, 1991, 1992, 2002).[1] In the words of Johnson-Laird and Byrne (1992, pp. 176–177): "The model theory predicts that any experimental manipulation that leads subjects to flesh out their models explicitly should also lead them to select the card representing the negated consequent of the conditional."

This is because the theory assumes that reasoners only consider for selection the cards that are represented in their models. Thus, from the models $<pq>\ <\dots>$, only cards p and q are available for selection—and reasoners are likely to select them both. However, from the models $<pq>$ $<\neg pq>\ <\neg p\neg q>$, all four cards are available for selection—although not all of them will be selected. The $\neg p$ card, for example, is present both in conjunction with q and with $\neg q$, and is likely not to be selected for that reason (and a similar reasoning applies to the q card).

> Thesis 2 (Selection task). *Individuals who flesh out their initial model of a conditional are more likely to select the $\neg q$ card in the Wason selection task.*

The truth table task

The truth table task consists of judging, for each of the combinations pq, $p\neg q$, $\neg pq$, and $\neg p\neg q$, whether it makes the conditional true (T), false (F), or whether it is irrelevant (I) to the truth value of the conditional. The four responses given to the four situations together form a "pattern" of answers to the truth table task. For example, the TFTT pattern corresponds to the answers "true" in the pq situation, "false" in the $p\neg q$ situation, "true" in the $\neg pq$ situation, and "true" in the $\neg p\neg q$ situation. This pattern is also understood as corresponding to a material interpretation of the conditional. The TFFT pattern corresponds to a biconditional interpretation, and the TFII pattern corresponds to a defective interpretation.

Mental model theory (Johnson-Laird & Byrne, 1991) assumes that the defective TFII pattern is given by reasoners who do not flesh out their initial model of the conditional. Since the $\neg p$ cases are not explicitly represented in this initial model, they are judged irrelevant to the truth value of the conditional.[2] However, reasoners who do flesh out their initial models into

---

[1]Mental model theory also considers other routes to the $\neg q$ card, and Johnson-Laird and Byrne (2002) insist in particular that one of these is the construction of an explicit counterexample to the rule.

[2]As observed by Oberauer et al. (2007, footnote 3), recent notational changes in the mental model theory (Johnson-Laird & Byrne, 2002) make it difficult to predict why the F case of the TFII pattern is not an I. We do not address this issue here, as we are concerned with the two uncontroversial I cases.

the explicit set $<pq>$ $<\neg pq>$ $<\neg p\neg q>$ have no ground to judge the $\neg p$ cases irrelevant.

> Thesis 3 (Truth table task). *Individuals who flesh out their initial model of a conditional are less likely to judge the $\neg p$ cases as irrelevant.*

Summary and consequences

We have argued that mental model theory commits to three essential theses concerning the fleshing-out process: (a) individuals vary in their propensity to flesh out their initial model; (b) individuals who flesh out their initial model are more likely to select the $\neg q$ card; and (c) individuals who flesh out their initial model are less likely to judge the $\neg p$ cases as irrelevant.

These theses have observable consequences from the perspective of individual differences. If the theory is correct, we should be able to model individual differences in the tendency to select the $\neg q$ card, as well as in the tendency to judge the $\neg p$ cases as irrelevant, and to find out that these two tendencies show *negative covariation*—because they both relate to an underlying tendency to flesh out the initial model of a conditional, albeit in opposite directions. If really individuals differ in their tendency to flesh out their initial models $<pq>$ $<\dots>$ into the explicit set $<pq>$ $<\neg pq>$ $<\neg p\neg q>$; if this tendency is (to some extent) responsible for selecting the $\neg q$ card; and if this tendency is (to some extent) responsible for not judging the $\neg p$ cases as irrelevant, then a psychometric model of individual differences in the selection task and the truth table task should display a negative covariation between the factor standing for the tendency to select the $\neg q$ card, and the factor standing for the tendency to judge the $\neg p$ cases as irrelevant.

PSYCHOMETRIC MODELLING

A multivariate measurement design allows the modelling of reliable individual differences related to (a) the tendency to select the $\neg q$ card, and (b) the tendency to judge the $\neg p$ cases as irrelevant. In a confirmatory factor-analytic model, these two tendencies can be modelled through two correlated continuous latent variables, also called factors, as shown in Figure 1. The model has two correlated factors s (for selection task) and t (for truth table task), which are each measured by a set of indicators, respectively Sis and Tjs. Si indicates the tendency to select the $\neg q$ card for item i, while Tj indicates the tendency to judge the $\neg p$ case as irrelevant for item j. The latent variables $eS\,i$ and $eT\,j$ represent measurement error. We assume that measurement error variables are uncorrelated.

Confirmatory factor analysis enables to test the statistical plausibility of the best-fitting model. If the best-fitting model holds, its estimated
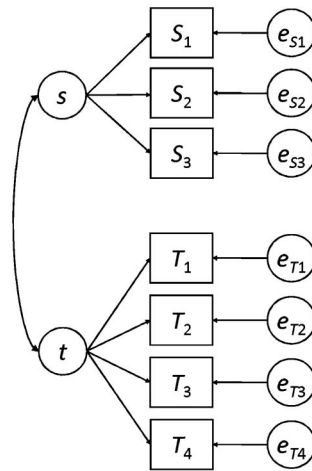
Figure 1. Path diagram of the confirmatory factor analytic model used to assess the correlation between the tendency s to select the $\neg q$ card and the tendency t to judge the $\neg p$ cases as irrelevant.

parameters can be used to investigate quantitative issues like reliability and correlation (corrected for measurement error). The most important issue for our purpose is the assessment of the correlation $\rho(s, t)$ between the two factors s and t. Mental model theory predicts that the value of the correlation $\rho(s, t)$ is negative. As measurement error is explicitly removed from the factors, there is no need to correct the estimate of $\rho(s, t)$ for attenuation due to measurement error. If the 95% confidence interval for $\rho(s, t)$ does not include negative values, the hypothesis of a negative covariation between the two factors can be rejected.

Taking advantage of the fact that this modelling strategy can accommodate categorical as well as continuous response formats, we conducted two independent studies using two different responses formats. We now describe these two studies and the model estimations they yielded. We then provide a joint discussion of the results of the two studies.

STUDY 1: CATEGORICAL RESPONSES

Method

Participants.   A total of 486 participants (248 men; mean age 30.9 years, $SD = 12.5$) were recruited by third-year psychology students as a course requirement. Each student made a list of several men and women who were older than 18, not studying psychology, and willing to take part in a survey

on reasoning—no other restriction applied, e.g., family members were permitted. Each student then randomly selected one male and one female participant from this list. It was expected that this recruitment procedure would promote variety in age, occupation, and education, while ensuring similar proportions of male and female participants. No incentive was offered to participants. In the rare cases when a randomly selected participant did not consent to take part in the survey, the student made a second random selection from his or her list.

Materials.   Participants solved three indicative selection tasks and two indicative truth table tasks, yielding three indicators of the tendency to select the $\neg q$ card, and four indicators of the tendency to judge $\neg p$ cases as irrelevant. The contents of all tasks were taken from Thompson (2000). Here is one example of selection task:

> You work as a bank officer. According to the interest rate formula: If the interest rates average at least 10%, then the value of a savings account doubles in less than 10 years. You wish to check out that this is indeed the case.

Participants judged the usefulness of selecting the p, $\neg p$, q, and $\neg q$ cases, using a categorical response format. The $\neg q$ question was phrased thus:

> You know of a saving account whose value did not double in less than 10 years. Is it useful to check out whether its interest rate averaged at least 10%?

Response options were "useful" (coded 1) and "not useful" (coded 0). The two other tasks featured the rules "If the interest rates are high, then many small businesses go bankrupt", and "If a person works in contact with asbestos, then they will develop lung cancer".

Here is an example of a truth table task:

> You work in a cancer clinic. According to health surveys: If a person is exposed to radiation, then they will develop cancer. Do the following situations conform to that rule?

Participants then considered the four situations pq, $p\neg q$, $\neg pq$, and $\neg p\neg q$. The situations $\neg pq$ and $\neg p\neg q$ were phrased thus, respectively: "a person is not exposed to radiation and develops cancer"; "a person is not exposed to radiation and does not develop cancer".

Response options were "the situation conforms to the rule" (coded 0), "the situation does not conform to the rule" (coded 0) and "none of the above" (coded 1). The other truth table task featured the rule: "If the licence plate number starts with 'M', then it is registered to a diplomat."

Analysis.   As the factor indicators are binary variables, the model was estimated by using a robust weighted least squares estimator as implemented in the Mplus software (Muthén & Muthén, 2007). With this estimator, the indicators are regressed on their factor by means of the probit function. The reliability of each indicator may be assessed by considering that each binary indicator reflects an underlying continuous indicator. The variance of the underlying indicator is decomposed as the sum of the variance accounted for by the factor, and the residual variance. Hence, the reliability of the indicator is its determination coefficient R^2.

Results

Although they are not the focus of our analysis, we provide some descriptive data before we move on to confirmatory factor analysis. Participants' responses followed an unsurprising pattern on the two tasks. In the truth table task, 14% of response patterns were of the TFTT conditional type, 24% were of the TFFT biconditional type, a total of 37% were of the defective (TFII) or semi-defective type (TFIT, TFTI, TFFI), and various irregular patterns accounted for the remaining 25%. Unsurprisingly, 92% participants chose the T response for the pq cases, and 80% chose the F response for the $p\neg q$ cases. The two other cases (critical to our present purpose) showed much more variation. In the $\neg pq$ case, the T, F, and I responses accounted for 28%, 38%, and 34% of responses, respectively. In the $\neg p\neg q$ cases, these frequencies were 65%, 4%, and 31%, respectively. Note in particular that a large enough subset of participants chose the I response for our correlational analysis to be reliable. Response patterns in the selection task were also unsurprising, with many conjunctive pq selections (28%) and biconditional $p\neg pq\neg q$ selections (21%), few conditional $p\neg q$ selections (4%), and even fewer single p card selections (3%). Note that in addition of being selected in the 21% all-card selections, the $\neg q$ card was selected, overall, in 47% of cases. This is in line with what Thompson (2000) observed using the same task and the same materials, and again it ensures that there is sufficient variability in our dependent variable for the correlational analysis to be reliable.

The uncorrected correlation between our two dependent variables is .02, which is already quite suggestive. Nevertheless, we need to ensure that this low correlation is not essentially the result of measurement error, for which we need to conduct a confirmatory factor analysis. The model fitted the data very closely, $\chi^2(N = 486, \ 11) = 17.47$, $p = .095$, RMSEA = 0.035. The reliability of the indicators Sis ranges from .26 to .43. The reliability of the indicators Tjs ranges from .23 to .79. These low reliability indices stress the need for the modelling of latent variables in order to correct the estimates for measurement error. The estimated value of the correlation

$\rho(s, t)$ is .054, with a 95% confidence interval of $-.12$ to $+.23$. The corrected correlation is thus more than twice the uncorrected correlation, but it remains extremely low. Although the hypothesis of a negative correlation cannot be rejected, this finding suggests that the correlation between the factors s and t, if different from zero at all, is more likely to be positive rather than negative.

STUDY 2: SCALE RESPONSES

Method

Participants.   Using a similar recruitment procedure as in Study 1, an independent sample of 486 participants was recruited for Study 2 (238 men; mean age 31.2 years, $SD = 13.1$).

Materials.   The materials used in Study 2 were similar to the materials used in Study 1, but the response options and the coding schemes were different. The $\neg q$ selection question was phrased thus (for the interest rate problem):

> You know of a saving account whose value did not double in less than 10 years. To what extent is it useful to check out whether its interest rate averaged at least 10%?

Participants answered on a 7-point scale anchored at *Not useful at all* and *Totally useful*, and their responses were accordingly coded from 1 to 7. The truth table questions were phrased: "To which extent do the following situations conform to the rule?" Participants answered on a 7-point scale anchored at *Not at all* and *Totally*. Because the tendency to judge a situation as irrelevant is maximal for the *midpoint* of the scale, responses were coded $<1, 2, 3, 4, 3, 2, 1>$ instead of $<1, 2, 3, 4, 5, 6, 7>$.

Analysis.   As the factor indicators are treated as continuous variables, the model was estimated by using maximum likelihood with robust standard errors as implemented in the Mplus software (Muthén & Muthén, 2007). With this estimator, the indicators are regressed on their factor by means of the linear function. The reliability of each indicator may be assessed by its determination coefficient R^2.

Results

Participants unambiguously judged that situation pq conformed to the rule ($M = 6.6$ on a scale from 1 to 7, $SD = 0.9$), and that situation $p\neg q$ did not ($M = 1.9$ on a scale from 1 to 7, $SD = 1.5$). Their judgments in the $\neg pq$ and

$\neg p \neg q$ cases were both more variable and closer to the midpoint of the scale ($M = 4.1$, $SD = 2.5$; and $M = 5.9$, $SD = 1.7$, respectively). In the selection task, participants showed a clear tendency to select the p and q cards ($M = 5.5$, $SD = 2.3$; and $M = 5.3$, $SD = 2.3$, respectively), and a tendency *not* to select the $\neg p$ card ($M = 2.3$, $SD = 2.3$). The greatest variability was observed in their tendency to select or not the $\neg q$ card ($M = 3.7$, $SD = 2.5$). The uncorrected correlation between our target variables is .02. The model fitted the data very closely, $\chi^2(N = 486, \ 13) = 20.13$, $p = .092$, RMSEA $= 0.034$. The reliability of the indicators S_is ranges from .18 to .20. The reliability of the indicators T_js ranges from .10 to .20. These results suggest that the raw indicators used in the continuous response format are markedly unreliable, and that there is a real need again for the modelling of latent variables in order to get estimates corrected for measurement error. The estimated value of the correlation $\rho(s, t)$ is .07, with a 95% confidence interval of $-.11$ to $+.25$. Although the hypothesis of a negative correlation cannot be rejected, this finding suggests that a negative correlation between the factors s and t is unlikely. Again, if at all different from zero, this correlation is more likely to be positive rather than negative.

DISCUSSION

Remarkably, the two studies yielded almost the same confidence interval for the correlation between the latent tendency to select the $\neg q$ case in the selection task and the latent tendency to judge the $\neg p$ cases as irrelevant in the truth table task. The two 95% confidence intervals include the value zero, and are centred on a slightly positive value, about .06 in both cases ($-.12$ to $+.23$ for categorical responses; $-.11$ to $+.25$ for scale responses). Thus the two studies strongly suggest that, even when it is corrected for measurement error, the correlation between the two tendencies is unlikely to be negative. The need for correcting for measurement error is made clear by the low reliability of the indicators, especially with the scale response format. As previously suggested (Bonnefon et al., 2007; Stanovich, 1999), reasoning tasks do not have the same psychometric qualities as standardised tests; and researchers must carefully consider the possibility that modest correlations across different tasks may be due to measurement error. Our strategy of modelling latent variables helps to alleviate this concern.

We note that the poor reliability of the indicators might be due to low inter-individual variability in the tendencies we sought to measure. This, however, would not bode well for mental model theory. Indeed, the variability in those tendencies is assumed to derive in part from variability in working memory capacity—and our recruitment procedure (that promotes large variations in age, education, and professional occupation) makes it likely that large variations exist in our sample in terms of general intelligence

and working memory capacity. However, to back up this claim we would need appropriate measurements of capacity, which we could not conveniently carry out in this study given our recruitment strategy. A stronger test of our case would be achieved if we could include measures of memory capacity as a covariate in our analyses, perhaps in conjunction with abstract tasks.

Indeed, our studies did not use basic, abstract conditionals, but rather conditionals featuring thematic indicative contents. As a consequence, mental model theory may explain away our results by appealing to the "principle of pragmatic modulation" (Johnson-Laird & Byrne, 2002). That is, it could be argued that the models participants constructed (and the responses they subsequently gave) were determined not by participants' propensity to flesh out their initial models, but by their background knowledge about the semantic contents of the conditionals. We acknowledge this possibility, but would find it unsatisfying from a metatheoretical point of view. By shifting all the burden of explanation to pragmatic modulation, this account would effectively deny any significant role to individual differences in the tendency/ability to flesh out initial models into explicit models. This would essentially amount to turning "mental model theory" into "pragmatic modulation theory". This is, we believe, a dangerous temptation. Indeed, the concern has been repeatedly raised that the mental model part of mental model theory may not actually explain much of reasoning, compared to the knowledge and semantic part (Bonnefon, 2004; Bonnefon & Hilton, 2004; Evans & Over, 2004; Evans, Over, & Handley, 2005). Since its introduction by Johnson-Laird and Byrne (2002), the principle of pragmatic modulation has mainly been used to shield the theory from falsification, and has not received to date, and to our knowledge, any convincing empirical support—perhaps, in fact, because it seems hardly falsifiable itself. Thus, we wish to consider, as a conclusion to this article, how mental model theory may try to accommodate our results without appealing to the controversial principle of pragmatic modulation.

PERSPECTIVES

As we laid bare in our introduction, the prediction of a negative correlation between the tendency to select the $\neg q$ case (in the selection task) and the tendency to judge the $\neg p$ cases as irrelevant (in the truth table task) derives from three theses: (a) Individuals vary in their propensity to flesh out their initial model; (b) Individuals who flesh out their initial model are more likely to select the $\neg q$ card; and (c) Individuals who flesh out their initial model are less likely to judge the $\neg p$ cases as irrelevant.

Our finding that the correlation is unlikely to be significantly negative implies that at least one of these three theses should be amended. The first thesis should arguably be maintained, for it is an essential tenet of mental

model theory. The second thesis might be the one mental model theory is ready to let go. Indeed, recent expositions of the model theory of the selection task (Johnson-Laird & Byrne, 2002), although they do not explicitly rule out the original fleshing out account, now choose to emphasise a different route to the $\neg q$ card—namely, the explicit consideration of the counterexample model $<p\neg q>$.

However, that does not mean that the third thesis is off the hook. Indeed, recent data (Evans et al., 2007) show that general intelligence (which draws heavily on working memory capacity) is *positively* correlated with the tendency to judge the $\neg p$ cases as irrelevant in the truth table task. Although the main purpose of Evans et al. (2007) was different from ours, their results partly speak to the same issue as ours. The main purpose of their study was to investigate the link between general intelligence, the conditional probability interpretation of conditional statements, and the defective truth table. In line with their predictions, they observed that individuals who reach a conditional probability interpretation of the conditional (measured through a probabilistic truth table task) tend to be higher in cognitive ability, and tend to give defective truth table tasks. This result is replicated with realistic conditionals in Evans et al. (in press). Because these results disprove Thesis 3, it might be that in order to adapt to recent findings of ours and of Evans and colleagues, mental model theory will have to abandon its original explanations of both the selection task and the truth table task.

To conclude on the perspectives of this research, we note that the suppositional theory of conditionals (Evans & Over, 2004) appears to be agnostic with respect to the correlation between the tendency to select the $\neg q$ card and the tendency to judge the $\neg p$ cases as irrelevant in the truth table task. The theory does predict that high-capacity individuals will tend to judge the $\neg p$ cases as irrelevant, a prediction that is supported by experimental results. It does not predict, anyway, any simple linear relation between cognitive capacity and the tendency to select the $\neg q$ card. Indeed, Evans and Over (2004) suggest that high ability is not sufficient to select the $\neg q$ card; an additional necessary condition for doing so being to overcome the attentional bias towards the matching cards. As it turns out, the data reported in Evans et al. (2007) suggest that the relation between ability and sensitivity to matching bias is curvilinear rather than simply linear, with higher sensitivity to matching bias at the middle of the ability continuum. If really ability is correlated with the construction of the defective truth table, but not necessarily with the selection of the falsifying $\neg q$ card, then there is no compelling reason to expect any correlation between behaviour on the two tasks.

REFERENCES

Barouillet, P., & Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning, 5*, 289–302.

Bonnefon, J. F. (2004). Reinstatement, floating conclusions, and the credulity of mental model reasoning. *Cognitive Science, 28*, 621–631.

Bonnefon, J. F., Eid, M., Vautier, S., & Jmel, S. (2008). A mixed Rasch model of dual-process conditional reasoning. *Quarterly Journal of Experimental Psychology, 61*, 809–824.

Bonnefon, J. F., & Hilton, D. J. (2004). Consequential conditionals: Invited and suppressed inferences from valued outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 28–37.

Bonnefon, J. F., Vautier, S., & Eid, M. (2007). Modelling individual differences in contrapositive reasoning with continuous latent state and trait variables. *Personality and Individual Differences, 42*, 1403–1413.

Evans, J. S. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. Hove, UK: Psychology Press.

Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory and Cognition, 35*, 1772–1784.

Evans, J. S. B. T., Handley, S. J., Neilens, H., & Over, D. E. (in press). Understanding causal conditionals: A study of individual differences. *Quarterly Journal of Experimental Psychology*.

Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review, 112*, 1040–1052.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Johnson-Laird, P. N., & Byrne, R. M. J. (1992). Modal reasoning, models, and Manktelow and Over. *Cognition, 43*, 173–182.

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.

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 680–703.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus version 5*. [computer program].

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farelly, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology, 57A*, 33–60.

Oberauer, K., Geiger, S. M., Fischer, K., & Weidenfeld, A. (2007). Two meanings of "if"? Individual differences in the interpretation of conditionals. *Quarterly Journal of Experimental Psychology, 60*, 790–819.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76*, 209–268.

Wason, P. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 106–137). Harmondsworth, UK: Penguin.