Journal of Experimental Psychology:
Learning, Memory, and Cognition
2007, Vol. 33, No. 4, 680–703

Copyright 2007 by the American Psychological Association
0278-7393/07/$12.00 DOI: 10.1037/0278-7393.33.4.680

The Abstract Selection Task: New Data and an Almost Comprehensive Model

Karl Christoph Klauer and Christoph Stahl
Albert-Ludwigs-Universität Freiburg

Edgar Erdfelder
University of Mannheim

A complete quantitative account of P. Wason's (1966) abstract selection task is proposed. The account takes the form of a mathematical model. It is assumed that some response patterns are caused by inferential reasoning, whereas other responses reflect cognitive processes that affect each card selection separately and independently of other card selections. The model parameters assess the contributions of different interpretational, inferential, and heuristic factors that jointly determine performance in the selection task. The interpretation of most of the model parameters in terms of these different factors is validated experimentally. This model of the selection task is the first to account for the observed frequencies of all 16 possible response patterns that can arise.

Keywords: reasoning, conditional reasoning, Wason selection task, rationality, multinomial model

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

In 1966, Peter Wason devised the famous four-card selection task, or Wason selection task (WST). Reasoners are told that cards have a number on one side and a letter on the other side. A rule is then introduced, such as "If there is an *A* on the letter side, then there is a *3* on the number side," along with four cards that represent instances of the antecedent and the consequent as well as instances of their negations on the visible sides. For example, the four cards might show *A*, *B*, *3*, and *4*. The reasoners' task is to decide which cards would have to be turned in order to test whether the rule is true or false. The selection task has instigated an enormous amount of research (for recent reviews, see Evans, Newstead, & Byrne, 1993, chap. 4; Evans & Over, 2004, chap. 5; Oaksford & Chater, 2003a).

One reason for this interest is the great difficulty of the task for human reasoners. Assuming that the above rule is understood as a conditional in which the antecedent is sufficient for the consequent, the logically correct response is to select the card showing an *A* and the card showing a *4*. However, the combination of these two cards is selected only by a small minority of participants, typically fewer than 10%, whereas by far the most frequent choices are to select the card with an *A* or the two cards showing an *A* and a *3*.

Sometimes selection behavior follows logical prescriptions much more closely when the rule is embedded in semantically rich or deontic contexts (e.g., Cheng & Holyoak, 1985; Cosmides, 1989). In the present article, however, we focus on basic determinants of card-selection behavior other than the semantic embedding of the rule and therefore on the abstract task.

Many findings suggest that some amount of reasoning is involved in the WST. As summarized by Evans and Over (2004, chap. 5), evidence for a role of reasoning stems from verbal protocols, from individual-differences data, and from data in which two rules with alternative antecedents are shown, among others. For example, Stanovich and West (1998) found a correlation between participants' general cognitive ability and selection behavior. Those who identify the logically correct solution are among the participants with the highest general ability scores (see also Newstead, Handley, Harley, Wright, & Farrelly, 2004).

There are, on the other hand, important findings that speak against a role for reasoning, including discrepancies between results obtained with the WST and conditional-inference tasks, inspection-times data for the cards in the WST, and data from the so-called negations paradigm. For example, in the negations paradigm (Evans & Lynch, 1973), rules with negated components such as "If there is an *A*, then there is not a *3*" are shown. It turns out that people more frequently select cards with characters that are explicitly mentioned in the rule than cards showing previously unmentioned characters. This phenomenon has been termed the *matching-bias effect*, and it may reflect the operation of nonanalytic heuristics (Evans, 1998).

The different theories of the WST are discussed in subsequent sections of this article. All of them account for the most frequent selection patterns (the *A* card alone or the *A* card and the *3* card) and for the individual card-selection frequencies in one way or another. Most studies have focused on these data.[1] There are, however, 16 possible selection patterns that can occur, and current theories are incomplete in that they do not account for the frequencies with which the different patterns are observed.

Karl Christoph Klauer and Christoph Stahl, Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany; Edgar Erdfelder, Lehrstuhl für Psychologie III, University of Mannheim.

The research reported in this article was supported by Grant Kl 614/31-1 from the Deutsche Forschungsgemeinschaft to Karl Christoph Klauer. We thank Sieghard Beller, Nick Chater, Jonathan Evans, and Mike Oaksford for helpful comments on a previous version of this article.

Correspondence concerning this article should be addressed to Karl Christoph Klauer at the Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, Freiburg D-79085, Germany. E-mail: klauer@psychologie.uni-freiburg.de

[1] Exceptions were Pollard (1985) and Oaksford and Chater (1994), who looked at pairwise associations between card selections, and Klauer (1999), who attempted to account for the rank order of selection patterns (see also Oaksford & Chater, 2003a; Perham & Oaksford, 2005).

The purpose of the present article is to develop and validate a mathematical model for the 16 possible selection patterns. Such a model is a complete account of the WST in the sense that it explains not only the modal selection patterns and the individual card-selection frequencies but also dependencies and contingencies of any order between the cards. None of the current accounts of selection behavior has approached this goal (but see Klauer, 1999, and Perham & Oaksford, 2005, for steps towards this goal).

In the next section, a family of models of the WST is formulated that is grounded in major theories of the WST. Model selection criteria are applied to identify the model that strikes the best compromise between statistical goodness of fit and parsimony in real data sets. Another statistical criterion considered is the model's goodness of fit. The winning model's psychological viability is then examined in a series of experiments. The experiments target the different model parameters one by one to validate their interpretation in terms of the postulated underlying cognitive processes.

The present article thereby identifies and psychologically validates a set of factors that, taken together in appropriate combination, account for WST performance completely and quantitatively. The resulting model is a member of the family of dual-process theories that has recently received increased attention in the reasoning literature (e.g., Evans, 2006, in press; Stanovich & West, 2000). That is, the model specifies how processes or systems of different natures interact to produce observed WST data without committing to strong assumptions about the precise representation-process pairs (e.g., in terms of rule theories or mental models) that realize the different processes or systems (although we sketch how algorithmic realizations might look like where appropriate).

Notation

In most instances, the rules considered in this article have the form "If p, then q"; for example, "If there is an *A* on the letter side of the card, then there is a *3* on the number side." Four cards are presented for selection showing, for example, *A*, *B*, *3*, and *4* on the visible side. These cards are referred to by their logical relationship to the propositions p and q in the rule as, respectively, the p card, the $\bar{p}$ card, the q card, and the $\bar{q}$ card. The probabilities of selecting one of these cards, irrespective of other card choices, are denoted by italicized letters as, respectively, p, $\bar{p}$, q, and $\bar{q}$. *A* selection of cards is $(x_p, x_{\bar{p}}, x_q, x_{\bar{q}})$, with $x_r = 0$ meaning that the r card was not selected, and $x_r = 1$ meaning that it was selected, where r is any of the four cards. Alternatively, the selection is referred to by the cards that were selected. For example, the so-called partial-insight pattern (Johnson-Laird & Wason, 1970) is denoted by $(p, q, \bar{q})$ or, alternatively, by (1, 0, 1, 1). There are 16 different patterns of card selections $(x_p, x_{\bar{p}}, x_q, x_{\bar{q}})$ that can occur.

Two valid and two invalid inferences are frequently studied in conditional reasoning given the major premise "If p, then q" and an additional minor premise.

Modus ponens (MP): Given p, it is concluded that q

Modus tollens (MT): Given $\bar{q}$, it is concluded that $\bar{p}$

Denial of the antecedent (DA): Given $\bar{p}$, it is concluded that $\bar{q}$

Affirmation of the consequent (AC): Given q, it is concluded that p.

For the rule "If p, then q," the rule "If q, then p" is called the *converse*; the rule "If not p, then not q" is called the *inverse*; and the rule "If not q, then not p" is called the *contrapositive*.

Models

A number of theories assume that individual card choices are stochastically independent events; an assumption that is embodied in what we call the *independence model*. Other theories of the WST assume that reasoners select those cards for which they derive testable consequences for the invisible side of the card, on the basis of different interpretations of the rule. This idea is formalized in the inference model. As suggested by Evans (1977) in his statistical theory of reasoning, one way to combine both ideas is to conceive of the independence model and the inference model as alternative processing paths in a more general reasoning model, which we call the *inference-guessing model*. Yet another integrative model was originally suggested by Evans (1984) in his heuristic-analytic theory. In this framework, the independence model captures heuristic processes that act as a relevance filter determining the input to the inference mechanisms in the second stage of the reasoning process. Because the inference-guessing model and the heuristic-analytic model do not contradict each other, it is of course possible to integrate them in the framework of a model providing a relevance filter in the first stage of reasoning followed by inference and guessing as alternative processing paths in the second stage. This relevance-inference-guessing model is the most general of the models that we consider in the present article.

The purpose of the present article is to identify and validate a model accounting for the frequencies with which each of the possible 16 selection patterns occurs. The models considered in this section are candidates for such a quantitative account, and two statistical criteria, model selection and model fit, are used to test whether one of them provides an adequate approximate description of the data. The interpretation of the winning model's parameters in terms of underlying cognitive processes is then tested for (almost) all of the parameters in separate experiments.

Setting aside this primary purpose, how do the model analyses bear on extant theories of the WST? First, with the exception of the optimal data selection (ODS) approach (Oaksford & Chater, 1994, 2003a), none of the existing theories is specified to the point where quantitative predictions for the selection patterns are possible. Only the ODS model is therefore a direct competitor of the model developed here, and we also present direct comparisons of the present account and the account by ODS.

Second, as already outlined above, the inference-guessing model and the heuristic-analytic model can be seen as different instances of dual-process or dual-systems models originally proposed by Evans (1977, 1984). In such models, heuristic processes interact with analytic processes in determining performance in reasoning tasks. As discussed by Evans (in press), quantitative specifications of dual-process models must consist of (a) submodels representing heuristic processes, (b) submodels representing analytic processes, and (c) a model of how the submodels interact. In the heuristic-analytic model, the independence model, repre-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

senting heuristic processes, determines the contents on which inferential processes, represented by the inference model, operate. It can therefore be seen as a quantitative specification of Evans' (1984) heuristic-analytic theory, as elaborated further below. In contrast, the inference-guessing model represents heuristic and analytic processes as alternative processing paths, each of which is taken with a certain probability. This removes the strictly sequential nature of the interaction of heuristic and analytic processes built into the heuristic-analytic model and is consistent with Evans' (2006) revised version of the heuristic-analytic theory.

Thus, the inference-guessing model and the heuristic-analytic model can be seen as quantitative specifications of existing dual-process theories of the WST. As already noted, these theories have previously not been specified to the point where quantitative predictions for the selection patterns are possible. This means that it cannot be tested whether these theories are capable of explaining WST data beyond accounting for a few salient findings and effects such as the high frequency of certain patterns and the low frequency of others. If the inference-guessing model or the heuristic-analytic model succeeds in accounting for the data, the new conclusion can be drawn that the underlying theory is consistent with the WST data in the sense that a specification of it exists that can account for the data. On the other hand, if the model must be refuted, not as much is won, because other specifications of the theory, using other sets of auxiliary assumptions in specifying the theory, may exist that provide better accounts of the data. This latter point is a general limitation of model-comparison studies in the reasoning field (e.g., Oaksford & Chater, 2003a; Oberauer, 2006).

The Independence Model

Independent card choices can arise as a consequence of card selections being driven by (a) cardwise relevance judgments, (b) expected information gain as in the recent formulation of ODS (e.g., Hattori, 2002; Oaksford & Chater, 2003a), and (c) by independent guessing.

Evans (1984, 1989, 1995) has proposed that card selections reflect relevance judgments governed by heuristic cues. In this view, a card is selected if and only if it is perceived as relevant. A card's relevance is determined to a large extent by two heuristics operating automatically and preattentionally: According to the *if* heuristic, the p card will be highly relevant because it is referred to in the *if* part of the rule; according to the *matching* heuristic, cards showing characters that are mentioned in the rule will be more relevant than cards showing unmentioned characters. Although the theory has not been cast in mathematical form, it suggests as one possibility a model in which the individual card-selection probabilities are monotone functions of perceived relevance and combine via stochastical independence to account for card-selection patterns.

The first model to make the independence assumption explicitly is Evans' (1977) statistical theory of reasoning (see also Krauth, 1982). The independence assumption is also built into recent statements of the account by ODS (Hattori, 2002; Oaksford & Chater, 2003a). According to that account, problem solvers see the WST as a task of selecting the most informative pieces of data for deciding between two statistical hypothesis, namely the hypothesis that the conditional rule is valid and a null hypothesis postulating independence of p events and q events. The information value of each card is quantified by an expected information gain statistic.

Finally, guessing comes into play when reasoners are uncertain about the appropriate response but have to make a response nevertheless. In independent guessing, decision makers guess for each card separately and independently whether it has to be selected. The cards may nevertheless have different probabilities of being chosen depending upon pragmatic cues and surface characteristics of the cards themselves but also depending on the wider task context. We refer to the model of independent card choices as the *independence model*.

The independence model thus can stand for processes of different kinds: Preattentional heuristics determining a card's relevance, processes determining a card's potential information gain as in ODS, and heuristics involved in making more or less informed guesses about whether a card should be selected or not. The commonality of these processes is that they have been argued to, or can plausibly be assumed to, occur for each card locally and independently of other cards. The independence model acquires the flexibility to stand for these different processes, because only the final output of the processes is modeled in terms of different cardwise selection probabilities. The independence model in fact allows the cardwise selection probabilities to take on any value. Its major assumption is that the processes that determine the cards' selection probabilities run for each card locally and independently of the other cards.

Formally, the independence assumption states that the probability, $P[(x_p, x_{\bar{p}}, x_q, x_{\bar{q}})]$, of selecting the card combination $(x_p, x_{\bar{p}}, x_q, x_{\bar{q}})$ can be expressed by means of the individual card-selection probabilities as the following product:

$$P[(x_p, x_{\bar{p}}, x_q, x_{\bar{q}})] = p^{x_p}(1-p)^{(1-x_p)}\bar{p}^{x_{\bar{p}}}(1-\bar{p})^{(1-x_{\bar{p}})} \times q^{x_q}(1-q)^{(1-x_q)}\bar{q}^{x_{\bar{q}}}(1-\bar{q})^{(1-x_{\bar{q}})}$$

This basic model uses four parameters to describe the 16 pattern probabilities and thereby provides a relatively parsimonious model of the pattern frequencies.

The Inference Model

If can be interpreted in different ways (e.g., Evans & Over, 2004; Johnson-Laird & Byrne, 2002). Regarding the WST, a number of researchers have argued that a large variety of different misunderstandings of the rule accounts for performance and individual differences in the WST (e.g., Beattie & Baron, 1988; Gebauer & Laming, 1997; Margolis, 1987; Osman & Laming, 2001).

Comprehension and interpretation of the rule make available a number of inferences that reasoners might apply to the visible, but only rarely to the invisible, sides of the cards. Following Johnson-Laird (1995), we assume that reasoners select those cards for which they deduce a constraint for the invisible side. Constraints are deduced by means of one or two relatively spontaneously available inferences. This general approach is consistent with a mental-logic framework (e.g., Braine & O'Brien, 1991), a mental-models framework (e.g., Johnson-Laird & Byrne, 2002), and the suppositional account of the meaning of *if* (Evans & Over, 2004).

Interpretation and available inferences. A number of different interpretational parameters govern which of the conditional infer-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ences, MP, DA, AC, and MT, become available. The parameters are direction; perceived sufficiency versus necessity; conditionality versus biconditionality; and kind of biconditional interpretation, that is, bidirectional or case-distinction interpretation.

The rule is normally understood as one inviting or warranting forward inferences from letters to numbers (MP or DA), but sometimes its direction can be reversed, leading to backward inferences from numbers to letters (AC or MT; Evans, 1993; Oaksford & Chater, 2003b; Oberauer, Hörnig, Weidenfeld, & Wilhelm, 2005). For example, if the rule "if p, then q" is replaced by the equivalent statement "p, only if q," the rule is more frequently seen as one warranting backward inferences. Direction is quantified by the parameter *d* that is the probability with which the rule is seen as one warranting forward inferences. A reversal occurs with probability $1 - d$. It is possible that inferences in both directions become available spontaneously; this case is considered below among the biconditional interpretations of the rule.

Although the antecedent is seen as sufficient in the modal interpretation, it is sometimes instead perceived as necessary; that is, "if p, then q" is misunderstood as "only if p, then q" (e.g., in the interpretation dubbed "enabling" by Johnson-Laird & Byrne, 2002, or "reversed conditional" by Staudenmayer, 1975; see also Ahn & Graham, 1999; Thompson, 1994). If the rule is seen as one warranting forward inferences and the antecedent is seen as sufficient, then the warranted inference is MP; if the antecedent is seen as necessary, it is DA. If the direction of the interpretation is reversed (backward inferences from numbers to letters), the warranted inferences are AC and MT, respectively. Figure 1 shows how direction and perceived sufficiency versus necessity taken together map onto the four conditional inferences. Perceived sufficiency versus necessity is quantified by the parameter *s* that is the probability with which the antecedent condition in the perceived direction is seen as sufficient. With probability $1 - s$, it is seen as necessary. It is possible that the antecedent is seen as both necessary and sufficient; this case is considered below among the biconditional interpretations of the rule.

So far, we assumed that "if p, then q" is understood as a conditional rather than a biconditional rule ("if and only if p, then q"). Many authors (e.g., Gebauer & Laming, 1997; Johnson-Laird & Wason, 1970; Liberman & Klar, 1996; Margolis, 1987; Smalley, 1974) have argued that the rule is frequently understood biconditionally rather than conditionally. The parameter *c* quantifies the probability of a conditional interpretation rather than a biconditional one ($1 - c$).

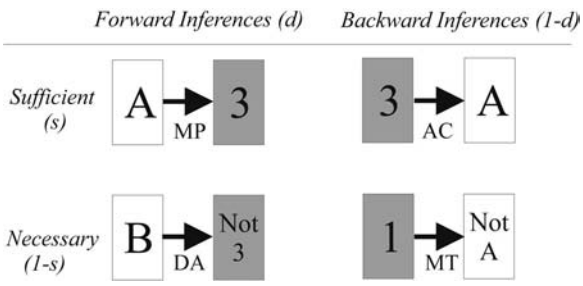

Figure 1. Influence of perceived sufficiency versus necessity and direction of inference on invited inferences for the rule "If there is an *A* on the letter side, then there is a *3* on the number side." For each of the four interpretations, the card left of the arrow predicts the information to the right of the arrow for the other side of this card. Letters below the arrow symbols denote common labels for the four invited inferences: MP = modus ponens; AC = affirmation of the consequent; DA = denial of the antecedent; MT = modus tollens. In the experiments, letter sides showed a capital letter in black on a white card; number sides showed a number in black on a grey card, just as shown here.

Biconditional rules can be understood as conjunctions of conditional premises in different ways. One option is to represent the biconditional "if and only if p, then q" by the bidirectional implication "if p, then q, and if q, then p." A second option is to represent it by the case distinction "if p, then q, and if not p, then not q" (or if direction is reversed, by the case distinction "if q, then p, and if not q, then not p"). We assume that with probability *x*, the interpretation is bidirectional, meaning that the same inferences are invited from letters to numbers as from numbers to letters (that is, in the reversed reading of the rule). For example, if MP is drawn both from the original reading of the rule, "If p, then q," as well as from the converse, "If q, then p," both p card and q card are selected.

Conversely, with probability $1 - x$, the biconditional rule is interpreted as a case distinction. For example, Rumain, Connell, and Braine (1983) have proposed that a conditional such as "If there is an *A* on the letter side, then there is a *3* on the number side" sometimes invites its inverse "If there is not an *A* on the letter side, then there is not a *3* on the number side." In addition, in their theory MP is part of the reasoner's immediately available repertoire of inference rules, suggesting in this example that the rule invites both the MP inference and, via its inverse, the DA inference, leading to the selection of p and $\bar{p}$. If direction is reversed, the AC and MT inferences are analogously invited (see Schroyens, Schaeken, & D'Ydewalle, 2001, for an extended discussion of different biconditional interpretations and inferences thereby invited).

Oberauer (2006) has recently proposed dual-process models for a conclusion-acceptance paradigm, in which reasoners are shown premises and conclusions for each of the four conditional inferences and are asked to evaluate each conclusion as valid or invalid, given the premises. The models incorporate similar interpretative possibilities as the inference model, but they are couched at a more algorithmic level. In particular, Oberauer draws on the distinction between two systems, System 1 and System 2, engaged in conditional reasoning (Stanovich & West, 2000). System 1 is described as heuristic, context dependent, fast, and automatic; System 2 is described as analytic, context independent, slow, and controlled. System 1 takes the liberty to replace and add premises through pragmatic implicature and background knowledge. In Oberauer's (2006) models, System 1 can generate the converse, the inverse, and the contrapositive of the given conditional statement, but it can only draw MP inferences from the generated conditionals. The interpretational possibilities just discussed can readily be mapped on such a System 1 algorithm. For example, if System 1 generates the inverse of the given rule, adds it as a second premise, and draws MP inferences from both premises, the resulting inferences are the same as that obtained under what we called the *case-distinction interpretation.*

Reversible and irreversible reasoning. The just-described parameters characterize the process of interpretation by which a few conditional inferences become available, either a single inference in the case of a conditional interpretation or pairs of inferences in

case of a biconditional interpretation. It is assumed that these are applied to the visible sides of the cards. An idea of old standing (Johnson-Laird & Wason, 1970; Wason & Johnson-Laird, 1972, chap. 15) is that people often do not reason from the invisible sides of the cards and in this sense, reasoners treat the cards as irreversible. Thus, with probability *i*, the available inferences are applied only to the visible sides of the cards. Alternatively, with probability $1 - i$, reasoners also consider the invisible sides and apply the available inferences to the invisible sides. For example, if MP is available, considering the possibilities for the invisible sides eventually entails applying MP to a supposed (instance of) p on the invisible side of the $\bar{q}$– card (e.g., to a supposed *A* on the invisible side of the card showing the number 4). This leads to a contradiction with the visible side and, hence, to the constraint that there must not be p on the invisible side and thereby to the selection of this card.[2]

According to Oberauer (2006), reasoning tactics of this kind are ascribed to System 2, so that the distinction between irreversible and reversible reasoning can be seen as that between System 1 and System 2. In this regard, the present model is again structurally similar to Oberauer's (2006) dual-process models for a conclusion-acceptance paradigm, in particular to his suppositional model (exclusive variant) that also incorporates a parameter for this distinction, although the different paradigms imply many differences between the inference model and Oberauer's models. According to Oberauer, the suppositional model is a formalization of Evans and Over's (2004) suppositional theory of conditional reasoning, and like Oberauer's suppositional model, the present inference model is consistent with the recent suppositional theory by Evans and Over (2004): People are assumed to make MP inferences from the interpretation of the rule that they endorse. MT inferences are made rarely and are made through a suppositional strategy that is captured by parameter *i*. On the other hand, the recent version of the mental-models theory (Johnson-Laird & Byrne, 2002), when augmented by a directional component to capture the distinction made by parameter *d* for direction, can also emulate the interpretational distinctions and reasoning tactics built into our inference model. For this reason, it seems likely that a quantitative specification of mental-models theory could be derived that results in a model with the same or a very similar outcome space as the inference model. In this sense, the inference model is also consistent with the mental-models theory.

The inference model per se is silent about the important question in what way pragmatic influences and/or background knowledge elicit the different interpretations and spontaneous inferences. In this regard we refer to the arguments and findings from the cited conditional-reasoning literature.

Summary of the inference model. Figure 2 shows a processing-tree representation of the inference model. The first branching determines whether a conditional or biconditional interpretation is followed. In the case of a conditional interpretation, the rule can be seen as suggesting forward inferences (i.e., from letters to numbers) or backward inferences (i.e., from numbers to letters). Independently, the rule can be perceived as describing a sufficient relationship in the perceived direction or a necessary one, mapping onto the different inferences as shown in Figure 1. The inference is then applied either only to the visible sides or to the visible and invisible sides, leading to the card selections shown as terminal nodes of the different processing paths in Figure 2.

For example, following the upper branch in each case corresponds to an interpretation of the rule in which p is seen as sufficient for q. This invites the MP inference from p to q (see Figure 1). According to the upper branch in the final branching (irreversible versus reversible reasoning), MP is applied only to the visible sides, which is possible for the p card and leads to the selection of that card on the grounds that the rule implies a constraint for the invisible side of that card via MP, namely that there must be q on the invisible side. The probability of a path from the root of the tree to a terminal selection pattern is the product of the parameters on the branches of the path; the probability of a specific selection pattern is the sum of the probabilities of all paths ending in that selection.

The inference model uses five parameters, four of them characterizing the interpretation of the rule (conditionality vs. biconditionality *c*, bidirectionality vs. case distinction *x*, direction *d*, and perceived sufficiency vs. necessity *s*), one (irreversibility *i*) referring to the reasoning process itself. The model uses one more parameter than the independence model.

The sequence with which the different parameters occur in Figure 2 is not important, although we believe that it represents the most plausible processing sequence. The same model could be depicted by a processing tree in which the different distinctions are reordered, and thus, the outcome of the model analyses below do not permit inferences as to the processing sequence.

Unlike the independence model, the inference model has little chance of fitting real data, because some of the 16 possible patterns such as the partial insight pattern (1, 0, 1, 1) should never arise according to the model. Yet, all possible patterns have sometimes been reported, if with widely differing frequencies; but even rare occurrences of "forbidden" patterns constitute very strong evidence against the model.

Combined Models

This is not true of models that combine the inference model with the independence model. As already outlined above, we consider three combined models, the heuristic-analytic model, the inference-guessing model, and the relevance-inference-guessing model.

The heuristic-analytic model can be seen as one possible specification of Evans' (1984, 1989) heuristic-analytic theory of reasoning. According to that theory—applied to the WST—preattentional heuristics determine, via perceived relevance, those

[2] Note that irreversibility, that is, reasoning only from the visible sides, has basically the same effect as a certain misunderstanding of the rule postulated by Gebauer and Laming (1997; see also Osman & Laming, 2001; Wason & Johnson-Laird, 1972, p. 176) in which the rule "If there is an *A* on one side of the cards, then there is a *3* on the other side" is misunderstood as "If there is an *A* on top, then there is a *3* underneath." In formulating the rules for the present experiments, we tried to avoid a misunderstanding in terms of top/underneath by referring to a letter side and a number side of the cards instead of to just "one side" versus the "other side." As pointed out by Oaksford and Chater (2003a), it is also difficult to see how the misinterpretation idea can account for the fact that there is relatively little difference between the standard version of the WST and versions in which all information is on one side of the card (e.g., Goodwin & Wason, 1972).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

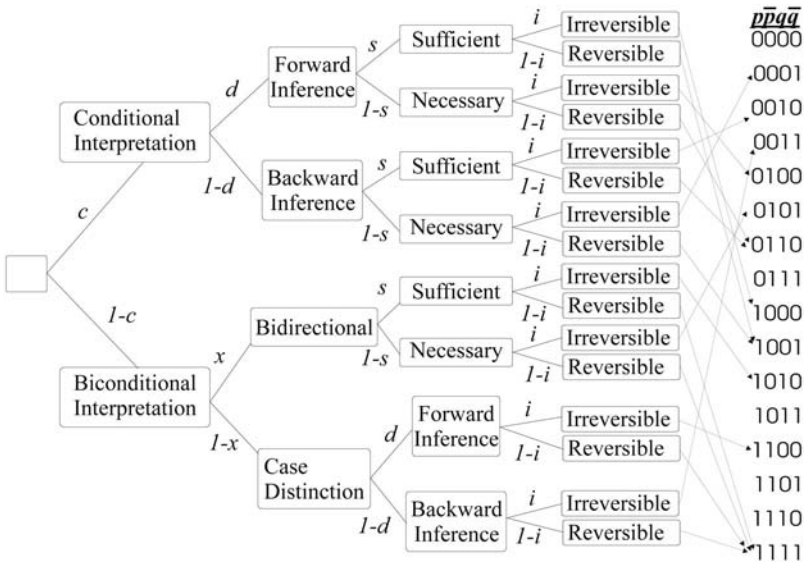

Figure 2. Processing-tree representation of the inference model. The model uses five parameters, defined as follows: c = conditionality vs. biconditionality; x = bidirectionality vs. case distinction; d = direction; s = perceived sufficiency vs. necessity; i = irreversibility. Four cards are presented for selection and are referred to as, respectively, the p card, the $\bar{p}$ card, the q card, and the $\bar{q}$ card according to their logical relationship to the propositions p and q in the rule "If p, then q."

contents upon which inferential processes operate. For example, a person with a biconditional interpretation and reversible reasoning would normally select all four cards but may not perceive the $\bar{p}$ and $\bar{q}$ cards as relevant in the first place. This means that these cards are not considered further and that they are not selected, leading to the selection (p, q). In our heuristic-analytic model, the independence model acts as a filter determining what is passed on to the inference model. In this use, the card parameters of the independence model quantify the probabilities with which the different cards become subjectively relevant. A card is selected if it is subjectively relevant and if the inference part of the model leads to the selection of that card.[3] The heuristic-analytic model uses nine parameters, four of them pertaining to the independence model and five to the inference model.

The inference-guessing model combines heuristic and analytic processes in the manner proposed by Evans' (1977) statistical theory of reasoning, and it can be seen as one possible specification of Evans' (2006) revised heuristic-analytic theory and related dual-process theories that remove the strictly sequential nature of the interaction of heuristic and analytic processes or systems (Evans, in press). It combines the same two submodels as the heuristic-analytic model but combines them as alternative processing paths rather than a single processing sequence. In the inference-guessing model, card selections are governed either by the inference submodel or by the independence model. Which submodel takes precedence is determined by a new parameter a. With probability a, card selections follow the inference submodel; with probability $1 - a$, they follow the independence model, in each case for all four cards. In this model, it is assumed that some people simply guess which cards to select, perhaps due to a lack of motivation or after failed efforts to understand what is required of them, or that they let themselves be governed only by subjective relevance (Evans, 1995) or expected information gain (e.g., Oaksford & Chater, 2003a)—possibilities that are captured by the independence model. However, a certain proportion of response patterns is based on attempts to evaluate the rule on the basis of its testable consequences and thus on inferential processes as described by the inference model. Note, however, that these inferential processes again comprise a mixture of more shallow reasoning processes (irreversible reasoning) that can be ascribed to System 1 and deeper reasoning processes (reversible reasoning) ascribed to System 2. The model uses 10 parameters, 4 for the independence model, 5 for the inference submodel, and a new parameter a for the proportion of responses that are governed by the inference submodel rather than the independence model.

Finally, there is the possibility that (a) preattentional heuristic processes guide attention toward certain cards and away from certain others, followed (b) by reasoning from the cards seen as relevant, accompanied by (c) an alternative branch of selections described by the independence model. That is, the inference submodel (b) is combined with the independence model acting both as a filter (a) and, with possibly different card probabilities, as an alternative processing tree (c). This model thereby allows for relevance judgments to operate early on in determining perceived relevance of the individual cards and for response bias at a late stage of output processes in the form of guessing a selection. In the following, we refer to this model as the *relevance-inference-*

[3] Note that Evans (e.g., Evans, 1995) has applied his heuristic-analytic model differently in that he argued that only the heuristic part is involved in causing selection-task data. To distinguish this point of view from the one built into the heuristic-analytic model, we refer to Evans' (1995) position as his *account by relevance.*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

guessing model. The model uses 14 parameters, and given that there are 15 independent selection frequencies to be described, it is an almost saturated model, leaving one degree of freedom for the test of model fit.

Summary: The Set of Models

Figure 3 shows the models discussed so far. The links between two models in the figure signal that the lower model is a submodel of the upper model, that is, the lower model derives from the upper model through constraints on some of its parameters.

The inference submodel incorporates the assumption that direction, sufficiency versus necessity, and conditionality versus biconditionality combine independently to determine the available inferences. In particular, it is assumed that the same parameter s for perceived sufficiency versus necessity applies independently of perceived direction and conditionality. This assumption keeps the model simple, which is desirable for theoretical and statistical reasons.

Nevertheless, it is an arbitrary assumption for which there is little justification a priori. If the simple models do not fit, a natural question to ask is therefore whether a model that relaxes this assumption fits, that is, a model with different parameters s_f, s_b, and s_{fb} for sufficiency versus necessity depending upon, respectively, whether the rule is seen as one warranting forward inferences (i.e., from letters to numbers), backward inferences (i.e., from numbers to letters), or inferences in both directions. That is, s_f is the probability with which p is seen as sufficient for q in forward inferences; s_b the probability with which q is seen as sufficient for p in backward inferences; and s_{fb} the probability with which p is seen as sufficient for q and simultaneously q for p in a bidirectional, biconditional reading of the rule.

The relaxed models are also shown in Figure 3 for the heuristic-analytic model and the inference-guessing model; relaxing the assumption for the relevance-inference-guessing model leads to a supersaturated model with more parameters than there are independent pattern frequencies, and we therefore omit this model from further study.

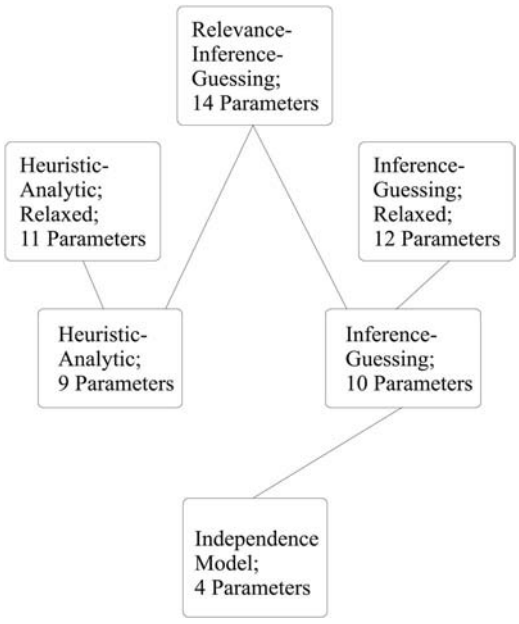

Figure 3. The family of models considered for model selection.

Statistical Evaluation of the Models

The models are members of the large family of multinomial processing tree models. As explained by Batchelder and Riefer (1999), multinomial processing tree models are designed for specific experimental paradigms such as the WST that yield only a limited set of response categories. Moreover, parameters for a model need to capture the main cognitive factors involved in the paradigm to have a chance to fit observed data. As a consequence, there is usually a large number of parameters to account for a small number of categories, leaving few degrees of freedom for testing the model's fit. For example, as pointed out by Evans (in press), a drawback of dual-process models such as the inference-guessing model or the heuristic-analytic model is

> the number of parameters that are required to model any given data set. Put conceptually, you need to have (a) a theory of how heuristic processes affect responding, (b) a theory of analytic reasoning on the given task, which as we have seen cannot simply be equated with logic, and (c) a theory of how and to what extent the two processes compete for control of the response. (Evans, in press, p. 6)

For models with many parameters, many researchers feel that achieving a good fit of the model to the data is relatively trivial. We addressed this concern in several ways. As shown in the following, the results of (a) model selection criteria, (b) goodness-of-fit tests, (c) simulation studies, (d) assessments of possible effects of interindividual variability, and (e) parameter validation experiments converge in showing that one of our models is clearly superior to its competitors and fits the available data in a nontrivial way.

We began the evaluation and selection exercise by using model selection criteria that penalize nonparsimonious models (Myung, 2000). In particular, we computed Akaike's information criterion (AIC) and the Bayes information criterion (BIC) for each model in Figure 3 and each independent data set that we collected. The two criteria are defined as $\mathrm{AIC} = -2\log L + 2S$ and $\mathrm{BIC} = -2\log L + S \log n$, where L is the likelihood of the data (computed at the maximum likelihood estimates of the parameters), S is the number of model parameters, and n is the number of data points in the data set. Note that AIC and BIC increase as the number of parameters is increased, implying a penalty for lack of simplicity of the model. The model with the smallest criterion values provides the best compromise between data fit and simplicity.

AIC and BIC are based on different normative approaches to correcting for model complexity. AIC follows a generalization-based approach, meaning "that a model is evaluated in terms of its ability to fit not only observed data, but also unseen (e.g., future) data from the same process" (Myung, 2000, p. 199). Complex models will tend to fit given data better than simple models, but the greater flexibility of complex models turns into a disadvantage in predicting future data because of the greater danger of capitalizing on measurement error in fitting observed data. That is, AIC assesses the goodness of the model for predicting future data from the model as fitted to the observed data. BIC, on the other hand,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

follows an explanation-based approach evaluating each model in terms of the expected likelihood of only the present data, averaged over all possible parameter values. A penalty for complex models is again implied because the larger outcome space of complex models means that the likelihood of the given data will deviate more widely from its maximum for many parameter values. BIC can be grounded in Bayesian statistics so that the difference in BIC values of two models is an approximation of the (logarithm of the) so-called Bayes factor of the two models, that is, the ratio of the posterior probabilities of the models, given the observed data. Theoretical results obtained in the context of regression models and covariance structure modeling show that BIC is consistent in the sense that it will correctly identify the true model among a set of competing models as sample size goes to infinity, whereas AIC performs perfectly in selecting the model that is the closest approximation to the true model if the true model is not among the ones considered (Myung, 2000).

Table 1 shows the selection frequencies for the data sets that we collected. Table 2 shows the AIC and BIC values for the six models of Figure 3 for each WST data set that we collected.[4] It was necessary to collect new data, because published WST data sets are usually too small to provide usable estimates of the probabilities of the individual response patterns (with one exception shown in the row labeled *Old* in Tables 1 and 2). The data sets themselves and the procedure by which they were obtained are described later. In all, there were 18 WST data sets, almost all of them with at least 300 participants, each participant contributing one response pattern in the WST. The columns labeled ODS and ODS_e in Table 2 refer to ODS models that are discussed later.

The six models were associated with significantly different AIC and BIC values, as shown by a Friedman nonparametric test, $\chi^2(5) = 48.22$, $p < .01$, and $\chi^2(5) = 66.60$, $p < .01$, respectively. As can be seen in Table 2, the independence model was consistently associated with the largest AIC and BIC values with a mean rank of 6.0, whereas the inference-guessing model received the lowest mean ranks. Separate pairwise Wilcoxon tests, with the 18 data sets defining the replications factor, revealed that the inference-guessing model was associated with significantly smaller AIC values than the heuristic-analytic model, the relevance-inference-guessing model, and the independence model (largest $Z = -2.07$, largest $p = .04$) and with significantly smaller BIC values than all models other than the heuristic-analytic model (largest $Z = -3.22$, largest $p < .01$). Overall, the inference-guessing model emerged as the winning model.

Model selection via AIC and BIC prevents the choice of an unnecessarily complex model that overfits the data and generalizes poorly (Myung, 2000). Model selection does not, however, imply an acceptable goodness of fit of the selected model because simplicity is traded against fit; for example, none of the models in the family might provide an acceptable description of the data. As can be seen in the last row of Table 2, the inference-guessing model and its cousin, the inference-guessing model with relaxed assumptions, could be maintained in log-likelihood ratio tests of goodness of fit (Batchelder & Riefer, 1999) for 16 of the 18 data sets at the 5% level of significance. Only the almost saturated relevance-inference-guessing model performed slightly better, whereas the heuristic-analytic model with and without relaxed assumptions had to be rejected substantially more frequently. Thus, the inference-guessing models provide a statistically satisfactory first approximation of the data.

Returning once more to the concern that achieving a good model fit is trivial: If a model is capable of fitting any data set that might plausibly arise from the WST task, on the basis of what is known about such data, goodness-of-fit of the model per se is not surprising or informative (Roberts & Pashler, 2000). We generated 1,000 random data sets with 300 data points each conforming to the typical trends found in WST data: (a) p alone and the (p, q) pattern each receive between 25% and 50% of all selections, (b) no other pattern is selected as frequently as either of these two patterns, (c) there are at most 10% (p, $\bar{q}$) selections; 4) the marginal card frequencies follow the order $p > q > \bar{q} > \bar{p}$. Although these data sets thereby conform to what is uncontroversially known about WST data, the inference-guessing model had to be rejected at the 5%-level for 95% of them; the inference-guessing model with relaxed assumptions had to be rejected for 85%. Thus, these models are able to fit only a subset of plausible data, implying that a good fit to real data is not trivial.

Taken together, there is little support for the independence model despite its simplicity. This conceptually replicates analogous findings by Pollard (1985) and Oaksford and Chater (1994) on a much larger data base. In contrast, a model of medium complexity, the inference-guessing model, received the highest support.

One reason for a simple model such as the independence model not to fit the data is that there may be pronounced interindividual differences in the parameters of the model. Although many statistical problems associated with individual differences are minimized when there is only one data point from each person (Klauer, 2006; Riefer & Batchelder, 1991) as in the present experiments, we investigated whether the poorly performing but parsimonious independence model described the data better when possible parameter heterogeneity was explicitly modeled. These analyses are presented in Appendix A. Taken together, they clearly show that the parsimonious independence model fails to fit the data even when possible parameter variability between individuals is taken into account.

Because the ODS model variant suggested by Oaksford and Chater (2003a) is a submodel of the independence model, the poor fit of the latter also applies to the former. Thus, not surprisingly, the independence ODS model by Oaksford and Chater (2003a) is associated with larger AIC and BIC values than the other models (see Table 2). Wilcoxon tests showed that it performed worse than the inference-guessing model in terms of both AIC and BIC (smallest $|Z| = 3.72$, largest $p < .01$). Considering goodness of fit, the model had to be rejected at the 1% level of significance for all 18 data sets.

It is important to acknowledge, however, that the original exposition of the account by ODS (Oaksford & Chater, 1994) incorporated a device for explaining nonindependence of card choices,

[4] Following common practice (Hu, 1991; Rothkegel, 1999), a constant of one was added to all cell frequencies in the event that there was one or more cells with observed zero frequency in a data set. The data in the row labeled *Inf.* stem from an inference task rather than from the WST (see General Discussion). They are therefore not included in the analyses presented in this section.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Frequencies of Different Selection Patterns

Group	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
Experiment 1																
CG																
Frequency	10	6	20	3	7	11	1	1	96	11	106	4	5	1	1	17
%	3	2	7	1	2	4	0	0	32	4	35	1	2	0	0	6
EG																
Frequency	5	4	29	4	8	18	6	0	87	22	91	.11	6	1	1	28
%	2	1	9	1	2	6	2	0	27	7	28	3	2	0	0	9
Experiment 2																
CG																
Frequency	13	7	20	4	11	13	4	1	104	12	113	1	6	2	0	11
%	4	2	6	1	3	4	1	0	32	4	35	0	2	1	0	3
EG																
Frequency	5	9	11	9	13	11	7	8	112	36	49	6	11	1	1	11
%	2	3	4	3	4	4	2	3	37	12	16	2	4	0	0	4
Experiment 3																
CG																
Frequency	9	11	28	4	10	20	4	2	95	16	109	5	4	2	1	15
%	3	3	8	1	3	6	1	1	28	5	33	1	1	1	0	4
EG																
Frequency	16	21	33	4	5	2	2	0	134	12	49	7	7	1	3	4
%	5	7	11	1	2	1	1	0	45	4	16	2	2	0	1	1
Experiment 4																
CG																
Frequency	11	8	17	4	4	17	1	2	92	10	111	4	5	0	2	12
%	4	3	6	1	1	6	0	1	31	3	37	1	2	0	1	4
EG																
Frequency	12	9	42	5	14	4	0	0	126	7	59	16	1	0	1	5
%	4	3	14	2	5	1	0	0	42	2	20	5	0	0	0	2
Experiment 5																
IT A,3																
Frequency	9	7	26	2	5	16	3	2	112	11	131	1	5	1	2	12
%	3	2	8	1	1	5	1	1	32	3	38	0	1	0	1	3
IT 3,A																
Frequency	9	9	24	3	5	19	0	3	132	11	101	2	3	1	5	12
%	3	3	7	1	1	6	0	1	39	3	30	1	1	0	1	4
OI A,3																
Frequency	8	7	67	11	5	10	2	1	54	23	99	13	5	2	1	16
%	2	2	21	3	2	3	1	0	17	7	31	4	2	1	0	5
OI 3,A																
Frequency	14	8	64	13	8	29	5	2	46	16	59	10	5	3	2	16
%	5	3	21	4	3	10	2	1	15	5	20	3	2	1	1	5
Experiment 6																
p																
Frequency	19	12	31	9	16	15	9	5	61	19	81	7	6	0	0	19
%	6	4	10	3	5	5	3	2	20	6	26	2	2	0	0	6
$\bar{p}$																
Frequency	15	13	24	7	9	5	5	1	63	30	93	15	2	0	1	17
%	5	4	8	2	3	2	2	0	21	10	31	5	1	0	0	6
q																
Frequency	13	17	16	7	11	17	5	1	85	18	70	6	19	8	4	21
%	4	5	5	2	3	5	2	0	27	6	22	2	6	3	1	7
$\bar{q}$																
Frequency	12	7	25	5	10	10	5	0	67	15	95	10	11	2	9	27
%	4	2	8	2	3	3	2	0	22	5	31	3	4	1	3	9
Old[a]																
Frequency	1	0	6	0	0	5	0	1	57	19	144	9	3	0	3	9
%	0	0	2	0	0	2	0	0	22	7	56	4	1	0	1	4
New[a]																
Frequency	1	2	11	2	2	9	3	2	74	20	89	2	5	1	1	9
%	0	1	5	1	1	4	1	1	32	9	38	1	2	0	0	4
Inf.[a]																
Frequency	8	3	2	1	8	3	3	1	115	29	102	4	3	0	4	14
%	3	1	1	0	3	1	1	0	38	10	34	1	1	0	1	5

Note. CG = control group; EG = experimental group; IT = if then; OI = only if; Inf. = inference task.
[a] Laboratory data compiled from the literature (Old); newly collected (New); collected from an inference task format (Inf.).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Model Selection: Akaike's Information Criterion and Bayes Information Criterion Values

Experiment	Group	RIG	IG	IG rel.	HA	HA rel.	Ind.	ODS	ODS_e
				AIC					
1	CG	29.65	25.28	28.36	21.88	25.09	147.15	171.40	106.98
1	EG	29.12	25.89	25.72	25.75	24.35	176.28	178.74	115.88
2	CG	28.32	21.29	24.59	24.92	27.89	152.02	199.11	95.31
2	EG	29.81	25.66	29.51	59.03	56.47	131.13	166.31	51.95
3	CG	28.59	22.85	24.90	25.39	25.27	185.41	212.64	102.81
3	EG	30.43	29.37	25.48	38.35	29.50	71.40	189.28	107.80
4	CG	29.56	29.74	26.66	30.73	28.16	160.49	183.93	101.03
4	EG	28.47	28.73	32.64	48.67	49.09	109.88	207.92	126.97
5	IT A, 3	31.77	27.97	25.24	26.84	24.01	183.92	225.71	119.89
5	IT 3, A	31.74	41.68	32.96	50.00	35.45	190.97	240.08	108.66
5	OI A, 3	31.31	29.46	30.50	36.24	37.19	118.27	149.40	113.89
5	OI 3, A	31.45	44.28	28.43	46.86	33.03	150.50	174.58	142.95
6	p	28.84	25.26	28.79	27.38	27.75	121.44	137.90	94.32
6	$\bar{p}$	30.52	25.19	25.94	28.12	28.19	85.70	98.75	78.69
6	q	28.13	26.92	27.48	20.96	22.25	109.67	134.67	47.62
6	$\bar{q}$	28.22	23.15	26.51	23.50	22.92	97.42	100.75	23.49
Old[a]		31.51	28.49	31.62	36.80	38.33	94.00	91.78	71.35
New[a]		32.42	30.11	31.02	36.71	35.21	122.12	125.58	83.19
Mean rank[b]		3.72	2.22	2.39	3.61	3.06	6.00		
				BIC					
1	CG	81.50	62.31	72.81	55.22	65.84	161.97	178.81	121.79
1	EG	82.60	64.09	71.56	60.13	66.37	191.56	186.38	131.16
2	CG	81.85	59.52	70.46	59.33	69.95	167.32	206.76	110.61
2	EG	81.67	62.69	73.96	92.37	97.21	145.94	173.72	66.77
3	CG	81.98	60.99	70.67	59.72	67.22	200.67	220.27	118.06
3	EG	83.01	66.93	70.55	72.16	70.81	86.42	196.79	122.82
4	CG	82.14	67.30	71.73	64.53	69.47	175.51	191.44	116.05
4	EG	81.09	66.32	77.74	82.50	90.44	124.91	215.44	142.00
5	IT A, 3	85.58	66.41	71.36	61.43	66.29	199.29	233.40	135.27
5	IT 3, A	85.95	80.40	79.43	84.85	78.05	206.46	247.82	124.15
5	OI A, 3	84.24	67.27	75.86	70.27	78.78	133.39	156.96	129.01
5	OI 3, A	83.30	81.32	72.88	80.19	73.77	165.32	181.99	157.77
6	p	81.81	63.10	74.20	61.43	69.37	136.58	145.47	109.45
6	$\bar{p}$	83.10	62.75	71.01	61.92	69.50	100.72	106.26	93.72
6	q	80.80	64.54	72.62	54.82	63.63	124.72	142.19	62.67
6	$\bar{q}$	81.23	61.01	71.95	57.58	64.58	112.56	108.33	38.64
Old[a]		82.04	64.58	74.93	69.29	78.04	108.43	99.00	85.78
New[a]		80.74	64.62	72.43	67.77	73.17	135.92	132.49	96.99
Mean rank[b]		4.78	1.94	3.22	1.94	3.11	6.00		
No. of data sets fitting at the 5% level		17	16	16	9	11	0	0	1

Note. RIG = relevance-inference-guessing model; IG = inference-guessing model; rel. = relaxed assumptions; HA = heuristic-analytic model; Ind. = independence model; ODS = optimal data selection model; ODS_e = extended ODS model; CG = control group; EG = experimental group; IT = if then; OI = only if.
[a] Laboratory data compiled from the literature (Old) and newly collected (New). [b] Excluding the ODS models.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

and it can be argued that the independence assumption is only an auxiliary assumption that is not central to the account by ODS. For this reason, we also fitted Oaksford and Chater's (2003a) ODS model in a way that releases it from the independence assumption. The extended ODS model also performed significantly worse than the inference-guessing model in terms of AIC and BIC (see Table 2, column ODS_e). Considering goodness-of-fit, it had to be rejected at the 1% level of significance for 17 of the 18 data sets. Details are described in Appendix A. It is possible, however, that another quantitative specification of the approach by ODS exists that provides a better account of the data.

In the experiments that follow, we move beyond summary statistical evaluations and examine the psychological viability of the inference-guessing model. To evaluate the proposed psychological meaning of the different parameters (such as direction, conditionality versus biconditionality, irreversibility, and so forth), we realized experimental comparisons that target the interpretive, inferential, and heuristic parameters built into the model one by one. For example, suppose we know from prior work or on theoretical grounds that a certain manipulation affects the likelihood with which the rule is seen as biconditional rather than conditional. If this manipulation is implemented in an experiment, an effect should be seen on the parameter *c* for conditionality versus biconditionality. The purpose of the experiments reported below is to validate the different model parameters in this manner. Table 3 provides a

Table 3
Parameters of the Inference-Guessing Model

Parameter	Meaning
a	Probability of response being based on the inference submodel
p	Probability of selecting card p under the independence submodel
$\bar{p}$	Probability of selecting card $\bar{p}$ under the independence submodel
q	Probability of selecting card q under the independence submodel
$\bar{q}$	Probability of selecting card $\bar{q}$ under the independence submodel
c	Probability of conditional rather than biconditional interpretation
x	Probability of bidirectional interpretation (e.g., the rule is interpreted as "if p, then q and if q, then p") rather than case-distinction interpretation, given biconditional interpretations (e.g., the rule is interpreted as "if p, then q and if not p, then not q")
d	Probability of inferences in the forward direction (from letters to numbers) rather than backward direction (from numbers to letters)
s	Probability of perceived sufficiency rather than necessity
i	Probability of irreversible reasoning (inferences only from the visible sides of the cards) rather than reversible reasoning (inferences from visible and invisible sides)

summary of the parameters and their intended psychological interpretation for later reference.

In the experiments, each participant was tested on only one Wason selection problem in between-participants designs, as was typical of early work using the WST. In this way, we can be sure that learning and transfer, and individual differences therein, play little role in the making of the data. The next section describes those parts of the methods and procedures that were common to all experiments.

General Method

The studies reported in this article were implemented as Internet-based experiments. Each participant performed only one WST, and experimental manipulations were implemented between participants. Participants were randomly assigned to the different experimental groups.

The experiments were advertised in several newsgroups, submitted to various search machines, and publicized in several Internet documents that collect links to online studies and experiments such as the Web Laboratory for Experimental Psychology (Reips, n.d.). The experiment was described as a short logic test with individualized feedback, conducted for scientific purposes.

The experiment consisted of a start page, an experimental page, and a feedback page. The experiment was offered in a German and an English version that were reached by different links. From the start page of each version, it was possible to reach the start page of the other version directly.

The start page asked participants whether they would like to participate in a short scientific study about reasoning of a duration of about 5 min. They were also asked to read the instructions carefully, if they were to participate.

Persons wishing to proceed to the problem indicated their intent by clicking on a link labeled *yes* on the start page. For each such participant, an experimental page was generated online. The experimental page comprised the instructions, the Wason selection problem, and a biographical questionnaire. In the questionnaire, participants were asked for demographic bits of information about themselves. Additional questions addressed the participant's language proficiency, prior experience with the just-completed task or similar card selection tasks, and whether the participant "had answered all questions carefully and participated for the first time" or whether he or she "just want[s] to see the results by way of trial without seriously participating in the study." These questions were used to screen out potentially suspicious data sets as explained next. The feedback page provided feedback about the participant's selection and the normatively correct selection. The rationale for the normative selection (p, $\bar{q}$) was explained.

A number of studies have discussed potential problems and advantages of Internet research (e.g., Kraut et al., 2004; Reips, 2002). In the present context, important problems are the low experimental control over the participants' situational circumstances and behavior, the problem of possible multiple participation, and the potential problem of selective dropout. Selective dropout is a problem if dropout affects some of the experimental groups more strongly than others, thereby compromising the comparability of the experimental groups.

Several techniques have been proposed to minimize such problems (Reips, 2002). Following these recommendations, submissions were accepted for data analysis in the present studies only if no submission had been previously received from the same Internet protocol (IP) address. For this purpose, a cumulative record was kept of the IP addresses of submissions throughout the present series of experiments to screen out any participant who might already have participated in the current or a previous experiment in the series. Furthermore, data were accepted only of those participants who stated that they had responded to all questions carefully and submitted data for the first time and that they were not familiar with the problem or similar card selection problems. Finally, participants were excluded who stated that their English (or German in the German version) was poor. These measures aimed at minimizing the potential problems of multiple participation, lack of seriousness, motivation, and comprehension. In addition, a few participants were excluded because they claimed to be older than 90 years.

Demographic information about the samples of participants is given in Appendix B for each experiment. Participants were mostly in their late teens to late 30s with a mean age of 27 years. Women were slightly in the majority. Participants' educational and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

occupational status was above that of the general population; for example, participants reported having spent an average of 13 years at school (including college/university).

A given experiment remained online until there were at least 300 participants who fulfilled the above criteria in each experimental group. To assess the possibility of selective dropout, we tested as a first step in the data analysis of each study whether the numbers of accepted submissions were significantly different between the experimental groups.[5] Chi-squared tests revealed that there was no significant difference between the number of participants in each experimental group for any of the experiments (smallest $p = .16$).

Experimental Validation of the Inference-Guessing Model

Six experiments were conducted to validate the intended interpretations of the parameters of the inference-guessing model (see Table 3). Experiments 1 and 2 used helpful hints. The hint used in Experiment 1 was designed to enhance reversibility of reasoning and thereby targeted parameter *i* for irreversibility; an additional hint used in Experiment 2 discouraged a bidirectional, biconditional interpretation, leading to the expectation that parameters *c* (conditionality versus biconditionality) and *x* (bidirectionality versus case distinctions) should be affected. Experiments 3 and 4 introduced a second rule presenting either an alternative antecedent or an alternative consequent. It was expected that parameter *c* would again be affected but also parameter *s* for perceived sufficiency versus necessity. Finally, Experiment 5 contrasted rules phrased with *if* and rules with *only if* to manipulate parameter *d* for direction of warranted inferences. In Experiment 6, we used an extended array of cards in an attempt to manipulate each of the parameters p, $\bar{p}$, q, and $\bar{q}$, of the independence submodel selectively.

Experiments 1 and 2: Irreversibility and Conditionality

The first two experiments aimed at improving reasoning performance in the WST through instructions. In both experiments, standard instructions were used for a control group, whereas members of an experimental group received potentially helpful hints.

In the first experiment, only one hint was given that specifically targeted the irreversibility parameter *i* as follows: "It is necessary to consider for EACH card what the possibilities are for the invisible side, and to consider for EACH such possibility whether it is consistent with the rule or refutes it."

This hint was also used in Experiment 2 along with two additional hints, one of which directly questioned a bidirectional, biconditional interpretation of the rule as follows (assuming the rule is "If there is an *A* on the letter side, then there is a *3* on the number side"): "Please note that the rule to be tested is not equivalent to the following reverse rule: 'If there is a *3* on the number side, then there is an *A* on the letter side.' In other words: A card with a *3* on its number side may have any letter on its letter side." Note that the last sentence is also part of Platt and Grigg's (1993) famous explicated instruction. A third hint, "The visible sides of the cards displayed are all different. This need not be true of the backsides," was added in Experiment 2 because one of us believed that it prevented a frequent misunderstanding of the rule.

We hypothesized that the irreversibility parameter should be decreased in the experimental group in both experiments because of the hint targeting irreversibility and that the hint in Experiment 2 that questioned a bidirectional, biconditional interpretation would lead to an increase in the conditionality versus biconditionality parameter *c* and that it would shift remaining biconditional interpretations away from bidirectionality towards case distinction interpretations, leading to a decrease in parameter *x*. The third hint in Experiment 2 (the backsides need not all be different) was expected to direct more attention to the invisible sides of the cards. Thus, like the first hint, it should help to decrease the irreversibility parameter *i*.

Method

Participants. Participants were sampled via the Internet as already described. In Experiment 1, there were 300 and 321 participants in the control group and the experimental group, respectively; in Experiment 2, these numbers were 322 and 300. Demographic information about the samples of participants is given in Appendix B.

Procedure. The experimental page began with the following standard instruction:

> Below you see a number of cards from a set of cards. Each card in the set has a capital letter on one side and a digit on the other. Naturally, only one side is visible in each case. For the set of cards, a rule has been stated. It is: . . .

This was followed by a rule with a randomly sampled capital letter in the antecedent (excluding the letters I, O, and V because of their similarity to numerals) and a randomly sampled number between 1 and 9 in the consequent, for example, "If there is an *A* on the letter side of the card, then there is a *3* on the number side."

In the next paragraph, participants were informed that "you must decide which card(s) displayed would have to be turned over in order to test the truth or falsity of the rule. Please use the mouse to check the card(s) that would have to be turned over. Do not check cards that would not have to be turned. You may take as long as you like."

Members of the experimental groups were told that many studies have shown that the task is solved correctly only by a minority of people, and that they would therefore receive a hint (Experiment 1) or three hints (Experiment 2) that may be helpful to avoid typical errors. In Experiment 1, this was followed by the hint already described. In Experiment 2, participants received the hint from Experiment 1, preceded by the two additional hints already quoted above. Participants in the experimental groups of both experiments were then reminded that they were to check those cards that must be turned over in order to test the truth or falsity of the rule.

Below this, four cards were displayed in a row. Letter sides showed a capital letter in black on a white card; number sides

[5] Like before (see footnote 4), a constant of one was added to all cell frequencies in the event that there was one or more cells with observed zero frequency in a data set for the multinomial-model analyses. In order to avoid differential treatment of the groups, we then added the constant to the data from all groups of a given experiment.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

showed a number in black on a grey card (as in Figure 1). The four cards displayed the letter mentioned in the rule, another randomly sampled letter (excluding the letters I, O, and V), the number mentioned in the rule, and another randomly sampled number in random order. Below each card, a box could be checked to signal selection of the card. No action was required if a card was not to be selected. All randomizations were carried out for each participant anew.

Results and Discussion

The pattern frequencies are shown in Table 1. Descriptively, the hints led to a modest increase in (p, $\bar{q}$) selections in both experiments. In Experiment 1, there was furthermore a modest increase in the selection of all four cards, and in Experiment 2, there was also a strong decrease in selections of (p, q).

The inference-guessing model was fitted to the data of each experiment with separate parameters for each group. The chi-squared log-likelihood ratio test of model fit, G^2, indicated that the model described the data well: In Experiment 1, $G^2(10) = 10.85$, $p = .37$; in Experiment 2, $G^2(10) = 5.87$, $p = .83$. Table 4 shows the values of the different model parameters in Experiments 1 and 2 along with confidence intervals and chi-squared tests for differences between control group and experimental group for each parameter.

As can be seen, the proportion of responses under the inference submodel, *a*, was approximately 75% in all conditions; the remaining 25% followed the independence model. Responses governed by the inference model are approximately equally often guided by a conditional as by a biconditional interpretation (i.e., $c \approx .50$ in all groups except the experimental group in Experiment 2). If a biconditional interpretation was taken, it was almost always a bidirectional one rather than a case distinction (parameter *x*). The direction of warranted inferences (parameter *d*) was predominantly the forward direction, from letters to numbers; a reversal occurred infrequently. Finally, the antecedent was mostly seen as sufficient rather than necessary (parameter *s*), and reasoning was most of the time irreversible (parameter *i*). In summary, three quarters of the responses arose from the inference part of the model. These were split more or less equally between responses based on a bidirectional, biconditional interpretation and a conditional interpretation with p sufficient for q; reasoning only rarely took the invisible sides of the cards into account.

There were also clear differences between the control groups and the ones with hint(s). The hint in Experiment 1 targeted the irreversibility parameter, and *i* was in fact significantly reduced; that is, there was more reasoning from the invisible sides of the cards in the experimental group than in the control group. Moreover, this was the only significant difference between the two groups.

In Experiment 2, this hint was used along with two other hints, one of which explicitly compromised a bidirectional, biconditional interpretation. The results replicate those from Experiment 1 in that irreversibility *i* was significantly reduced in the experimental group relative to the control group. In addition, the conditional interpretation was significantly more frequent in the experimental group than in the control group (parameter *c*), and if a biconditional interpretation was taken, it was significantly less frequently the bidirectional one (parameter *x*). Pointing out that the rule and its converse are not equivalent effectively conveyed that a bidirectional, biconditional interpretation is not appropriate, but it might also have had the effect to decrease the likelihood of reversals in direction, given a conditional interpretation of the rule. In fact, parameter *d* for direction increased from the control group to the experimental group (i.e., reversals became less frequent), but the increase was not significant ($p = .11$), perhaps because of a ceiling effect, given that most reasoners see the rule as one warranting forward inferences (i.e., from letters to numbers) to begin with.

Table 4
Parameter Estimates of the Inference-Guessing Model in Experiments 1 and 2

	Control		Hint(s)			
Parameter	Estimate	CI	Estimate	CI	χ^2 (1)[a]	*p*
			Experiment 1			
p	.50	0.27, 0.72	.73	0.52, 0.93	2.27	.13
$\bar{p}$	.22	0.05, 0.39	.15	0.01, 0.28	0.45	.50
q	.42	0.22, 0.62	.53	0.31, 0.76	0.54	.46
$\bar{q}$	.38	0.20, 0.56	.41	0.17, 0.64	0.03	.86
a	.78	0.68, 0.87	.76	0.64, 0.89	0.03	.87
c	.47	0.40, 0.54	.50	0.43, 0.57	0.32	.57
x	.97	0.92, 1.00	.94	0.88, 1.00	0.42	.52
d	.87	0.76, 0.98	.76	0.64, 0.88	1.92	.17
s	.92	0.87, 0.97	.86	0.80, 0.92	3.12	.08
i	.90	0.85, 0.94	.81	0.75, 0.88	5.20	.02
			Experiment 2			
p	.35	0.12, 0.57	.33	0.17, 0.48	0.02	.89
$\bar{p}$	.34	0.12, 0.55	.39	0.25, 0.54	0.17	.68
q	.31	0.13, 0.49	.54	0.39, 0.68	3.33	.07
$\bar{q}$	.40	0.24, 0.57	.62	0.46, 0.79	3.35	.07
a	.79	0.71, 0.87	.73	0.65, 0.82	0.89	.35
c	.49	0.43, 0.56	.69	0.63, 0.76	17.76	<.01
x	.97	0.91, 1.00	.83	0.72, 0.94	5.49	.02
d	.87	0.78, 0.95	.96	0.89, 1.00	2.62	.11
s	.93	0.87, 1.00	.92	0.86, 0.98	0.13	.72
i	.92	0.88, 0.96	.81	0.75, 0.87	9.74	<.01

Note. CI = 95% confidence interval.
[a] Chi-square test for differences between the control group and the experimental group.

Experiments 3 and 4: Conditionality and Sufficiency

Experiments 3 and 4 were inspired by experiments by Feeney and Handley (2000) and by Handley, Feeney, and Harper (2002) in which the standard rule was accompanied by a second rule that specifies an alternative antecedent. For example, the two rules may be "If there is an *A* on the letter side, then there is a *3* on the number side" and "If there is a *B* on the letter side, then there is a *3* on the number side." Participants' task was to test the truth or falsity of the first rule. It was found that selection of the q card was strongly suppressed as well as, to a lesser extent, selection of the $\bar{p}$ card.

Introducing an alternative antecedent is likely to block a biconditional interpretation of the first rule (Rumain et al., 1983), both in terms of a bidirectional and a case distinction interpretation. In addition, specifying an alternative antecedent strongly suggests that the antecedent of the first rule is not a necessary condition for

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the consequent although it may be sufficient. In terms of the present model parameters, we therefore predict effects on the parameter *c* for conditionality versus biconditionality and on the parameter *s* for perceived sufficiency versus necessity. Both parameters should increase when an alternative antecedent is introduced.

In Experiments 3 and 4, a control group with one rule is compared with an experimental group with two rules. In Experiment 3, the second rule specifies an alternative antecedent as just exemplified; in Experiment 4, it specifies an alternative consequent, for example, "If there is an *A* on the letter side, then there is a *4* on the number side."

Specifying an alternative consequent should again block a biconditional interpretation; in addition, specifying an alternative consequent strongly suggests that the antecedent is not a sufficient condition, although it may be necessary. Thus, effects are again expected on the parameter *c* for conditionality versus biconditionality as well as on the parameter *s* for perceived sufficiency versus necessity. In Experiment 4, *c* should increase like in Experiment 3, but in contrast to Experiment 3, *s* should decrease.

Remember that the inference-guessing model with relaxed assumptions allows for different parameters s_f, s_b, and s_{fb} for perceived sufficiency versus necessity as a function of perceived direction (s_f is the parameter for perceived sufficiency versus necessity of p for q if the rule is seen as inviting forward inferences; s_b for perceived sufficiency versus necessity of q for p if the rule is seen as inviting backward inferences; s_{fb} for perceived sufficiency versus necessity of p for q and q for p under a bidirectional, biconditional interpretation). In terms of these parameters, an increase (a decrease) is, strictly speaking, expected in Experiment 3 (in Experiment 4) only for s_f, the perceived sufficiency versus necessity of p for q in forward inferences, but not for s_b, the perceived sufficiency of q for p in backward inferences. This might cause lack of fit for the inference-guessing model that we use as default model, but if the inference-guessing model fits, the joint parameter *s* is dominated by s_f because reversals from forward to backward direction occur infrequently, so that we can expect to see the effect in parameter *s*.

Method

Participants. Participants were sampled via the Internet. In Experiment 3, there were 335 and 300 participants in control group and experimental group, respectively; in Experiment 4, these numbers were 300 and 301. Demographic information about the samples of participants is given in Appendix B.

Procedure. The procedures followed those of the control group of Experiment 1 unless where explicitly mentioned otherwise. In the experimental groups, the instructions were changed as follows. The sentences "For the set of cards, a rule has been stated. It is: . . ." were replaced by "For the set of cards, two rules have been stated. They are: . . . ." The additional second rule specified an alternative antecedent in Experiment 3 and an alternative consequent in Experiment 4. Participants in experimental groups were told that their task was to test the truth or falsity of the first rule rather than of only "the rule." The words "first rule" were set in boldface. Following Feeney and Handley (2000, Experiment 1), the characters shown on the four cards included the letter specified as alternative antecedent in Experiment 3 and the number specified as alternative consequent in Experiment 4. All used letters and numbers were randomly sampled for each participant anew with the same restrictions as in Experiments 1 and 2.

Results and Discussion

The pattern frequencies are shown in Table 1. Descriptively, the strongest effect of a second rule appears to be a reduction in (p, q) selections in favor of selections of the p card alone. In addition, in Experiment 3, selections involving the $\bar{p}$ card are generally decreased. In Experiment 4, the selection of the q card alone is increased in the presence of a second rule.

In Experiment 3, the individual card-selection frequencies for the p, $\bar{p}$, q, and $\bar{q}$ cards were, in order, 247, 58, 168, and 75 in the control group ($n = 335$) and 217, 24, 102, and 51 in the experimental group ($n = 300$). Previous results for an additional rule with alternative antecedent were replicated (Feeney & Handley, 2000; Handley et al., 2002): In the experimental group, selections of the q card and the $\bar{p}$ card were significantly reduced relative to the control group, $\chi^2(1) = 16.89$, $p < .01$, and $\chi^2(1) = 12.21$, $p < .01$, respectively, whereas selections of the p card and the $\bar{q}$ card were not significantly affected, $\chi^2(1) = 0.16$, $p = .69$, and $\chi^2(1) = 2.89$, $p = .09$, respectively.

Model analyses for Experiment 3. The inference-guessing model described the data well, $G^2(10) = 11.96$, $p = .29$. The parameter values are shown in Table 5 along with chi-squared tests for differences between control group and experimental group. As can be seen, an alternative antecedent strongly increased parameter *c* for conditionality versus biconditionality. In addition, there was the expected significant increase in perceived sufficiency (parameter *s*). Specifying a second rule with alternative antecedent also significantly decreased the proportion *a* of responses governed by the inference submodel, reflecting perhaps a greater degree of confusion in the condition with two rules. Finally, there was a nonsignificant tendency for an additional rule to increase irreversibility *i*. Thus, the effects of the second rule were to make the interpretation of the first rule more similar to a conditional with p sufficient for q; the reasoning process itself was not improved. If anything, deeper reasoning (i.e., reasoning from the invisible sides) was slightly less frequent with two rules and the proportion of responses governed by the independence model increased.

Model analyses for Experiment 4. Fitting the inference-guessing model to the data of Experiment 4 led to a goodness-of-fit statistic just outside the conventional 5% region of acceptable model fit, $G^2(10) = 18.48$, $p = .047$. For this reason, the inference-guessing model with relaxed assumptions was fitted. This model achieved an acceptable goodness of fit according to conventional criteria, $G^2(6) = 11.30$, $p = .08$. The parameter values are also shown in Table 5 along with significance tests for differences between control group and experimental group.

As can be seen, there is again a pronounced and significant effect on the parameter *c* for conditionality versus biconditionality as expected. Furthermore, there is the expected significant decrease in the perceived sufficiency of p for q (parameter s_f) when there is an alternative consequent. Like in Experiment 3, reasoning is more shallow given two rules than with one rule (parameter *i*), and the proportion of responses governed by the inference submodel is decreased, but both of these trends did not attain significance.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5
Parameter Estimates of the Inference-Guessing Model in Experiments 3 and 4

	Control		Two rules			
Parameter	Estimate	CI	Estimate	CI	$\chi^2(1)$[a]	*p*
			Experiment 3			
p	.49	0.30, 0.67	.47	0.32, 0.63	0.01	.92
$\bar{p}$	.28	0.12, 0.45	.19	0.08, 0.30	0.79	.37
q	.38	0.22, 0.53	.34	0.21, 0.47	0.15	.70
$\bar{q}$	.51	0.33, 0.68	.44	0.31, 0.56	0.37	.55
a	.77	0.69, 0.85	.62	0.50, 0.74	4.50	.03
c	.49	0.43, 0.56	.75	0.68, 0.83	25.90	< .01
x	.99	0.94, 1.00	.91	0.77, 1.00	0.79	.37
d	.79	0.70, 0.88	.83	0.75, 0.91	0.50	.48
s	.89	0.83, 0.94	.99	0.94, 1.00	5.61	.02
i	.91	0.86, 0.95	.97	0.92, 1.00	3.55	.06
			Experiment 4			
p	.45	0.24, 0.66	.64	0.49, 0.80	2.14	.14
$\bar{p}$	.22	0.06, 0.39	.11	0.01, 0.20	1.64	.20
q	.49	0.24, 0.49	.61	0.42, 0.80	0.49	.48
$\bar{q}$	.33	0.12, 0.57	.35	0.14, 0.55	< 0.01	.97
a	.75	0.64, 0.87	.62	0.45, 0.78	2.04	.15
c	.43	0.35, 0.56	.80	0.67, 0.93	30.68	< .01
x	.97	0.92, 1.00	1.00	0.88, 1.00	0.32	.57
d	.90	0.76, 0.95	.79	0.69, 0.89	1.69	.19
s_f	.98	0.92, 1.00	.89	0.83, 0.95	4.50	.03
s_b	.67	0.00, 1.00	.85	0.61, 1.00	0.15	.69
s_{fb}	.86	0.79, 1.00	.88	0.72, 1.00	0.05	.83
i	.92	0.88, 0.96	.98	0.93, 1.00	3.05	.08

Note. CI = 95% confidence interval.
[a] Chi-square test for differences between the control group and the experimental group.

Discussion

The model parameters again provided a psychologically meaningful mapping of the experimental manipulation. According to the present analysis, an additional rule with an alternative antecedent affects the interpretation of the first rule as suggested by Rumain et al. (1983), Feeney and Handley (2000), and Handley et al. (2002), among others. Specifically, the rule was more frequently seen as conditional rather than biconditional, and perceived sufficiency increased. An additional rule with an alternative consequent similarly biased interpretation away from a biconditional interpretation and decreased the perceived sufficiency of p for q as expected.

Note that the present analyses go above previous studies in several respects. Experiment 4 is the first to introduce a rule with an alternative consequent as far as we know. More important, the model accounts for all 16 pattern frequencies as well as for the changes therein rather than for only the individual card frequencies. In addition, by incorporating the independence model as an alternative processing path, fitting the inference-guessing model implies a test of the question of whether the second rule has effects on the inference part of the model or alternatively on the independence part. Oaksford (2002) has argued that Feeney and Handley's (2000) results can alternatively be accommodated by their ODS model that is a submodel of the independence model. As it turns out, the effects of an additional rule map on the inference part of the inference-guessing model, whereas there are no effects on the independence part.

Experiment 5: Direction

Experiment 5 was focused on the parameter *d* for direction of warranted inferences. Two experimental manipulations were aimed at altering direction from the predominant forward direction (i.e., from letters to numbers) toward the reversed one (i.e., from numbers to letters), that is, at decreasing parameter *d*. One manipulation that has this effect in conditional inference tasks is to rephrase the rule "if *A*, then *3*" as "*A*, only if *3*" (Evans, 1993; Evans et al., 1993, chap. 2; Evans, Legrenzi, & Girotto, 1999, Experiment 3). We also tried a manipulation of phrase order, namely to mention the consequent first rather than as usual second, although there was little evidence in the literature for an effect of phrase order. In all, there were four groups in Experiment 5 with different phrasings of the rule. They were

> If *A*, *3*: If there is an *A* on the letter side, then there is a *3* on the number side.
>
> *3*, if *A*: There is a *3* on the number side, if there is an *A* on the letter side.
>
> *A*, only if *3*: There is an *A* on the letter side, only if there is a *3* on the number side.
>
> Only if *3*, *A*: Only if there is a *3* on the number side, there is an *A* on the letter side.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 6
Parameter Estimates of the Inference-Guessing Model in Experiment 5

Parameter	If A, 3	3, if A	A, only if 3	Only if 3, A	$\chi^2(3)$[a]	*p*
p						
Estimate	.41	.48	.67	.54	4.68	.20
CI	0.18, 0.64	0.28, 0.68	0.47, 0.87	0.36, 0.72		
$\bar{p}$						
Estimate	.40	.36	.15	.20	6.85	.08
CI	0.17, 0.62	0.17, 0.55	0.05, 0.26	0.09, 0.32		
q						
Estimate	.37	.47	.43	.46	0.57	.90
CI	0.16, 0.58	0.26, 0.67	0.28, 0.58	0.30, 0.62		
$\bar{q}$						
Estimate	.36	.36	.54	.45	2.63	.45
CI	0.15, 0.56	0.17, 0.55	0.32, 0.77	0.27, 0.64		
a						
Estimate	.82	.78	.68	.61	12.65	< .01
CI	0.74, 0.90	0.70, 0.87	0.57, 0.79	0.49, 0.73		
c						
Estimate	.47	.56	.48	.51	3.88	.27
CI	0.41, 0.54	0.49, 0.62	0.40, 0.56	0.42, 0.60		
x						
Estimate	.99	1.00	.92	.92	4.17	.24
CI	0.94, 1.00	0.94, 1.00	0.84, 1.00	0.80, 1.00		
d						
Estimate	.82	.85	.43	.41	24.21	< .01
CI	0.73, 0.91	0.77, 0.94	0.26, 0.59	0.24, 0.59		
s_f						
Estimate	1.00	.99	.90	.86	6.11	.11
CI	0.92, 1.00	0.94, 1.00	0.79, 1.00	0.69, 1.00		
s_b						
Estimate	.88	.76	1.00	1.00	2.23	.53
CI	0.59, 1.00	0.45, 1.00	0.82, 1.00	0.78, 1.00		
s_{fb}						
Esimate	.90	.85	.90	.63	17.60	< .01
CI	0.84, 0.97	0.77, 0.92	0.83, 0.97	0.49, 0.78		
i						
Estimate	.93	.93	.91	.89	2.06	.56
CI	0.89, 0.96	0.90, 0.97	0.87, 0.96	0.83, 0.95		

Note. CI = 95% confidence interval.
[a] Chi-square test for differences among the four groups.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Method

Participants. Participants were sampled via the Internet. In Experiment 5, there were 345, 339, 324, and 300 participants in the groups labeled "If *A*, *3*," "*3*, if *A*," "*A*, only if *3*," and "Only if *3*, *A*," respectively. Demographic information about the sample of participants is given in Appendix B.

Procedure. The procedures followed those of the control group of Experiment 1, the only difference being the use of differently phrased rules.

Results and Discussion

The pattern frequencies are shown in Table 1. Descriptively, selections of p alone and of (p, q) were substantially decreased in the *only if* groups. Simultaneously, there was a marked increase in the selection of q alone along with a less pronounced increase in the selection of (p, $\bar{q}$) in the *only if* groups. In the group with "Only if *3*, *A*" there is furthermore an increase in the ($\bar{p}$, $\bar{q}$) pattern. Phrase order appears to have had little effect.

The inference-guessing model did not describe the data well, $G^2(20) = 60.46$, $p < .01$. The inference-guessing model with relaxed assumptions did, however, provide an acceptable account of them, $G^2(12) = 19.34$, $p = .08$. The parameter estimates are shown in Table 6 along with chi-squared tests for differences between the different groups. As expected, *only if* led to a significant increase in reversals of direction of warranted inferences (i.e., to a decrease of parameter *d*).

In addition, there were significant effects on the parameters *a* and s_{fb}: In the groups with *only if*, fewer responses were guided by the inference submodel than in the groups with *if*, perhaps reflecting increased confusion when *only if* was used or a greater variety of idiosyncratic interpretations and response patterns. Klauer (1994) found that it took longer to understand rules using *only if* than rules using *if*, suggesting that it is more difficult to comprehend the former type of rule than the latter (see also Ormerod, Manktelow, & Jones, 1993). Finally, in the group with "Only if *3*, *A*," parameter s_{fb} was significantly reduced. That is, under a bidirectional, biconditional reading of the rule, p was more frequently seen as necessary for q and q as necessary for p than in all other groups, as though the rule "Only if *3*, *A*" was interpreted as "If not *3*, not *A*, and if not *A*, then not *3*," leading via MP to the selection of ($\bar{p}$, $\bar{q}$), given irreversible reasoning. In fact, Braine (1978) argued that "only X" means "no Y other than

X" and in his view, the rule "Only if *3*, *A*" is thereby a paraphrase of "If other than *3*, not *A*" meaning "If not *3*, then not *A*". According to the present analysis, such paraphrasing or translation occurs primarily under a biconditional interpretation of the rule (the effect on sufficiency was seen only for biconditional interpretations). The effect is consistent with many findings in the conditional reasoning literature suggesting that instances with negated antecedent and consequent are treated differently under *only if* than under *if* (e.g., Johnson-Laird, Byrne, & Schaeken, 1992); for example, ($\bar{p}$, $\bar{q}$) is more often selected as true in the truth table task for *only if* than for *if* (Ormerod et al., 1993).

Experiment 6: The Parameters of the Independence Model

In Experiment 6, we used nonstandard arrays of cards (e.g., Hardman, 1998; Oaksford, Chater, Grainger, & Larkin, 1997). Specifically, we presented five rather than four cards for selection. The additional fifth card was either another p card, another $\bar{p}$ card, another q card, or another $\bar{q}$ card, defining the four groups of Experiment 6. There was little reason to expect effects of this manipulation on the inference submodel. However, it seemed likely that heuristic processes involved in superficial card selections such as guessing would be affected. For example, when guessing which cards to select, decision makers might consider it sufficient to select only one card, if any, of the doubled kind, leading to lowered parameters for each of the doubled cards in the independence model.

Method

Participants. Participants were sampled via the Internet. In Experiment 6, there were 309, 300, 318, and 310 participants in the groups with an additional p card, $\bar{p}$ card, q card, and $\bar{q}$ card, respectively. Demographic information about the sample of participants is given in Appendix B.

Procedure. Procedures and instructions were as in the control group of Experiment 1 with the following modifications. The rule was "If there is a vowel on the letter side of the card, then there is an even number on the number side of the card." Numbers and letters were randomly sampled as before, but instead of four cards, five cards were presented for selection. Members of the p group saw two cards with different vowels, members of the $\bar{p}$ group saw two cards with different consonants, members of the q group saw two cards with different even numbers, and members of the $\bar{q}$ group saw two cards with different odd numbers. The order in which the five cards were arranged from left to right was randomized for each participant anew.

Results and Discussion

For the analyses, the rightmost additional card was discarded (e.g., in the p group only the first p card from the left is considered) so that the data are based on the remaining four cards and have the same format as in the previous experiments. The pattern frequencies are shown in Table 1. Descriptively, selection of patterns involving the doubled kind of card were depressed. For example, 62% of the members of the p group selected the p card, whereas that proportion exceeded 73% in all other groups. Similarly, for the $\bar{p}$ group, these proportions were 13% and 23%; for the q group, 41% and 52%; and for the $\bar{q}$-group, 25% and 29%.

The inference-guessing model provided a good description of the data, $G^2(20) = 19.62$, $p = .48$. The parameter estimates are shown in Table 7 along with significance tests for differences between the four groups. As can be seen, the manipulation had the desired effect on the parameters of the independence submodel. In particular, parameter p for the p card was significantly depressed in the p group, and parameter $\bar{p}$ for the $\bar{p}$ card was significantly depressed in the $\bar{p}$ group. There were nonsignificant trends for an analogous effect in the q group and in the $\bar{q}$ group.

In more focused analyses, it was found that for each card p, $\bar{p}$, q, and $\bar{q}$ considered separately, it was possible to set the corresponding parameter of the independence submodel equal across the three groups in which that (kind of) card was not doubled, largest $\Delta G^2(2) = 2.92$, smallest $p = .23$. In addition, for each card, the common parameter in these groups was significantly larger than the parameter in the group in which that card was doubled, smallest $\Delta G^2(1) = 5.51$, largest $p = .02$, so that there was evidence for effects on each of the targeted parameters of the independence submodel. Adding another instance of one of the four kinds of cards did not significantly affect any of the parameters of the inference part of the model.

General Discussion

The present data sets showed the pattern that is characteristic of the WST (see Table 1). There were few selections of the logically correct pattern; the most frequent selections were to select the p card alone or both the p card and the q card. Like in previous studies, helpful instructions (Experiments 1 and 2) entailed only modest increases of logically correct choices. The effect pattern documented by Feeney and Handley (2000) and Handley et al. (2002) for a second rule with alternative antecedent was exactly replicated in the frequency data (see Experiment 3); that is, selections of the q card and of the $\bar{p}$ card were depressed.

There were a number of new findings that can be appreciated without the model analyses. Introducing a second rule with alternative consequent had complementary effects to those of a second rule with alternative antecedent. An alternative consequent, like an alternative antecedent, reduced the selection of (p, q) and of ($\bar{p}$, $\bar{q}$), but unlike an alternative antecedent, it enhanced the selections of the q card alone and of the $\bar{p}$ card alone, that is, of the patterns (0,1,0,0) and (0,0,1,0) (see Experiment 4). Introducing an additional card that was either another p card, another $\bar{p}$ card, another q card, or another $\bar{q}$ card had the effect of decreasing the frequency of selections of the first exemplar (counted from left to right) of each of these cards (see Experiment 6).

The goal of the present article was to attempt a complete, psychologically substantiated account of selection data. For that purpose, a family of models was defined that is related to a number of different theories of the WST. The independence model is a supermodel (i.e., a model that relaxes assumptions of) the ODS model discussed by Hattori (2002) and Oaksford and Chater (2003a), and it is linked to Evans' (1995) relevance theory. The heuristic-analytic model can be seen as a quantitative specification of Evans' (1984) heuristic-analytic theory. The inference model can be seen as a quantitative specification of approaches that assume a role for deductive reasoning in the WST (e.g., Evans & Over, 2004, chap. 9; Johnson-Laird, 1995). The inference-guessing model is consistent with Evans' (2006) revised heuristic-analytic theory.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 7
Parameter Estimates of the Inference-Guessing Model in Experiment 6

Parameter	Double p	Double $\bar{p}$	Double q	Double $\bar{q}$	$\chi^2(3)$[a]	*p*
p						
Estimate	.32	.61	.50	.69	10.26	.02
CI	0.16, 0.48	0.46, 0.75	0.34, 0.66	0.52, 0.87		
$\bar{p}$						
Estimate	.23	.10	.37	.22	10.39	.02
CI	0.10, 0.35	0.02, 0.19	0.23, 0.52	0.09, 0.35		
q						
Estimate	.51	.46	.31	.51	5.84	.12
CI	0.36, 0.66	0.32, 0.59	0.19, 0.43	0.34, 0.68		
$\bar{q}$						
Estimate	.39	.43	.51	.24	7.14	.07
CI	0.26, 0.51	0.28, 0.58	0.38, 0.64	0.12, 0.37		
a						
Estimate	.63	.54	.61	.52	1.56	.67
CI	0.51, 0.74	0.40, 0.69	0.51, 0.72	0.34, 0.71		
c						
Estimate	.47	.45	.49	.39	1.88	.60
CI	0.38, 0.56	0.35, 0.56	0.40, 0.57	0.27, 0.51		
x						
Estimate	.95	1.00	.84	.94	5.55	.14
CI	0.88, 1.00	0.93, 1.00	0.73, 0.95	0.82, 1.00		
d						
Estimate	.84	.78	.86	.76	0.78	.85
CI	0.62, 1.00	0.60, 0.97	0.75, 0.97	0.46, 1.00		
s						
Estimate	.84	.90	.91	.86	1.39	.71
CI	0.76, 0.93	0.83, 0.97	0.80, 1.00	0.75, 0.96		
i						
Estimate	.82	.81	.86	.79	1.94	.58
CI	0.76, 0.88	0.73, 0.90	0.80, 0.93	0.69, 0.89		

Note. CI = 95% confidence interval.
[a] Chi-squared test for differences among the four groups.

Statistical Evaluation and Implications for Other Theories of the WST

The models were statistically evaluated with model selection indices AIC and BIC and goodness of fit as criteria. Model selection on the basis of AIC and BIC takes into account that the ability of a model to fit a given data set is increased as its complexity increases; the criteria seek to identify the model with the best tradeoff between parsimony and fit. The results were relatively clear:

1. Although most parsimonious, the independence model consistently received the poorest criterion values.

2. A model of medium complexity, the inference-guessing model, overall achieved the best tradeoff between parsimony and fit across the present data sets than the other models, including the almost saturated relevance-inference-guessing model. It also fitted 16 of the 18 WST data sets at the 5% level of significance.

3. The ODS model by Oaksford and Chater (2003a) and an extended version of it that removes the independence assumption built into that model performed significantly worse than the inference-guessing model in terms of AIC and BIC. The ODS model and the extended ODS model had to be rejected at the 1% level of significance for, respectively, 18 and 17 of the 18 WST data sets in goodness-of-fit tests.

The heuristic-analytic model did not perform as well as the inference-guessing model. Although the heuristic-analytic model can be seen as one possible quantitative specification of Evans' (1984) heuristic-analytic theory, its poorer performance does not imply that the theory it specifies is decisively refuted. Other specifications of it, using different auxiliary assumptions in specifying the theory, may exist that produce models with better fit. For example, heuristic relevance judgments might be made configurally rather than for each card locally, as currently assumed by the heuristic-analytic model. Similarly, although the present data do not support the ODS model variant proposed by Oaksford and Chater (2003a), other model variants of the account by ODS may exist that provide better accounts of the data.

On the positive side, the success of the inference-guessing model speaks to dual-process theories, such as Evans' (2006) revised heuristic-analytic theory, that remove the constraint of a strictly sequential interaction of heuristic and analytic processes as embodied in the heuristic-analytic model. The success of the inference-guessing model implies that such theories are consistent with the WST data: Specifications of them exist that account for the data.

The Inference-Guessing Model: Modal Parameter Values and Laboratory Data

With regard to the modal parameter values obtained for the inference-guessing model across the present experiments, approximately $a = 75\%$ of the responses stemmed from the inference

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

submodel, but only about $1 - i = 10\%$ of these reflected what we called reversible reasoning that takes the invisible sides of the cards into account in suppositional inferences. Twenty-five percent of the responses are governed by the independence submodel.

For the card-selection parameters p, $\bar{p}$, q, and $\bar{q}$ of the independence submodel, their means (standard deviations) across the analyzed WST data sets were, in order, .52 (.15), .26 (.11), .45 (.08), and .43 (.10). The effect of card type was significant, $F(3, 51) = 16.45$, $p < .01$, as were all pairwise comparisons between cards, smallest $t(17) = 2.13$, largest $p = .048$, except that between q and $\bar{q}$, $t(17) = 0.47$, $p = .64$. The observed order, $p > q > \bar{q} > \bar{p}$, is of course compatible with the idea that processes as captured in ODS are responsible for the responses governed by this submodel or that these responses are guided by relevance judgments for the individual cards as proposed by Evans (1995).

With regard to the interpretational parameters of the inference submodel, the rule was interpreted as biconditional almost as often as conditional ($c \approx .50$), it was usually seen as inviting forward inferences ($d \approx .85$), and the antecedent in the perceived direction was seen as sufficient for the consequent ($s \approx .90$). When the rule was interpreted biconditionally, it was usually seen as inviting the same forward as backward inferences; that is, it was interpreted bidirectionally rather than in terms of what we have called case distinctions ($s \approx .95$). In short, problem solvers were roughly equally divided between a bidirectional, biconditional interpretation and a conditional interpretation directed from numbers to letters with p sufficient for q.

Published data sets on the WST usually do not report the selection frequencies for all 16 patterns, and if they do, the sample size is usually too small to permit a model-based analysis. Klauer (1999, based on Oaksford & Chater, 1994) compiled data from nine published studies that provided complete records of pattern frequencies with a total of 257 participants for affirmative rules (see the row labeled *Old* in Table 1). In addition to these old data, we collected new data from 233 students from the Universities of Bonn, Freiburg, and Mannheim, Germany, with a questionnaire version of the standard WST condition of our Internet-based experiments (see the row labeled *New* in Table 1). How do these data that were collected by conventional means compare to the present data that were collected over the Internet? The inference-guessing model provided acceptable fits to the old and the new data, $G^2(5) = 8.49$, $p = .13$, and $G^2(5) = 10.11$, $p = .07$, respectively. Table 8 shows parameter estimates for the old and the new data, respectively, along with the estimates averaged over the control groups from Experiments 1 to 5 (the inference-guessing model fitted the data from each control group adequately). For the old and the new data, $a = 70\%$ and $a = 90\%$, respectively, of the responses were based on the inference submodel, but $i = 94\%$ and $i = 87\%$ of these reflect only shallow irreversible reasoning. Parameter $c = 25\%$ for conditionality versus biconditionality is unusually small for the old data; that is, biconditional interpretations prevailed in these data. See the unusually high frequency of (p, q) selections ($f = 144$) relative to selections of p alone ($f = 57$) in the raw data. The old data, being compiled over different studies with different procedures, should perhaps be interpreted with caution. The parameter estimates for the new data agree with the estimates for the Internet data reasonably well, given the confidence intervals shown in Table 8 and taking into account that the estimates for the control groups are additionally associated with estimation error.

Table 8
Parameter Estimates of the Inference-Guessing Model for Control Groups of Experiments 1–5 (Averages), for the Old and the New Data, and for the Inference Task

Parameter	Control group	Old	New	Inference task
p				
Estimate	.44	.89	.50	.61
CI		0.79, 0.99	0.19, 0.82	0.38, 0.84
$\bar{p}$				
Estimate	.25	.17	.44	.30
CI		0.03, 0.32	0.14, 0.74	0.10, 0.49
q				
Estimate	.37	.46	.44	.43
CI		0.29, 0.63	0.15, 0.73	0.22, 0.64
$\bar{q}$				
Estimate	.41	.42	.70	.32
CI		0.16, 0.67	0.38, 1.00	0.15, 0.50
a				
Estimate	.79	.70	.90	.77
CI		0.52, 0.87	0.84, 0.97	0.67, 0.88
c				
Estimate	.47	.25	.49	.55
CI		0.14, 0.36	0.42, 0.56	0.48, 0.62
x				
Esimate	.97	1.00	.95	1.00
CI		0.96, 1.00	0.90, 1.00	0.93, 1.00
d				
Estimate	.86	.89	.88	1.00
CI		0.76, 1.00	0.81, 0.95	0.92, 1.00
s				
Estimate	.91	.96	.96	.96
CI		0.93, 1.00	0.91, 1.00	0.92, 1.00
i				
Estimate	.91	.94	.87	.84
CI		0.89, 0.99	0.82, 0.92	0.79, 0.90

Note. CI = 95% confidence interval.

Experimental Validation of the Model Parameters

How did the experimental manipulations map on the model parameters of the inference-guessing model? The experimental manipulations that targeted the parameters c, x, d, s, and i of the inference submodel were successful in affecting these parameters.

1. A hint designed to encourage reversible reasoning had the expected effect on parameter i (see Experiments 1 and 2).

2. Rule clarification aimed at discouraging a bidirectional, biconditional interpretation increased conditionality at the expense of biconditionality (parameter c), and it shifted remaining biconditional interpretations towards case distinctions rather than bidirectional interpretations (parameter x; see Experiment 2).

3. A more subtle manipulation to discredit biconditional interpretations was to introduce a second rule with alternative antecedent (see Experiment 3) and alternative con-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

sequent (see Experiment 4). Both manipulations had the expected effects to increase parameter *c* for conditionality versus biconditionality.

4. A second rule with alternative antecedent increased perceived sufficiency (parameter *s*) as expected, whereas a second rule with alternative consequent decreased perceived sufficiency of p for q as also predicted (see Experiments 3 and 4).

5. Using *only if* rather than *if* had the expected effect of increasing the proportion of backward inferences (parameter *d*; see Experiment 5).

Thus, there is evidence that the different parameters validly capture the different psychological variables and processes that they are assumed to quantify.

We were also able to selectively influence each of the parameters p, $\bar{p}$, q, and $\bar{q}$ of the independence submodel (see Experiment 6), but the theoretical status of the experimental manipulation that led to selective influence is somewhat unclear. One possibility is to interpret it in terms of the perceived frequency or rarity of cards of different kinds. That is, adding an additional card of a given type means that cards of this type occur twice as often as cards from other categories, that is, that they are less rare. Perceived rarity is a major causal force in the account by ODS, because expected information gain generally increases as rarity of the card type is increased (see also Nickerson, 1996). The effects in Experiment 6 are thus broadly consistent with the idea that the processes tapped by these parameters are governed by expected information gain as postulated by ODS. Note that Experiment 6 thereby opens a natural route by which probabilistic effects on card selection (e.g., Green, Over, & Pyne, 1997; Kirby, 1994; Oaksford, Chater, & Grainger, 1999; but see Oberauer, Wilhelm, & Rosas Diaz, 1999) can influence selection data (other than by changing rule interpretation) in the present model, namely via the card-selection parameters of the independence submodel.

As noted in the introduction, one reason for the interest in the WST has been the low number of normatively correct responses, suggesting that the WST elicits processes that differ fundamentally from those involved in conditional-inference tasks (e.g., Evans, 1995; Lucas & Ball, 2005; Oaksford & Chater, 1994; O'Brien, 1995). In conditional-inference tasks, participants typically draw or evaluate conclusions on the basis of a conditional rule "if p, then q" and a minor premise such as p. However, low proportions of normatively correct responses have also been observed in traditional conditional-inference tasks when the format of the task is made more similar to the WST (Evans & Handley, 1999, Experiment 3). To illustrate this point, we collected another data set of 300 participants over the Internet in which we replaced the WST to test the truth or falsity of the rule by the inference task to check those cards for which "the rule predicts something about the letters or numbers on the invisible side of the card." This instruction was meant to elicit the constraint-seeking behavior postulated by the inference submodel of the inference-guessing model: According to that submodel, cards are selected for which a constraint for the invisible side is deduced from the rule. All procedural details were otherwise as in the control groups of the above WST experiments. The data are shown in the last row of Table 1. As can be seen, they exhibited all the characteristics of typical WST data. This suggests that procedural differences typically confounded with the comparison of WST and that inference tasks may be responsible for previously observed dissimilarities in outcome patterns rather than fundamental differences in process. Moreover, the inference-guessing model provided a good fit to the data, $G^2(5) = 6.97$, $p = .22$, with parameters that were reasonably well matched to what was observed for the WST data (see Table 8). This suggests that the inference-guessing model may also be useful in the analysis of data from inference tasks.

How does the inference-guessing model explain the low proportion of normatively correct responses, given that conditional inference constitutes a central aspect of it? Note first of all that a low proportion of normatively correct responses was also obtained in the inference task just discussed along with the other characteristics of typical WST data. According to the inference-guessing model, between 20% and 30% of the responses are governed by the independence model; they do not reflect conditional reasoning at all. These processes produce (p, $\bar{q}$) selections only randomly, typically with a probability of less than .10. In addition, as indicated by estimates of parameter *i*, about 90% of those participants who do engage in propositional reasoning forget to apply the spontaneously available inferences to the invisible sides of the cards; their behavior can be characterized as driven by System 1 along the lines suggested by Oberauer (2006; see The Inference Model section). This always results in card selections different from (p, $\bar{q}$). The remaining reasoners, typically 5% of the total population, reason from the invisible sides of the cards (reversible reasoning) using deeper reasoning tactics that would be ascribed to System 2 according to Oberauer (2006). However, these persons do not all share the standard conditional interpretation of the rule, resulting in less than 50% of (p, $\bar{q}$) selections for this group. Thus, somewhat paradoxically, a substantial proportion of the few (p, $\bar{q}$) selections in the WST are not caused by inferential reasoning. Rather, they arise from card selections of persons behaving as described by the independence part of our model. Nevertheless, the proportion of responses based on deeper reversible reasoning in (p, $\bar{q}$) selections is of course sufficiently high to account for the significant correlations that have been found between markers of general cognitive ability and performance in the WST (Stanovich & West, 1998).

Limitations and Open Questions

The present approach is roughly consistent with the individual-differences data suggesting three groups of reasoners (Newstead et al., 2004): A high-ability group with correct solution may correspond to the subset of responses with reversible reasoning under a conditional interpretation of the rule with p sufficient for q; a group of high but somewhat lower ability responding consistently across selection tasks may correspond to the subset of remaining responses governed by the inference submodel, most of them based on shallower irreversible reasoning; a third group with still lower ability and inconsistent responding across selection tasks may include responses governed by the independence model. It would be interesting to link the present model analyses with data on individual differences.

The present experiments were focused on validating the parameters of the inference-guessing model. It would now be interesting

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

to see, as a next step, how factors that have been manipulated in previous research on the WST such as kind of instruction, materials, time pressure, and so forth map on the different model parameters. For example, strong effects on selection behavior occur in the negations paradigm in which matching bias is found, and we are working on applying the inference-guessing model to the negations paradigm. Another obvious next step is to apply the model to the thematic selection task that has generated a vast literature in its own right.

Processing-tree models constrain, but do not completely determine, possible cognitive architectures that realize the modeled processes, as pointed out by Evans (in press). For example, processes governed by the independence submodel and the inference submodel might run in parallel with only one set of outputs eventually gaining access to output routines. Alternatively, the two sets of processes might be triggered exclusively or sequentially. The inference-guessing model is silent with respect to many, but not all, questions of temporal ordering. Similarly, the inference-guessing model per se is silent about the precise manner in which pragmatic influences and/or background knowledge elicit the different interpretations and spontaneous inferences, and in this regard we simply draw on the arguments and findings in the reasoning literature. The inference-guessing model and the present series of experiments identify a set of factors involved in WST performance that taken together in appropriate combination account for pattern selections quantitatively and completely.

Conclusions

In concluding, we can draw a couple of firm conclusions. Selection-task data are highly configural. Analyzing only individual card frequencies is a gross oversimplification ignoring large parts of the information that is present in such data. This is also true for effects that manipulations have on selection data. The frequency manipulation in Experiment 6 came closest to an effect pattern with simple main effects of card type on selection data. However, it is more typical for manipulations to have complex interactive effects. For example, an effect that maps on parameter *c* for conditionality versus biconditionality (e.g., see Experiment 3) is expected to involve shifts in the frequencies of selections of p alone, q alone, $\bar{p}$ alone, $\bar{q}$ alone, (p, $\bar{q}$), and ($\bar{p}$, q) relative to the selections of (p, q), (p, $\bar{p}$), ($\bar{p}$, $\bar{q}$), (q, $\bar{q}$), and all four cards.

The fact that different manipulations selectively affected different model parameters in the present experiments means that qualitatively different and empirically dissociable configural effects contribute to selection data. Independent of the model-based analyses, these dissociations are signatures of the existence of several empirically separable causal factors that any complete account of selection data has to accommodate. The series of experiments found evidence for the assumptions about the psychological interpretations that we postulated for the different model parameters.

The present model of WST performance is in all likelihood not the last word on the abstract WST; like all science, it is preliminary. It is, however, the first to account for complete sets of selection data in the abstract WST in a psychologically interpretable manner. It describes the ensemble of observed selection-pattern frequencies and the different complex and configural effects that experimental manipulations have on such data. We hope that it will be useful as a tool to interpret the effects of experimental manipulations on the WST in psychological terms and that it will serve as a baseline or benchmark in developing new, and perhaps simpler, models that provide even better fits of WST data.

References

Ahn, W., & Graham, L. M. (1999). The impact of necessity and sufficiency information in the Wason four-card selection task. *Psychological Science, 10,* 237–242.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 40*(A), 269–297.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1–21.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98,* 182–203.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187–276.

Erdfelder, E. (2000). *Multinomiale Modelle in der kognitiven Psychologie* [Multinomial models in cognitive psychology]. Unpublished habilitation thesis, Universität Bonn, Bonn, Germany.

Evans, J. St. B. T. (1977). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology, 29,* 621–635.

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75,* 451–468.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Evans, J. St. B. T. (1993). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition, 48,* 1–20.

Evans, J. St. B. T. (1995). Relevance and reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (pp. 147–172). Hove, England: Erlbaum.

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning, 4,* 45–82.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13,* 378–395.

Evans, J. St. B. T. (in press). On the resolution of conflict in dual process theories of reasoning. *Thinking and Reasoning.*

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 52*(A), 739–769.

Evans, J. St. B. T., Legrenzi, P., & Girotto, V. (1999). The influence of linguistic form on reasoning: The case of matching bias. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 52*(A), 185–216.

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64,* 391–397.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning.* Hillsdale, NJ: Erlbaum.

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

Feeney, A., & Handley, S. J. (2000). The suppression of q card selections: Evidence for deductive inference in Wason's Selection Task. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 53*(A), 1224–1242.

Gebauer, G., & Laming, D. (1997). Rational choices in Wason's selection task. *Psychological Research, 60,* 284–293.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Goodwin, R. Q., & Wason, P. C. (1972). Degrees of insight. *British Journal of Psychology, 63,* 205–212.

Green, D. W., Over, D. E., & Pyne, R. A. (1997). Probability and choice in the selection task. *Thinking and Reasoning, 3,* 209–235.

Handley, S. J., Feeney, A., & Harper, C. (2002). Alternative antecedents, probabilities, and suppression of fallacies in Wason's selection task. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55*(A), 799–813.

Hardman, D. (1998). Does reasoning occur on the selection task? A comparison of relevance-based theories. *Thinking and Reasoning, 4,* 353–376.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55*(A), 1241–1272.

Hu, X. (1991). Statistical inference program for multinomial binary tree models [Computer software]. Irvine, CA: University of California at Irvine.

Johnson-Laird, P. N. (1995). Inferences and mental models. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (pp. 115–146). Hove, England: 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.

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology, 1,* 134–148.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51,* 1–28.

Klauer, K. C. (1994). Zur Modelltheorie des aussagenlogischen Schluβfolgerns: Zeitliche Faktoren beim Konstruieren und Anwenden mentaler Modelle [The mental model theory of propositional reasoning: Temporal factors in constructing and applying mental models]. *Sprache & Kognition, 13,* 1–25.

Klauer, K. C. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review, 106,* 215–222.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika, 71,* 1–31.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *American Psychologist, 59,* 105–117.

Krauth, J. (1982). Formulation and experimental verification of models in propositional reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 34*(A) 285–298.

Liberman, N., & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange cheating detection or task understanding. *Cognition, 58,* 127–156.

Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking and Reasoning, 11,* 35–66.

Margolis, L. (1987). *Patterns, thinking and cognition: A theory of judgment.* Chicago, IL: University of Chicago Press.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

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

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2,* 1–31.

Oaksford, M. (2002). Predicting the results of reasoning experiments: Reply to Feeney and Handley (2000). *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55*(A), 793–798.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Oaksford, M., & Chater, N. (2003a). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10,* 289–318.

Oaksford, M., & Chater, N. (2003b). Probabilities and pragmatics in conditional inference: Suppression and order effects. In D. Hardman & L. Macchi (Eds.), *Psychological perspectives on reasoning, judgment, and decision making* (pp. 95–122). New York: Wiley.

Oaksford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5,* 193–243.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 441–458.

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology, 53,* 238–283.

Oberauer, K., Hörnig, R., Weidenfeld, A., & Wilhelm, O. (2005). Effects of directionality in deductive reasoning II: Premise integration and conclusion evaluation. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 58*(A), 1225–1247.

Oberauer, K., Wilhelm, O., & Rosas Diaz, R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking and Reasoning, 5,* 115–144.

O'Brien, D. P. (1995). Finding logic in human reasoning requires looking in the right places. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning* (pp. 147–172). Hove, England: Erlbaum.

Ormerod, T. C., Manktelow, K. I., & Jones, G. V. (1993). Reasoning with three types of conditional: Biases and mental models. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 46*(A), 653–677.

Osman, M., & Laming, D. (2001). Misinterpretation of conditional statements in Wason's selection task. *Psychological Research, 65,* 128–144.

Perham, N., & Oaksford, M. (2005). Deontic reasoning with emotional content: Evolutionary psychology or decision theory? *Cognitive Science, 29,* 681–718.

Platt, R. D., & Griggs, R. A. (1993). Facilitation in the abstract selection task: The effects of attentional and instructional factors. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 46*(A), 591–613.

Pollard, P. (1985). Nonindependence of selections on the Wason selection task. *Bulletin of the Psychonomic Society, 23,* 317–320.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256.

Reips, U.-D. (n.d.). *Experimental psychology lab.* Retrieved August 21, 2005, available from http://www.psychologie.unizh.ch/sowi/Ulf/Lab/WebExpPsyLab.html

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments.* New York: Springer.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 31,* 696–700.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: *If* is not the biconditional. *Developmental Psychology, 19,* 471–481.

Schroyens, W., Schaeken, W., & D'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic study in mental models and/or mental logic theory. *Thinking and Reasoning, 7,* 121–172.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Smalley, N. S. (1974). Evaluating a rule against possible instances. *British Journal of Psychology, 65,* 293–304.

Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning, 4,* 781–810.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645–726.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 55–79). Hillsdale, NJ: Erlbaum.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition, 22,* 742–758.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I* (pp. 135–151). Harmondsworth, England: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1972). *The psychology of reasoning: Structure and content.* London: Batsford.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix A

Parameter Heterogeneity, the Independence Model, and the Optimal Data Selection Model

Erdfelder (2000) and Klauer (2006) proposed to model heterogeneity by a latent-class extension of multinomial models in which the population of participants is partitioned into a few latent classes with different parameter values for each class. This leads to a tractable model that has been found to model existing heterogeneity adequately with as few as two or three latent classes for real data sets. We fitted the latent-class extension of the independence model with two and three latent classes to the present data sets; the two-class model used 9 parameters (4 per latent class and 1 for the proportional class sizes), and the three-class model with 14 parameters was almost saturated.

Both in terms of AIC and BIC, the latent-class models, placed in the set of models shown in Table 1, occupied the highest mean ranks below the independence model itself. Not surprisingly then, Wilcoxon tests showed that the latent-class models performed significantly worse in terms of BIC and AIC than the inference-guessing model (smallest $|Z| = 3.72$, largest $p < .01$). In terms of goodness of fit, the models with two and three classes had to be rejected at the 1% level for all 18 data sets. These results make it unlikely that the independence model performed poorly because of heterogeneity-related problems.

In related analyses, we fitted Oaksford and Chater's (2003a) optimal data selection (ODS) model in a way that releases it from the untenable independence assumption. Oaksford and Chater (1994) simulated nonindependence of card selections by letting the parameters of their model vary within certain bounds. For each set of parameter values and each card, Oaksford and Chater computed a so-called scaled information gain value predicting the card selection probability. They then computed correlations between the information gain values for any two cards, across the sampled sets of parameter values. These correlations adequately accounted for the signs (plus vs. minus) of the observed correlations between card selections for rules with affirmative components and somewhat less successfully for rules with negated components.

To implement this idea in a statistical model, we departed from Oaksford and Chater's (2003a) model and added a continuous distribution of the model parameters. Integrating over that distribution is the statistical analogue of the simulation performed by Oaksford and Chater (1994).

Oaksford and Chater's (2003a) ODS model uses the parameters a and b that correspond to the probabilities of the antecedent and consequent, respectively, of the rule "If p, then q." Furthermore, there is an exception parameter ε, fixed to the value .1, for the probability that the consequent does not occur given that the antecedent has occurred and a parameter $P(M_D)$, fixed to the value .5, for the prior probabilities of the two statistical models between which reasoners seek to decide according to the account by ODS. Parameters a and b are free to vary, and in their simulation of nonindependence, Oaksford and Chater (1994) sampled pairs of points for a and b at intervals of .025 satisfying a number of inequalities derived from (a) a so-called rarity assumption ($a \leq .2$ and $b \leq .2$ for affirmative rules), (b) logical restrictions on the parameter space, and (3) the assumption that b varies only within a narrow band close to a.

For the extension of Oaksford and Chater's (2003a) model, we assumed that a is distributed according to a beta distribution with parameters α_a and β_a, both of them real numbers larger than zero. The beta distribution is a family of distributions on the interval (0, 1) that is relatively tractable and accommodates a wide range of different distributions. In particular, its mean, $\mu = \frac{\alpha_a}{\alpha_a + \beta_a}$, can take on any value between 0 and 1, and its variance, $\mu(1-\mu)\frac{1}{1 + \alpha_a + \beta_a}$, can take on any value between 0 and the maximum possible variance, $\mu(1 - \mu)$, that a random variable in the interval (0, 1) with mean μ can have. Note that a thereby effectively satisfies the rarity assumption if it has a small mean and not more than a medium-sized variance.

In Oaksford and Chater's (2003a) model, a and b must satisfy the logical constraint $1 - a\varepsilon \geq b \geq a(1 - \varepsilon)$, and thus b can in principle range from $a(1 - \varepsilon)$ to $1 - a\varepsilon$. This means that the admissible values of b can be obtained from the equation $b = c(1 - a) + a(1 - \varepsilon)$, where the new parameter c ranges from zero to one. We therefore reparametrized the model by generating b from c, and we assumed a beta distribution for the new parameter c with parameters α_c and β_c. The resulting model uses four parameters: α_a, β_a, α_c, and β_c. Note that a and b can in principle take on any of the admissible values and that the regions within which a and b move with nonnegligible probability are determined by the parameters. For example, as the variance of c (determined by α_c and β_c) decreases, the size of the band within which b moves around a decreases. The position of this band relative to a is determined by the mean of c.

For each value of the parameters a and c and each card r, we computed the scaled expected information gain associated with card r, $SEI_g(r \mid a, c)$, as in Equation 7 of Oaksford and Chater (2003a), and transformed this to card selection probabilities via the selection tendency function given in their Equation 8. For each selection pattern, the probability of that pattern was then computed

from these cardwise probabilities via the independence assumption, followed by integration over the above-specified distribution for *a* and *c*. The integration introduces dependencies between the cards in a manner analogous to Oaksford and Chater's (1994) simulation of nonindependence. The resulting predicted pattern probabilities specify the extended ODS model. The four parameters, α_a, β_a, α_c, and β_c, of the resulting model were estimated from the selection frequencies of the 16 patterns with the maximum likelihood method, and AIC and BIC were computed from the results as before.

Both in terms of BIC and AIC, the extended ODS model, placed in the set of models shown in Table 2, occupied the highest mean ranks below the independence model itself (see Table 2). In particular, Wilcoxon tests showed that it performed worse than the inference-guessing model in terms of both AIC and BIC (smallest $|Z| = 3.51$, largest $p < .01$). When we considered goodness of fit, the extended ODS model had to be rejected at the 1% level of significance for 17 of the 18 data sets.

We also considered a model with five parameters in which the exception parameter ε was estimated from the data (with the restriction $0 < \varepsilon < 0.5$) rather than fixed at the value .1, but this model performed like the one with fixed ε in terms of model selection and model fit. The same was true of the original ODS model (i.e., the model as stated by Oaksford and Chater, 2003a, with independence assumption) with variable parameter ε.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix B
Demographic Information

Index	Experiment 1	2	3	4	5	6
			Demographic			
German version (%)	10	19	17	11	19	21
Familiar with formal logic (%)	52	50	51	46	51	47
Proportion male (%)	44	45	41	43	48	48
Age						
M	28.8	26.7	26.4	27.4	27	27.8
SD	13.2	12.7	12.2	11.9	12.5	13.6
Years at school[a]						
M	13.5	13.2	13.3	13.1	13	13.3
SD	5.7	7.3	5.5	5.5	5.4	6.1
			Occupation (%)			
College/university student	21	25	24	26	30	31
School student	18	22	22	21	19	17
White-collar worker	24	17	18	20	18	18
Blue-collar worker	6	5	6	5	6	4
Civil servant	2	3	4	2	3	3
Freelance	2	2	3	3	2	3
Self-employed	7	5	5	7	6	5
Other	20	20	19	17	17	19
			Field of academic studies (%)			
No college/university studies	20	21	23	20	21	18
Pedagogics	2	2	1	1	3	3
Humanities and arts	10	8	8	10	8	7
History	1	1	2	1	2	2
Computer science	6	8	9	9	9	8
Law	4	3	4	3	3	3
Mathematics	4	3	4	4	4	4
Medicine	6	6	5	5	4	5
(Natural) science	5	6	5	4	5	8
Philosophy	1	1	2	2	2	1
Social science	5	7	4	7	7	8
Sports	0	1	1	1	1	1
Language	2	3	2	2	2	1
Theology	1	1	0	1	1	1
Economy and management	9	9	8	7	9	10
Engineering	7	6	8	8	7	7
Other	16	15	15	15	14	14

[a] Includes years at college or university.

Received October 4, 2006
Revision received February 16, 2007
Accepted February 16, 2007 ■