American Psychological Association

Psychological Bulletin

0033-2909/18/$12.00

2018, Vol. 144, No. 8, 779–796
http://dx.doi.org/10.1037/bul0000146

On Selecting Evidence to Test Hypotheses: A Theory of Selection Tasks

Marco Ragni and Ilir Kola
University of Freiburg

Philip N. Johnson-Laird
Princeton University and New York University

How individuals choose evidence to test hypotheses is a long-standing puzzle. According to an algorithmic theory that we present, it is based on dual processes: individuals' intuitions depending on mental models of the hypothesis yield selections of evidence matching instances of the hypothesis, but their deliberations yield selections of potential counterexamples to the hypothesis. The results of 228 experiments using Wason's selection task corroborated the theory's predictions. Participants made dependent choices of items of evidence: the selections in 99 experiments were significantly more redundant (using Shannon's measure) than those of 10,000 simulations of each experiment based on independent selections. Participants tended to select evidence corresponding to instances of hypotheses, or to its counterexamples, or to both. Given certain contents, instructions, or framings of the task, they were more likely to select potential counterexamples to the hypothesis. When participants received feedback about their selections in the "repeated" selection task, they switched from selections of instances of the hypothesis to selection of potential counterexamples. These results eliminated most of the 15 alternative theories of selecting evidence. In a meta-analysis, the model theory yielded a better fit of the results of 228 experiments than the one remaining theory based on reasoning rather than meaning. We discuss the implications of the model theory for hypothesis testing and for a well-known paradox of confirmation.

Public Significance Statement
Our research shows that individuals select evidence to test hypotheses that almost always seeks an instance of the hypothesis to corroborate it and that less often seeks potential counterexamples to the hypothesis to refute it. The data indicate that individuals do not reason independently about the evidence; a result that helped to eliminate most of the 16 existing cognitive theories.

Keywords: hypothesis testing, mental models, paradox of confirmation, reasoning, selection task

Supplemental materials: http://dx.doi.org/10.1037/bul0000146.supp

> And the end of all our exploring
>
> Will be to arrive where we started
>
> And know the place for the first time.
>
> —T. S. Eliot, *Little Gidding*

The way to test a hypothesis depends on its nature and on the available evidence. The evidence may derive from testimony or from observation—casual in daily life, systematic in science. A general hypothesis ranges over a set of entities, e.g., If a person has the disease, then the person has the M1 virus. One sort of test examines people with the disease: if any of them do not have the virus, then the hypothesis is false. Another sort of test examines those who do *not* have the virus: if any of them has the disease, then the hypothesis is false. In practice, of course,

This article was published Online First May 21, 2018.

Marco Ragni and Ilir Kola, Cognitive Computation Lab, Technical Faculty, University of Freiburg; Philip N. Johnson-Laird, Department of Psychology, Princeton University, and Department of Psychology, New York University.

For more information on the model, visit https://mm-wason-selection-task.herokuapp.com/.

This research was supported by a DFG-Heisenberg fellowship DFG RA 1934 3-1 and RA 1934 4/1 to Marco Ragni. We thank the following individuals for advice and help: Linden Ball, Monica Bucciarelli, Ruth Byrne, Nick Chater, Jonathan Evans, Geoff Goodwin, Keith Holyoak, Sangeet Khemlani, Christoph Klauer, Markus Knauff, Paolo and Maria Legrenzi, Robert Mackiewicz, Juan Madruga, Ken Manktelow, Mike Oaksford, Klaus Oberauer, Cristina Quelhas, Lance Rips, Carlos Santamaría, Walter Schaeken, Walter Schroyens, Dan Sperber, Christoph Stahl, Valerie Thompson, and the late and much missed Vittorio Girotto. A special thank goes to Nicolas Riesterer and Moritz Rocholl who implemented and tested the Python programs. We also thank the editor, Dolores Albarracin, for a very helpful suggestion about the reframing of a draft of this article, Bettina von Helversen, and two anonymous reviewers for their insightful critiques.

Correspondence concerning this article should be addressed to Marco Ragni, Cognitive Computation Lab, Technical Faculty, University of Freiburg, 79110 Freiburg, Germany. E-mail: ragni@cs.uni-freiburg.de

the selection of evidence is a strategic and sometimes a statistical matter. People often suppose that instances of a hypothesis, such as people with both the disease and the virus, are crucial, but no matter how many instances one encounters, there remains the possibility of a counterexample, someone with the disease who does not have the virus. One counterexample suffices to falsify a hypothesis unless it is probabilistic. And the possibility of its falsification is one criterion that a hypothesis is scientific (Popper, 1959). Three psychological questions therefore stand in need of answers:

1. Do logically naive individuals grasp the need to check for potential counterexamples to a hypothesis?

2. What are the mental processes underlying their selections of evidence?

3. What factors in the situation modify their performance, so that, for example, they start to search for counterexamples?

A massive psychological literature exists on these topics, and its major stimulus was the late Peter Wason's invention of various tasks designed to discover whether naive individuals grasped the importance of counterexamples. In his "*2 4 6*" experiment, participants were given these three numbers as an instance of a general numerical rule. Their task was to discover the rule. They generated further triples, and the experimenter told them whether or not each triple was an instance of the rule. At any point, they could announce their hypothesis about the rule, and the experimenter told them whether or not it was correct (Wason, 1960). The participants tended to test instances of their hypotheses rather than counterexamples, and they often failed to discover the correct rule, which was *any three ascending numbers*. But, the test itself has some problems arising from the participants' need to generate their own hypotheses (Klayman & Ha, 1987; Poletiek, 1996) and the deceptive nature of the actual rule (Baron, 2008).

Another of Wason's tasks was seminal: his *selection* task. It has provoked more than 200 journal articles (Evans, 2017). Readers might therefore suppose that psychologists have definitive answers to the three questions above. Far from it. They have proposed at least 16 different theories about the testing of hypotheses. Such a number shows that no real understanding exists of how individuals select evidence. The present article therefore aims to eliminate as many of these theories as possible, and to make progress toward a definitive theory.

The article begins with an outline of the two main tasks for studying how people choose evidence: the selection task, and a task in which individuals receive feedback from their selections of evidence, the "repeated" selection task (Johnson-Laird & Wason, 1970a). It then describes a new theory of how individuals select evidence, which is an integration of mental models with an early dual-process theory (Johnson-Laird & Wason, 1970b). This "model" theory makes three principal predictions, and the article surveys the experimental findings corroborating them. These results include a novel analysis using Shannon's (1948) measure of information to show that individuals choose items of evidence in a dependent way, that is, one choice relates to another. The article postpones the discussion of alternative theories until after it has marshaled the main evidence about the selection task, because this evidence eliminates many of the alternative theories. The evidence leaves only two viable theories: the model theory and a theory based on inferences from hypotheses or on guesswork (Klauer, Stahl, & Erdfelder, 2007). The model theory relies on the *meaning* of a hypothesis, which at the very least distinguishes between those cases in which the hypothesis is true and those in which it is false. The inference-guessing theory relies on *inferences* from the hypothesis, which at the very least are conclusions that follow from it, and expressible in other assertions. The article reports an analysis of the goodness of fit of these two theories to the results of all 228 relevant experiments that we could find in the literature. Finally, it discusses the implications of these results for testing hypotheses.

The Selection Task

The selection task enables investigators to determine which potential evidence individuals think is relevant to finding out whether a hypothesis is true or false. In its original version (Wason, 1966, 1968), the experimenter explains to the participants that each card in a pack has a letter on one side and a number on the other side. Four cards chosen at random from the pack are placed on the table; for example,

E K 2 3

The experimenter then presents a hypothesis, such as "If there is a vowel on one side of a card, then there is an even number on the other side." The participant's task is to select all those cards, and only those cards, which would have to be turned over in order to discover whether the hypothesis is true or false about the four cards on the table. The first four experiments with the selection task used a conditional hypothesis of the sort, *if p then q*, such as the one above, and they yielded four *canonical* selections of cards, and a remainder of miscellaneous selections, each occurring less often than chance (Wason & Johnson-Laird, 1972, p. 182). Given a hypothesis, *if p then q*, we symbolize the four cards as follows, using a bar over a letter to represent that it is a negative instance of a clause in the hypothesis,

p (the E card for the hypothesis above, because it is a vowel)
$\bar{p}$ (the K card, because it is not a vowel)
q (the 2 card, because it is an even number)
$\bar{q}$ (the 3 card, because it is not an even number).

The canonical selections occurred in the following percentages in the four studies (n = 128):

pq: 46%
p: 33%
p$q\bar{q}$: 7%
p$\bar{q}$: 4%

other miscellaneous selections: 10%. As Nickerson (2015, p. 33) remarked, the majority of selections are either pq or p alone. They are instances of the hypothesis, whereas a "counterexample" selection is the pair p$\bar{q}$, and q is irrelevant, because no matter what is on its other side, it cannot falsify the hypothesis.

Piaget and his colleagues argued that to check the truth of a conditional, *if p then q*, individuals look for a counterexample, p and $\bar{q}$ (e.g., Beth & Piaget, 1966, p. 181). Yet, adults failed to choose $\bar{q}$ in the selection task. Skeptics therefore thought that the selection task was a trick (Cohen, 1981), over complicated

(Finocchiaro, 1980), or elicited a sensible neglect of $\bar{q}$ (Wetherick, 1970, 1995). Such views defend human rationality. Their proponents, rightly impressed by the invention of logic and mathematics, hold that humans are rational. Yet, the view that it is impossible to think illogically, as the logician Ramsey (1931/1990, p. 7) remarked long ago, is like arguing that it is impossible to break the rules of bridge, because, if you do, you are no longer playing bridge. As we show, people's errors in the selection task are systematic and predictable. Moreover, given feedback about the consequences of their selections, they soon switch to potential counterexamples—a point that the skeptics overlooked.

The original selection task used an abstract hypothesis about letters and numbers, but with a hypothesis that was an everyday generalization, "Every time I go to Manchester, I travel by train." Most participants tended to select the two cards corresponding to the potential counterexample (i.e., going to Manchester, traveling there by car and not train; Wason & Shapiro, 1971). Other studies examined deontic principles, which govern what is obligatory and what is permissible. The participants had to select, not evidence that tests a hypothesis, but evidence of a potential contravention of a principle. The first deontic study had a striking effect too (Johnson-Laird, Legrenzi, & Legrenzi, 1972). It used a postal regulation akin to one in force in the United Kingdom:

If a letter is sealed, then it has a 50 lire stamp on it.

Lire were the then Italian currency, which was a fact well known to the English participants. Nearly all of them selected counterexamples to the postal regulation, but hardly any of them selected counterexamples to an abstract hypothesis, and there was no transfer from one task to the other.

Some subsequent studies replicated the effect of an everyday generalization on counterexample selections (Bracewell & Hidi, 1974; Gilhooly & Falconer, 1974; Pollard, 1981; van Duyne, 1974). Others did not (Brown, Keats, Keats, & Seggie, 1980; Griggs & Cox, 1982; Stanovich & West, 1998a; Tweney & Yachanin, 1985; Yachanin & Tweney, 1982). For example, the hypothesis: "If I eat haddock, then I drink gin" is hardly abstract, but led to no reliable increase in counterexample selections (Manktelow & Evans, 1979). The same study also found no improvement with the generalization about traveling to Manchester by train. Other studies showed that a critical factor in the deontic rule about postage was familiarity with a British postal regulation that charged less postage for unsealed envelopes (e.g., Cheng & Holyoak, 1985; Golding, 1981). Another familiar deontic principle was effective (Griggs & Cox, 1982): "If a person is drinking beer, then the person must be over 18." This tended to elicit the selection of potential contraveners: a beer drinker of less than 18 (for similar results, see Cheng & Holyoak, 1985; Kroger, Cheng, & Holyoak, 1993).

We examine presently all the studies that we could find using abstract hypotheses, everyday hypotheses, or deontic principles. The implications of the early studies seemed clear, however. When naive individuals test a hypothesis, they tend to focus on those entities that the hypothesis refers to, and to select these instances of the hypothesis. They select evidence corresponding to a counterexample to the hypothesis only if they are exceptional individuals, or the contents are helpful, or they carry out the task we describe next.

The Effects of Feedback in the Repeated Selection Task

What happens when individuals select evidence to test a hypothesis and then get feedback about its consequences? The repeated selection task answers this question. In a typical study, participants tested a hypothesis, such as, "All the triangles are white," which is equivalent to the conditional, "If a shape is a triangle, then it is white" (Johnson-Laird & Wason, 1970a). There were two boxes of shapes, and, as the participants knew, one box contained 15 white shapes and one box contained 15 black shapes. On each trial they could ask either for a white shape (q) or for a black shape ($\bar{q}$), and they then saw whether it was a triangle (p) or not ($\bar{p}$). Most participants started with potential instances of the hypothesis: they requested a white shape. It was always an actual instance of the hypothesis: a triangle. They chose more white shapes. But, whether a white shape was a triangle or not, it was consistent with the hypothesis, and even if all the white shapes were triangles, the participants still wouldn't know whether the hypothesis was true. In contrast, if a black shape were a triangle, then it would refute the hypothesis. So, sooner or later, the participants started to choose nothing but black shapes, and kept doing so until they had exhausted the box. At which point, they knew whether or not the hypothesis was true. Some participants had a partial insight: they vacillated between the white and black shapes (see also Oakhill & Johnson-Laird, 1985). But, with a simple hypothesis, all the participants at some point switched to choosing only potential counterexamples to the hypotheses—a result often overlooked by critics attacking the selection task.

The repeated task is comparable to a standard selection task in which the choice is between only two items of evidence, q or $\bar{q}$. Such a "reduced array" improves the rate of counterexample selections in comparison to an array of all four items (Lunzer, Harrison, & Davey, 1972; Roth, 1979), perhaps because it reduces the load on working memory (Baron, 2008). The improvement does not transfer to the standard version with four cards (Wason & Green, 1984), and even some children can cope with the reduced array (Girotto, Gilly, Blaye, & Light, 1989; Girotto, Light, & Colbourn, 1988).

The Model Theory

The original insight theory. The first theory of the selection of evidence postulated that individuals varied in their insight into the importance of counterexamples to the hypothesis (Johnson-Laird & Wason, 1970b). It proposed an algorithm, which was in the form of a flowchart, but not programmed, because computers were not available to the authors in those days. Its basic assumption is that individuals use a representation of the meaning of the hypothesis to guide the selection of evidence. They may operate intuitively with no insight into counterexamples, or they may switch to such an insight, or in rare cases they may have this insight from the outset. Hence, the theory was a "dual process" one *avant la lettre* (cf. Wason & Evans, 1975; Wason & Johnson-Laird, 1970), and it was the first theory of hypothesis-testing to be described in an algorithm. Oddly, it seems to be still the only algorithmic theory of hypothesis-testing, because alternative theories describe what is computed rather than how it is computed. We have recently implemented the theory in a computer program,

with one crucial modification, and we now outline the theory and program.

The new model theory. The theory of mental models—*the model theory*, for short—was developed to explain reasoning in general (e.g., Johnson-Laird, 1983; Johnson-Laird, Khemlani, & Goodwin, 2015). It postulates that a conditional hypothesis, *if p then q*, refers to what is possible and what is impossible (Johnson-Laird et al., 2015). The conjunction, *p and q*, is possible for a true conditional, but impossible for a false conditional, whereas the conjunction, *p and not-q*, is impossible for a true conditional but possible for a false conditional. Cases of *not-p*, however, are possible whether or not the conditional is true. Experiments have shown that people make such judgments (Quelhas, Rasga, & Johnson-Laird, 2017). The model theory accordingly postulates that the conditional: "If there is an E on one side of a card, then there is a 2 on the other side" refers to a conjunction of possibilities. They are represented in two mental models, as depicted in the following diagram:

E 2

. . .

The first model represents the possibility of an E and 2, and the second model denoted by the ellipsis is a place-holder for possibilities in which there is not an E. Mental models underlie the intuitive system (system 1) of the program. However, they can be fleshed out into fully explicit models, which yield a conjunction of all the possibilities:

E 2

$\neg$E 2

$\neg$E $\neg$2

and, most important, of the counterexample to the hypothesis, which is impossible:

E $\neg$2

Fully explicit models underlie the deliberations (system 2) of the program, and because the two models containing $\neg$E are possible whether the hypothesis is true or false, the truth of the hypothesis implies just one possibility and one impossibility. A single possibility implies a factual claim: E and 2 co-occur, and a single impossibility implies another factual claim: E and $\neg$2 do not co-occur.

This semantics solves a paradox of confirmation that has long puzzled philosophers and psychologists (see, e.g., Hempel, 1945; Nickerson, 1996). The semantics of the original insight theory was from logic (see Jeffrey, 1981, chapter 1). And, in standard logic, a hypothesis, such as, "If it is a roc, then it is white" is equivalent to "If it not white, then it is not a roc." So, the observation of, for instance, a blue dahlia corroborates the hypothesis about rocs. It gets worse. The hypothesis is true in case the proposition "it is a roc" is false. In fact, the proposition is false because rocs are mythological birds, and so the hypothesis is true. As logicians say, it is "vacuously" true because rocs do not exist. Hence, in standard logic, a general hypothesis, *if p then q*, can be confirmed without establishing that instances of p exist. Likewise, in the repeated selection task with the hypothesis, "All triangles are white," one need only examine shapes that are not white, and the hypothesis could be true because there are no triangles. In contrast, the semantics of mental models ensures that the truth of the hypothesis depends on showing both that white triangles occur and that nonwhite triangles do not occur. It is therefore sensible to select some instances of white shapes in order to establish that white triangles exist, and to select all the instances of black shapes to establish that none of them are triangles. And the claim about rocs is not true merely because rocs do not exist. Conditional hypotheses mean something quite different from their analogs in logic, and this meaning does not yield the paradox of confirmation.

The algorithm for the selection task is the original one (Johnson-Laird & Wason, 1970b) but it now uses models of the hypothesis in place of a semantics from logic. Its first step is to make a list of those items of evidence to which the hypothesis refers—an anticipation of "matching" bias (see Evans, 1972, 1998). Given a conditional hypothesis, *if p then q*, the list is either p or pq. That is, with no insight into counterexamples, system 1 in the program selects as potential evidence any item on its list based on mental models. With a partial insight, system 2 of the program constructs the fully explicit models of the hypothesis, and adds any additional item that could verify the hypothesis, or, failing that, any that could falsify the hypothesis. So, if q is not on the opening list, it is selected now because it can verify the hypothesis. But, if q is already on the opening list, the program adds $\neg$q because it can refute the hypothesis, yielding the selection pq$\neg$q. With complete insight from the outset, the program's system 2 constructs fully explicit models of the hypothesis and selects only items that are potential counterexamples to the hypothesis: p$\neg$q. The theory is not deterministic, because its level of insight is probabilistic (pace Evans, 1977). The program uses probabilistic parameters to yield the level of insight, and its source code in the programming language Python is in the online supplemental materials. Readers can interact with a demonstration of the program at: https://mm-wason-selection-task.herokuapp.com/.

The algorithm's focus on the items in models is borne out in the finding that individuals spend more time inspecting those items that they go on to select than those they do not select (e.g., Ball, Lucas, Miles, & Gale, 2003; Ball, Lucas, & Phillips, 2005; Evans, 1995, 1996; Evans & Ball, 2010; Lucas & Ball, 2005; cf. Roberts & Newton, 2001). We report later on the theory's predictions and its fit to the results of experiments.

Other versions of the model theory. Schroyens, Schaeken, and d'Ydewalle (2001) formalized their own version of the model theory, which Schroyens and Schaeken (2003) showed gave a better account of inferences from conditional premises than a probabilistic model (Oaksford, Chater, & Larkin, 2000); and Oberauer (2006) corroborated this analysis and examined two further versions of the model theory. Koralus has likewise developed another version of the model theory (Koralus & Mascarenhas, 2013). But, these versions of the theory are not formulated for the selection task, and so we consider them no further.

The effects of negation on the selection task. The model theory explains a striking discovery due to Evans and his colleagues (e.g., Evans & Lynch, 1973). Given a conditional hypothesis with a negated *then*-clause, *if p then not q*, participants made a counterexample selection, pq, much more often than they made a counterexample selection, p$\neg$q, for an affirmative conditional. But, given a conditional with a negated *if*-clause, *if not-p then q*, or *if not-p then not-q*, participants showed little consensus about which

items of evidence to select. They seemed almost to be guessing. Evans described these effects in terms of heuristics, including the "matching" heuristic, that is, the tendency to select evidence that matches the item referred to in the *then*-clause of a hypothesis, whether or not it is negated (see, e.g., Beattie & Baron, 1988; Platt & Griggs, 1995; Reich & Ruth, 1982). As Evans (1983) predicted, this tendency to match ignoring negations was reduced if cards used explicit negations, such as, "not an E", instead of "K" (Evans, Clibbens, & Rood, 1996). A large study replicated the phenomena, but on a smaller scale (Stahl, Klauer, & Erdfelder, 2008).

The model theory explains the effects of negation using Wason's (1965) hypothesis that negation suggests the possibility of the corresponding affirmative proposition—a view anticipated by philosophers of language (e.g., Strawson, 1952, p. 8). Hence, the mental models of a conditional with a negated *then*-clause—If p then not-q—are as follows:

$$\begin{array}{ll} p & \bar{q} \\ & q \end{array}$$

Individuals can select a card only if their models of the conditional represent it. Hence, the preceding conditional should be more likely to elicit a counterexample selection of pq. The mental models of a conditional with a negated *if*-clause—If not-p, then q—are as follows:

$$\begin{array}{ll} p & \\ \bar{p} & q \end{array}$$

They do not represent $\bar{q}$, but are equivalent to a disjunction, *p or q*, (see Evans, 1993) and disjunctions are a well-known source of a greater variety of responses (Johnson-Laird & Tagart, 1969; Wason & Johnson-Laird, 1969). A crucial result corroborates the theory's account. Goodwin and Wason (1972) detected different levels of insight in participants' remarks as they explained their selections of evidence. For an affirmative conditional, *if p then q*, they referred to selecting q because p on its other side would verify the hypothesis. But, for a negative hypothesis, *if p then not q*, they referred to selecting q because p on its other side would *falsify* the hypothesis (see also Evans, 1995).

Deontic interpretations of conditionals. The model theory of deontics postulates that models can represent permissible situations (Bucciarelli & Johnson-Laird, 2005), but it is incompatible with modal logics, of which there are many, because individuals make inferences that are invalid in modal logics, and reject inferences equally robustly that are valid in modal logics (Ragni & Johnson-Laird, 2017).

Knowledge can modulate the interpretation of conditionals and other assertions (e.g., Quelhas, Johnson-Laird, & Juhos, 2010). It leads people to interpret a conditional such as "If the animal is a lion, then the female is a lioness" as equivalent to a biconditional, which is equivalent to "If, *and only if*, the animal is a lion, then the female is a lioness." They do so because they know that for no other animal is the female a lioness. Knowledge also modulates deontic assertions. For example, the deontic conditional "If you tidy your room, then you may go out to play" is likely to be treated as a biconditional, "If, and only if, you tidy your room, then you may go out to play." It therefore has two counterexamples:

$$\begin{array}{rr} \text{tidy room} & \text{do not go out to play}: \ p\bar{q} \\ \text{do not tidy room} & \text{go out to play}: \ \bar{p}q \end{array}$$

In general, participants should select all four instances: $p\bar{q}\bar{p}q$, as potential counterexamples to the conditional. But, the speaker (presumably a parent) and the listener (presumably a child) are two protagonists each concerned with just one of the two counterexamples. So, those taking the point of view of the child should select only the first pair: the child tidied the room, yet she didn't go out to play, whereas those taking the point of view of the parent should select only the second pair: the child didn't tidy the room, yet she went out to play (Light, Girotto, & Legrenzi, 1990). Generalizations that are not deontic can also be interpreted as biconditionals (e.g., Quelhas et al., 2010), and then point of view has similar effects too (Almor & Sloman, 2000; Fairley, Manktelow, & Over, 1999; Staller, Sloman, & Ben-Zeev, 2000).

Two observations informed the theory's account of tests of deontic principles. First, children learn what they should *not* do earlier than what they should do (Gralinski & Kopp, 1993). Second, adults are more sensitive to counterexamples to deontic principles than to instances of them (Bucciarelli & Johnson-Laird, 2005). In tests of moral principles, individuals should therefore be more likely to select potential counterexamples than instances.

The model theory's predictions. The model theory makes three principal predictions about the selections that individuals should make to test hypotheses or deontic principles:

> Prediction 1: The only reliable selections of potential evidence for tests of conditional hypotheses should be the four canonical ones. Other selections should be haphazard, and therefore occur no more often than chance.
>
> Prediction 2: Choices of items of evidence to test conditional hypotheses should be dependent: intuitions are based on mental models and should yield selections that include q only if they include p; and deliberations are based on fully explicit models and should yield selections that include $\bar{q}$ only if they include p.
>
> Prediction 3: Any manipulation that makes a counterexample to a hypothesis more salient should increase selections of potential falsifications (Johnson-Laird & Byrne, 1991, p. 80; Johnson-Laird & Byrne, 2002). Hence, deontic principles should be more likely to yield counterexample selections than other generalizations do, and everyday hypotheses should be more likely to yield them than abstract ones. Counterexamples can be made salient for abstract hypotheses in the instructions and in the framing of the task.

The Empirical Evaluation of the Model Theory

To assess the model theory's three predictions, we assembled all the papers we could find on the testing of conditional hypotheses. We excluded those studies that did not report the frequencies of the four canonical selections p, pq, $p\bar{q}$, and $pq\bar{q}$. We classified all the resulting experiments according to their three sorts of generalization: abstract, everyday, and deontic. There were 228 experiments that satisfied our criteria, and their individual results are available in the online supplemental materials.

Our first task was to check whether the results of the 228 studies were homogenous enough to be worth analyzing. We assessed their homogeneity for each of the three sorts of generalization. Experiments with abstract hypotheses were of three sorts requiring

Table 1
The Concordances (Kendall's W) in the Frequencies of the Four Canonical Selections of Evidence Over 228 Different Experiments Examining the Three Main Sorts of Conditional Generalization: Abstract Hypotheses, Everyday Hypotheses, and Deontic Principles, and for the Three Sorts of Instruction for the Abstract Hypotheses

Three sorts of generalization	Instructions: the evidence should determine that the hypothesis:	Number of experiments	Number and percentage of participants making a canonical selection	Kendall's W over the 4 canonical selections, its $\chi^2(df = 3)$, and probability		
				W	χ^2	*p*
Abstract	Is true or false	55	3,497 (68%)	.52	86	<.001
	Is false	29	1,503 (73%)	.36	32	<.001
	Holds	20	307 (71%)	.31	18	<.001
	Overall	104	5,307 (69%)	.34	107	<.001
Everyday	Overall	44	2,451 (67%)	.25	33	<.001
Deontic	Overall	80	2,547 (77%)	.54	129	<.001

the participants to test whether the hypothesis was true or false, whether it was false, or whether it holds, and we tested the homogeneity of results for these different instructions too. Table 1 summarizes the results. The rank orders of the frequencies of the four canonical selections of evidence showed a robust concordance for each sort of generalization (as Kendall's W shows because it ranges from 0 for no consensus to 1 for perfect consensus). We therefore assessed the model theory's predictions.

Prediction 1: Individuals tend to make a canonical selection of evidence. The model theory predicts that individuals should make one of the four canonical selections. Some participants guess, make haphazard responses, or fail to tackle the task properly, and so the prediction excludes such responses. But, what selections were guesses, errors, or failures? One touchstone is that each instance of such responses should be idiosyncratic and rare, that is, it should occur no more often than a selection made at random. There are 16 possible selections, and therefore a selection that occurs less than 6% in the overall results occurs less often than a random one. Within our database, 99 experiments reported the percentages of all 16 selections, which include the choice of no items at all. Table 2 presents each of the selections that occurred 6% or more for at least one of the three sorts of generalization in these 99 experiments. They include the four canonical selections: pq, p, p$\bar{q}$ and pq$\bar{q}$. They also include a selection that reflects either a biconditional interpretation (e.g., Manktelow & Over, 1991) or a precautionary choice of all four cards in place of a thoughtful choice: p$\bar{p}$q$\bar{q}$, and a selection made in online studies with special instructions to consider each card carefully (Klauer et al., 2007): q. Overall, the test-bed corroborated the prediction that individuals tend to make the canonical selections of evidence.

Prediction 2: The dependence of selections of evidence. Individuals could select any of the four items, p, $\bar{p}$, q, and $\bar{q}$, independently of the others. But, the model theory predicts that selections are dependent, for example, the selection of q is dependent on the selection of p. The issue matters, because independent selections call for an analysis only of the frequencies of choice for each of the four individual items of evidence, and some studies report only these data, whereas dependent selections call for an analysis of the frequencies of the different selections as a whole. So, are selections independent or dependent?

As Evans (1977, p. 635) remarked, the issue is an empirical one. He tested the correlation between the presence of q and of $\bar{q}$ in the selections. A reliable correlation would show, he argued, that their choices of items of evidence were dependent on one another. Conversely, the lack of a reliable correlation would support the null hypothesis of independent selections. Because the expected values in the cells of the 2 × 2 table were for some studies less than 5, Evans assessed the correlation with the Fisher–Yates exact test, though it calls for unrelated data (Siegel, 1956, p. 96). In multiple conditions of six studies, no significant correlation occurred, and so Evans accepted the null hypothesis of independence. He likewise proposed a stochastic theory in which each item of evidence is chosen independently from the others.

Pollard (1985) readdressed the issue using the same method, but he examined all six possible correlations between the four items: p with $\bar{p}$, p with q, p with $\bar{q}$, and so on, in studies of all four possible

Table 2
The Percentages of the Selections of Potential Evidence to Test the Hypothesis, If p Then q, That Occurred for at Least One of the Three Sorts of Generalization (Abstract, Everyday, and Deontic) Among the 99 Experiments Reporting all 16 Possible Selections

Three sorts of selection task	Number of experiments	pq	p	p$\bar{q}$	p$\bar{p}$q$\bar{q}$	q	pq$\bar{q}$	The remaining 10 selections each made < 4%
Abstract	43	29	29	7	4	9	2	19
Everyday	30	26	15	21	9	2	7	20
Deontic	26	9	5	61	4	1	1	18
Overall	99	26	22	17	6	6	4	19

hypotheses based on conditionals with affirmative or negative clauses (Evans & Lynch, 1973; Manktelow & Evans, 1979). His results led him to reject independence in the selections of evidence. His findings have been replicated in other studies (Oaksford & Chater, 1994; Klauer et al., 2007).

A better assessment than pairwise correlations would provide a single measure taking into account each selection as a whole for all the selections in an experiment. We devised a procedure to make such an assessment. It combines Shannon's (1948) measure of informativeness (or unpredictability) and the computer simulation of thousands of experiments. Its rationale is straightforward. Suppose that the selections in an experiment are more redundant—less informative—than selections based only on the individual probabilities with which each of the four items of evidence were selected in the experiment. It follows that something is constraining the selections over and above these four independent probabilities. Hence, the selections are dependent.

The first step in our procedure is to compute the amount of information in the selections in an actual experiment. Shannon's measure, which is symbolized as H, has the following equation in units that are bits:

$$H = -\sum P_i \log_2 P_i$$

where P_i is the probability of the *ith* selection of evidence in an experiment, which depends on its frequency of occurrence in the experiment, and $\log_2$ is a logarithm to the base 2. In general, the greater the number of different selections, and the more evenly distributed their frequencies, so the value of H increases, and it is harder to predict the selections. If participants chose each item of evidence independently of the others in making a selection, the value of H for an experiment would not differ reliably from its value for selections based on sampling each item of evidence according to its probability in the experiment. But, if the value of H for the selections in the experiment is reliably smaller than this theoretical value, then we can reject the null hypothesis of independent selections. In other words, the redundancy in an experiment's smaller value of H reflects the dependence of its selections.

We analyzed the redundancy of each of the experiments in the database of 99 experiments reporting all 16 possible selections of evidence, because pooling different low frequency selections into a single "miscellaneous" category underestimates the value of H. If each of the 16 selections has the same frequency of occurrence, then H has a maximum value of 4 bits. The mean over the 99 experiments (see Table 2) was 2.14 bits. We implemented a computer program to compare the redundancy of each experiments' selections with that of simulations of the experiment based only on the probabilities of choices of the four items, and its source code in R is in the second section of the supplemental materials. Its main steps are as follows for each experiment in a set:

1. Compute N, the number of participants in the experiment, and the probabilities with which each of the four items of evidence, p, $\bar{p}$, q, and $\bar{q}$, occurred in the experiment's selections.

2. Compute Shannon's informativeness, H, for the selections in the experiment.

3. Carry out 10,000 simulated experiments based on the four probabilities of selecting each item of evidence, assigning a selection based only on these probabilities to each of *N* participants.

4. Return the H value of the actual experiment and the mean H value of the simulated experiments based on the assumption of independence.

For each of the three main sets of experiments, we then tested whether the difference between the pairs of actual values of H and the mean simulated values differed reliably (using a nonparametric test, the Wilcoxon's signed-ranks matched pairs test).

As an illustrative example, we use the selections of evidence reported in Stahl et al.'s (2008) Experiment 2, which we chose because it had 300 participants. Here are the frequencies of their selections, in which 11 participants selected no evidence whatsoever:

pq	p	$p\bar{q}$	q	$p\bar{p}qq$	none	$p\bar{q}$	$\bar{q}$	$p\bar{p}$	$pq\bar{q}$	$q\bar{q}$	$\bar{p}$	$p\bar{p}q$	$p\bar{q}\bar{q}$	$\bar{p}q$
111	92	17	17	12	11	10	8	5	4	4	4	2	2	1

Shannon's informativeness, H, for the experiment is 2.66 bits, and it yields the following probabilities of choosing the four items of evidence: $p = .787$, $\bar{p} = .143$, $q = .510$, $\bar{q} = .190$. The mean informativeness of the 10,000 simulated experiments each based on the same four probabilities and 300 participants was 3.00 bits. All 10 thousand simulations of the experiment had a higher value of H than the original experiment (Binomial test, $p = .5^{10,000}$). Hence, the selections in the experiment are more redundant than the simulated ones: the choices of the four items are dependent on one another.

Table 3 presents the mean informativeness, H, of the 99 experiments reporting all 16 possible selections and the mean values of H of each of their 10,000 simulations, and the results of Wilcoxon's tests and their p-values comparing the simulations to the actual results for the three sets of experiments. The redundancy of the real experiments over the simulated ones shows that individuals choose items of evidence to test a hypothesis in a dependent way. This result corroborates the model theory and eliminates any theory that predicts that selections are independent.

Prediction 3: Any manipulation that makes a counterexample to a hypothesis more salient should increase corresponding selections in tests of the hypothesis. A counterexample to a hypothesis, *if p then q*, is the case of $p\bar{q}$. Hence, when counterexamples are salient, individuals should be more likely to make such a selection. Three methods of manipulating salience are, first, to change the contents of the task, second, to alter its instructions, and, third, to reframe the experiment as a whole. Some manipulations of the contents have enhanced the rate of counterexample selections, but not all have. Likewise, manipulations of the instructions in experiments have sometimes altered performance and sometimes had no reliable effects. Consider, for instance, whether the instruction to try to falsify the hypothesis or to detect violations of it, lead to an increase in counterexample selections. Some studies reported reliable effects (e.g., Chrostowski & Griggs, 1985; Dominowski, 1995; Green, 1995; Griggs, 1995; Griggs & Cox, 1983; Platt & Griggs, 1993). Other studies reported no reliable

Table 3
The Mean Informativeness, H, (in Bits) of 99 Experiments Reporting the 16 Possible Selections of Evidence (for Abstract, Everyday, and Deontic Generalizations), the Mean Values of H of Sets of 10,000 Simulations of Each Experiment, and Wilcoxon Tests and Their p-values of the Differences Between the Pairs of Real and Simulated Experiments in Each Set

The three sorts of generalization	Mean value of H for real experiments	Mean value of H for sets of 10,000 simulations	Wilcoxon's *W* and *p* value for difference in H between the real and simulated experiments
Abstract	2.33	2.60	$W = 89, p < .001$
Everyday	2.43	2.75	$W = 8, p < .001$
Deontic	1.57	1.80	$W = 24, p < .001$

effects (Augustinova, Oberlé, & Stasser, 2005; Griggs, 1984; Valentine, 1985; Yachanin, 1986).

To assess the overall influence of contents and instructions on performance, we analyzed studies reporting the four canonical selections for abstract, everyday, and deontic generalizations. Table 4 summarizes the results. The three sorts of generalization had a reliable effect on the percentages of counterexample selections (Kruskal-Wallis one-way analysis of variance by ranks, $\chi^2 = 54.6$, $df = 2$, $p < .001$). Likewise, the three sorts of instruction for tests of an abstract hypothesis also had a reliable effect on the percentages of counterexample selections (Kruskal-Wallis, $\chi^2 = 62.3$, $df = 2$, $p < .001$). Deontic principles, as the model theory predicts, were the most effective in eliciting counterexample selections, and, for abstract hypotheses, the instructions to test whether the hypothesis was false, or holds, were the most effective in eliciting counterexample selections of evidence.

Various ways of reframing the selection of evidence affect performance (e.g., Griggs & Jackson, 1990; Margolis, 1987). A most revealing study is due to Sperber, Cara, and Girotto (1995). Their prediction was that if a counterexample is both relevant and easy to envisage, participants should be more likely to make a falsifying selection. The term *relevant* here does double duty: it has both its everyday meaning and a technical meaning (Sperber & Wilson, 1995). The investigators carried out four experiments corroborating their hypothesis, and the fourth of them illustrates the results. The four groups of participants ($n = 21$ per group) corroborated the prediction: 57% of those in the group in which the counterexample was relevant and accessible made counterexample selections; only 5% (one participant) did so in the group in which the counterexample was neither relevant nor accessible; and the participants in the other two groups had an intermediate performance: 38% made counterexample selections (see also Girotto, Kemmelmeier, Sperber, & Van der Henst, 2001).

The salience of counterexamples by itself appears to increase the corresponding selections. Love and Kessler (1995) used the conditional hypothesis: "If there are Xow, then there must be a force field," where the participants knew that the Xow are strange crystal-like organisms that depend for their existence on a force field. When the context suggested the possibility of counterexamples, that is, mutant Xows that can survive without a force field, the participants were more likely to make a counterexample selection than in a control condition. Likewise, Liberman and Klar (1996) showed that underlying Cosmides's (1989) idea that people check generalizations to find out whether others are cheating them is the participants' grasp of counterexamples (see also Brown & Moore, 2000; Cheng & Holyoak, 1989; Gigerenzer & Hug, 1992; Politzer & Nguyen-Xuan, 1992).

Overall, salient counterexamples have the predicted effect on testing hypotheses. And when the framing of the experiment—its contents, instructions, background story—make counterexamples relevant and easy to access, participants are more likely to select only potential refutations of the hypothesis. As the example of the Xow shows, this phenomenon goes beyond the mere recall of a familiar counterexample to a hypothesis.

Alternative Theories of the Selection of Evidence

The evidence in the previous section helps to evaluate all the theories of the selection of evidence. At least 16 such theories exist, and they are based on the meanings of hypotheses, on

Table 4
The Percentages of the Four Canonical Selections for the Three Sorts of Generalization in the 228 Experiments and for the Three Different Instructions for the Abstract Contents

The three sorts of generalization	Instructions. The evidence to select should show that the hypothesis:	Number of experiments	Canonical selections			
			pq	p	$pq\bar{q}$	$p\bar{q}$
Abstract	Is true or false	55	45	41	4	**11**
	Is false	29	29	30	7	**34**
	Holds	20	30	13	13	**44**
	Overall	104	39	36	5	**19**
Everyday	Overall	44	39	23	11	**28**
Deontic	Overall	80	19	13	4	**64**

Note. The counterexample selections are shown in bold.

inferences using logical or content-specific rules, on innate modules for reasoning, on heuristics, on probabilities, or on neural models. Table 5 summarizes the theories, and the online supplemental materials describe each of them in detail. The table states each theory's name, its provenance, and its basis for selecting evidence to test generalizations. It also states whether the theory can account for the three principal findings reported in the previous section: the dependence of choices of items of evidence, their membership of the canonical set, and the increase in counterexample selections given pertinent experimental manipulations. We gave theories the benefit of the doubt in cases of uncertainty. Only five theories make at least two of the three predictions. However, the relevance theory (Sperber et al., 1995) was not formalized in a way that allows tests of its fit to experimental data (Dan Sperber, personal communication, January 26, 2017). The multiple-interpretations theory (Stenning & van Lambalgen, 2008) predicts independent selections for deontic principles (cf. Table 3, which reports their robust dependency), and it is also not formulated in a way that allows tests of goodness of fit. Of the three remaining theories, one is the model theory described earlier. We now outline the other two theories.

Optimal data selection. Oaksford and Chater (1994) made the radical proposal that the selections of p, pq, and $p\bar{q}$, are all rational, whether the hypothesis is abstract or an everyday generalization, because for nearly everyone the testing of such hypotheses concerns, not their truth or falsity, but the statistical dependency, which according to these authors, these hypotheses express (see Theory 13 in Table 5). The details of the theory are described in the online supplemental materials. It predicts the correct rank order of the frequencies with which the four items are chosen in selections:

$$p > q > \bar{q} > \bar{p}$$

It also correctly predicts whether the six correlations between choices of items (p with $\bar{p}$, p with q, etc.) are positive or negative. Over the years, the authors clarified their theory (Oaksford & Chater, 1996). But, it could not make quantitative predictions until Hattori (2002) developed it in a new version (see Theory 14 in Table 5). Oaksford and Chater (2003, 2007, p. 172) have endorsed this revision to their theory (see also Oaksford & Wakefield, 2003). But, as Hattori (2002, p. 1262) pointed out, the new theory implies that the choices of items of evidence are independent of one another. So, the later theory of optimal data selection no longer makes binary predictions about positive or negative correlations between pairs of items, but instead it makes independent numerical predictions about the probabilities of choosing each of the four items of evidence.

The original theory provoked various reactions. It elicited criticisms that it is mistaken in its normative assumptions (Evans & Over, 1996), in its Bayesian presuppositions (Laming, 1996), in its account of results (Evans & Over, 1996; Handley, Feeney, & Harper, 2002; Oberauer, Weidenfeld, & Hörnig, 2004; Oberauer, Wilhelm, & Rosas-Diaz, 1999), and in its adequacy as a theory (Almor & Sloman, 1996). Its authors replied to these criticisms (Oaksford & Chater, 1996). Their theory is a brilliant integration of Bayes's theorem and Shannon's informativeness, and it has addressed most of the phenomena of the selection task. Its recent version in which the four cards are selected independently is based, not on a central tenet of the theory, but on one made for a useful index of fit (Mike Oaksford, personal communication, January 20, 2017). One of the advantages of the theory is that it is a special case of a general Bayesian approach to cognition (see, e.g., Oaksford & Hall, 2016). However, its proponents have recently argued that brains need not represent or calculate probabilities at all, and are poorly adapted to do so (Sanborn & Chater, 2016). Estimates instead call for the sampling of information. And the decision of whether or not to choose a potential item of evidence likewise depends on sampling its possible outcomes, and on sampling their informativeness (Nick Chater, personal communication, January 22, 2017).

Table 5
The 16 Theories of the Selection of Evidence: The Theory's Name, Provenance, Basis of Its Predicted Selections, and Whether It Can (+) or Cannot (−) Predict Dependent Selections, Canonical Selections (p, pq, $pq\bar{q}$, $p\bar{q}$), and the Role of Salient Counterexamples in Selections

Name of theory	Provenance	Basis of selections	Dependent selections	Canonical selections	Salient counterexamples
1. Defective truth tables	Wason (1966)	Meaning	−	+	−
2. Insight and models	Johnson-Laird and Wason (1970a)	Meaning	+	+	+
3. PSYCOP	Rips (1994)	Inference	−	−	−
4. Relevance	Sperber, Cara, and Girotto (1995)	Inference & relevance	+	−	+
5. Multiple interpretations	Stenning and van Lambalgen (2008)	Inference & meaning	+	+	−
6. Pragmatic schemas	Cheng and Holyoak (1985)	Inference	+	−	−
7. Innate modules	Cosmides (1989)	Innate module	−	−	+
8. Stochastic theory	Evans (1977)	Heuristics & inference	−	−	−
9. Matching & verifying	Krauth (1982)	Heuristics & meaning	−	−	−
10. Heuristic-analytic	Evans (1984, 1989, 2006)	Heuristics & inference	+	−	−
11. Inference-guessing	Klauer, Stahl, and Erdfelder (2007)	Guessing & inference	+	+	+
12. Probability & utility	Kirby (1994)	Expected utility	−	−	−
13. Optimal information gain, version 1	Oaksford and Chater (1994)	Likely information gain	+	−	+
14. Optimal information gain, version 2	Hattori (2002), Oaksford & Chater (2003, 2007)	Adds logistic selection function	−	−	+
15. Parallel distributed processes	Leighton and Dawson (2001)	Backwards propagation of error	−	−	−
16. Neurons	Eliasmith (2005)	Vectors	−	−	−

The theory has three drawbacks. First, it describes what individuals are supposed to compute in the selection of potential evidence, but provides no algorithm for how they do so. The omission is exacerbated if brains do not calculate probabilities. Second, it presupposes that the experimental instructions and participants' remarks about their selections are irrelevant. The instructions call for an evaluation of the hypothesis as true or false, or variations thereof, and the participants' remarks about the task are likewise about truth and falsity, not probabilities (e.g., Evans, 1995; Goodwin & Wason, 1972; Lucas & Ball, 2005). The contrast between conditionals of the sort used in the selection task and those containing the word "probably", as Goodwin (2014) has shown in multiple experiments, is stark. And highly intelligent individuals do assess the truth or falsity of hypotheses in the selection of evidence: they select potential counterexamples (Stanovich & West, 1998a). Oaksford and Chater (2007, p. 211) counter that only a small percentage of individuals are competent enough to make such selections, perhaps as little as 1% of the population. They may use the meanings of hypotheses, but the rest of us rely on optimal data selection. The mystery then is why everyone converges on potential counterexamples to conditional hypotheses in the "repeated" selection task. Third, the original version of the theory fails to predict one of the canonical selections: pq$\bar{q}$. The more recent version fails to predict dependent selections. Hence, it is not possible to fit either version to the frequencies of canonical selections in our test bed of experiments.

The inference-guessing theory. Klauer et al. (2007) proposed a set of related models, including one that postulates that individuals make one or two inferences, or else use cues to guess or match, in order to select evidence. This model is formulated, not as an algorithm for mental processes, but as a multinomial processing-tree (Riefer & Batchelder, 1988). Figure 1 presents the tree for the inferential component of the model. For example, given the card, p, and a hypothesis, *if p then q*, individuals who infer that q should be on the other side of the card, should select the p card. The inferences depend on the interpretations of the *if–then* hypothesis, which depend on the parameters in Figure 1. The guessing component, which is not shown in in the figure, selects each of the four cards depending on the value of its own parameter. Hence, one parameter governs whether individuals make inferences or guesses, five parameters control the sort of inference that they make, and four parameters control their guesses. As Figure 1 shows, the inferential component yields 11 out of the 16 possible selections. The others can be made from the guessing component of the model. Klauer et al. (2007) fitted the model to their own results. It provided a better fit of selections of single cards or pairs of cards than the stochastic model (see Theory 8 in Table 5), the heuristic-analytic model (see Theory 10 in Table 5), or the second version of optimal data selection (see Theory 14 in Table 5).

The chief problems with the theory are that it provides no account of the mental processes of inference, which, its authors

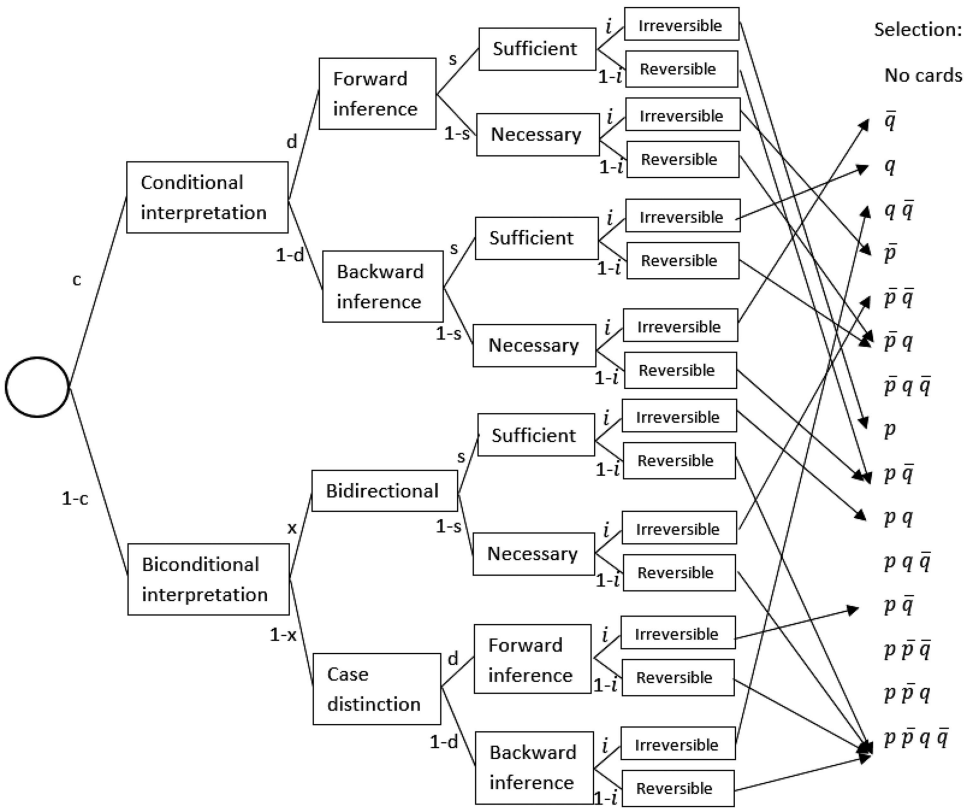

Figure 1. The multinomial processing-tree for the inferential component of the inference-guessing model (from Klauer et al., 2007).

say, could depend on logical rules, mental models, or suppositions, and that its inferential component cannot predict the canonical selection, pq$\bar{q}$ (see Figure 1 in which there is no path to this selection). Yet, this component predicts five selections that occur so rarely—below chance in the 228 experiments—that they should be attributed to the theory's guessing component, that is, the selections: $\bar{p}$, $\bar{q}$, $\bar{p}\bar{q}$, p$\bar{p}$, and q$\bar{q}$ (see Figure 1). The theory offers no explanation for other phenomena, such as the dependence of counterexample selections on their salience to the hypothesis. Each of a theory's free parameters is an abrogation of explanation in favor of description. So, critics might argue that the theory's goal is to fit results rather than to explain them. Nonetheless, the model is impressive, because it yields a better fit than both the heuristic-analytic theory and the optimal data selection theory when it plays the game by their rules.

An immediate snag in fitting the model to the results from the 228 experiments is that standard algorithms cannot fit models with more parameters than categories of results: the inference-guessing model has 10 parameters but the results have only four canonical selections. We therefore had to cut the parameters down to four. We describe the original theory and the function of the boxes in Figure 1 in the Supplemental materials. We describe the simplified theory in the next section.

The Goodness of Fit of the Two Theories

We fitted the model theory (based on Johnson-Laird & Wason, 1970b) and the simplified inference theory (based on Klauer et al., 2007) to the frequencies of the canonical selections in the database of 228 experiments. Figure 2 presents a multinomial processing-tree for the model theory, which mimics the algorithm described earlier.

To simplify the inference-guessing theory, we dropped the four parameters for guessing. Such parameters provide little psychological understanding of how individuals test hypotheses, other than to index the difficulty of the task. So, no profound loss occurs in dropping them. But, it raises a problem: no way then exists for the theory to make the canonical selection, pq$\bar{q}$, which it hitherto could only guess. Because there is now no need for the parameter to choose between inference and guessing, we assigned it instead to the probability of guessing this canonical selection. As we mentioned earlier, its inferential component predicts several selections that occur less often than chance, and so we made sure that our simplification yields only canonical selections and depends only on four parameters. We refer to it as the "simplified inference" model, and Figure 3 summarizes its multinomial processing-tree. It is much simpler than the original theory. Its proponents might well object that we have pruned it too much, but we had to eliminate six of its parameters, because only the canonical selections are relevant.

Fitting the two theories to data. We optimized the values of the two theories' parameters so that they would predict the total numbers of each the four canonical selections (1) for the abstract hypotheses, (2) for the every day hypotheses, (3) for the deontic principles, and (4) for all 228 experiments. The output was therefore a single set of three optimal values for the parameters of the model theory, and a single set of four optimal values for the parameters of the simplified inference theory. To find these optimal values, we used the standard routine for minimization, the L-BFGS-B[1] algorithm from the Python SciPy package for scientific computation. Our script can be downloaded from the online supplemental materials.

To assess the goodness of these fits, we used the single set of optimal values for a theory to calculate the root mean square errors (RMSE) over all the experiments in a set. Table 6 presents these RMSEs, and shows that models fit the overall data well and the data for the abstract, everyday, and deontic sets of experiments. But, the model theory yielded a more accurate fit than the simplified inference theory. To take into account the different complexity of the trees for the two theories, we calculated the Bayesian information criterion (BIC) and the Bayes factor, using the maximum-likelihood method computed with the Python program (and checked with the R-package for multinomial processing-trees, MPTinR of Singmann & Kellen, 2013). The BIC indicates how much information is lost when a tree represents the process that generates the data, taking into account both its goodness of fit and number of parameters. It penalizes theories with a greater number of their parameters, and the smaller the BIC, the better the fit between a theory's tree and the results. The Bayes factor (Schwarz, 1978) also compares the different models. It approximates the difference between the BIC values of the model theory and the simplified inference theory (as computed in MPTinR). The higher the Bayes factor the stronger it supports one theory over another, and a value between 30 and 40 indicates strong evidence for the tree with the better fit (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Table 6 also reports these results. The BIC values and the Bayesian information criteria enhanced the advantage of the model theory, because it has one less parameter than the simplified inference theory.

We searched the space of all possible multinomial processing-tree models for up to three parameters and with no repeating parameter in the search path, and there were 16 other possible trees for optimal fits for the 228 experiments. The model theory was closest to the best model, deviating from it only in one decimal place. It therefore appears that no alternative theory could yield a better fit for these experiments. Yet, as we argue in the final section of the paper, a superior fit to performance in a particular task does not guarantee a superior theory in general.

Doubtless, the 10 parameters of the original inference-guessing model theory would provide a better fit than the model theory for the results of the 99 experiments reporting the data for all 16 possible selections. But, apart from the canonical selections, those that remain are the result of guesses, happenstance, and failures to carry out the task, such as not selecting any evidence. The one possible exception is the selection of all four items of evidence, which could reflect a biconditional interpretation or a prudent choice of all the items of evidence, perhaps to avoid having to think. Hence, such an analysis would be to overfit the data—to fit low frequency noise rather than real selections. As von Neumann

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin_l_bfgs_b.html

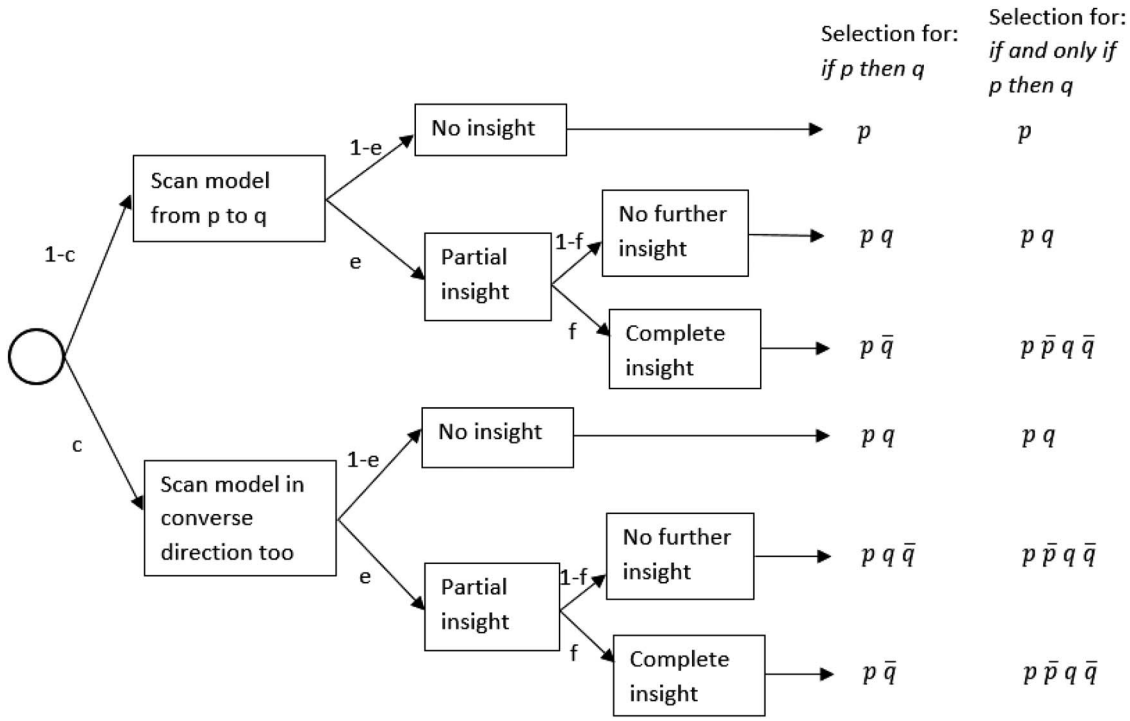

Figure 2. A multinomial process-tree for the model theory (based on Johnson-Laird & Wason's, 1970b, algorithm), which predicts selections of evidence to test conditional and biconditional hypotheses. Its parameters govern the interpretation of hypotheses, and the shift from system 1 using mental models to system 2 using fully explicit models (see the account in the text).

famously remarked, "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" (Dyson, 2004).

General Discussion

Do naive individuals grasp the role of falsification in testing hypotheses? What mental processes underlie their selections of evidence? What factors lead them to select potential counterexamples? We began with these questions, and now we can answer them.

Falsification is not intuitive for most individuals, and so they overlook possibly falsifying evidence, as shown in their failure to select the potential counterexample, $\bar{q}$, in testing a conditional hypothesis, *if p then q*, that is, a case in which *q* does not hold. Yet, if $\bar{q}$

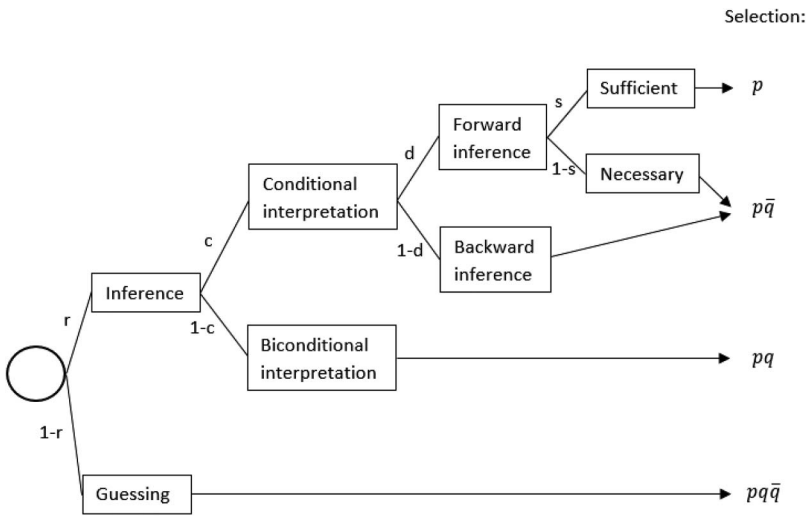

Figure 3. A multinomial process-tree summarizing the simplified inference theory's predictions of evidence selected to test a conditional hypothesis, *if p then q*.

Table 6
The Model Theory's and the Simplified Inference Theory's Goodness of Fit With the Canonical Selections of Evidence for Abstract Hypotheses, Everyday Hypotheses, and Deontic Principles, and for the Overall Results of 228 Experiments: The Root Mean Square Errors (RMSE) for Their Predictions Based on the L-BFGS-B Algorithm, the Bayesian Information Criterion (BIC), and the Bayes Factor for the Better Fitting Model

The numbers of experiments of the three sorts	The two models				Bayes factor of the model theory vs. the simplified inference theory
	The model theory (3 parameters)		The simplified inference theory (4 parameters)		
	RMSE (10^{-6})	BIC	RMSE (10^{-6})	BIC	
Abstract: 104	15.5	50.4	20.0	59.0	73.7
Everyday: 44	6.4	45.2	8.3	53.1	51.9
Deontic: 80	6.1	46.4	12.3	54.2	49.2
Overall: 228	19.2	54.7	32.6	63.9	99.4

co-occurs with p, then the hypothesis is false. This egregious omission in early studies of the selection task shocked philosophers and psychologists, and led both to skepticism about the task and to the large number of its replications. However, given feedback about the consequences of their selections in the repeated selection task, almost everyone realizes sooner or later the crucial need to select $\bar{q}$.

The mental processes underlying the selection of evidence to test conditional hypotheses have provoked at least 16 different theories (see Table 5). Some theories fail because they cannot explain the dependence of selections. Some theories fail because they cannot predict the four canonical selections: pq, p, p$q\bar{q}$, and p$\bar{q}$. Some theories fail because they cannot answer the third question above—the contents of hypotheses, instructions, and the framing of the task, which make counterexamples salient, can all lead to falsifying selections. And some theories fail because they cannot be fitted to experimental results.

Only Two Theories Are Viable

The model theory is based on replacing logic with models in an algorithm due to Johnson-Laird and Wason (1970b). The theory argues that the selection of evidence depends on intuitions based on the meaning of the hypothesis, or, in part or whole, on deliberations using that meaning to yield a counterexample to the hypothesis.

The inference-guessing theory (Klauer et al., 2007) argues that selections depend either on inferring the consequences of the hypothesis or on cues that trigger guessing or other noninferential processes.

The model theory yields a better fit to the 228 experiments in our meta-analysis, and one that is almost optimal. However, we were forced to cut down the 10 parameters of the inference-guessing model in order to fit it to results concerning the four canonical selections. The model theory has three main advantages.

First, it presents an algorithmic account, implemented in a computer program, of how individuals use the meanings of hypotheses to guide their tests of evidence, either intuitively or deliberatively. People appear to rely on intuitions in selecting potential instances of the hypothesis, because they tend not even to look at potential counterexamples in the evidence. Likewise, they normally take several trials in the repeated selection task before they switch to selecting such items.

Second, its predictions are corroborated by the experiments in the literature. They yield consistent results (see Table 1). They bear out its prediction of canonical selections (see Table 2). They yield dependent selections of evidence, for example, individuals choose q only if they choose p, which is corroborated in the greater redundancy of real experiments over that of 10,000 of simulations of each experiment (see Table 3). They show that contents affect the likelihood of counterexample selections, which are most likely with deontic principles, less likely with everyday hypotheses, and least likely with abstract hypotheses (see Table 4).

Third, the model theory avoids a puzzle and a paradox for theories based on logic (e.g., see Theories 3 through 5 in Table 5). Such theories postulate that individuals combine a hypothesis, such as the following: "If there is an E on one side of a card, then there is a 2 on the other side of the card" with an item of evidence, such as the card, E, to infer that there is a 2 on its other side. They therefore choose the card satisfying the clause in the hypothesis: "There is an E on one side of a card." How do they identify the card that fits this description? Proponents of inferential theories, we assume, envisage that they use the meaning of the clause to identify the appropriate card. But, if they can do so, why cannot they use the meaning of the conditional hypothesis itself to determine which cards to select? That is the puzzle. The meaning of hypotheses, as the model theory shows, can be used to select evidence to test their truth or falsity. No compelling reason exists either for the need to use inferences in selecting evidence or against the use of meaning to do so.

The paradox for theories based on logic, as we pointed out earlier, is that a "unicorn" hypothesis, such as,

"All unicorns are invisible" (in logic: for any x, if x is a unicorn then x is invisible) is equivalent in logic to

"No visible things are unicorns" (for any x, if x is visible then x is not a unicorn). So, the observation of a grizzly bear corroborates the unicorn hypothesis: it is visible and not a unicorn (Hempel, 1945). In fact, the unicorn hypothesis is bound to be true in logic. Unicorns do not exist, and so no counterexamples to the hypothesis exist, either—there are no visible unicorns. But, in the model theory, the truth of conditionals demands an additional

corroboration. It is necessary to establish the existence of at least one instance of the hypothesis. The unicorn hypothesis awaits the discovery of an invisible unicorn.

What our account has ruled out is more important than what it has ruled in. Readers can be confident in the following conclusions: (a) No need exists to defend impeccable rationality in the testing of hypotheses. People err in making decisions (see, e.g., Kahneman, 2011), they err in reasoning (see, e.g., Khemlani & Johnson-Laird, 2017), and they err in the selection of evidence (see Table 4). Defenders of rationality have made valiant efforts to explain away the errors. But, the errors are robust, frequent, and predictable. (b) No need exists to invoke probabilities in order to explain the selection of evidence to test conditional hypotheses (pace Kirby, 1994; Oaksford & Chater, 1994). Individuals distinguish quite sharply between *if p then q* and *if p then probably q* (Goodwin, 2014). Despite the ingenuity of probabilistic theories, they have to posit that participants ignore the instructions to the selection task. They call for the selection of evidence to test whether the hypothesis holds for four items (the four cards on the table). And the participants' introspections about performance hardly ever refer to probabilities, but they do refer to truth and falsity. If true means a probability of 1 and false means a probability of 0, then probability becomes an idle wheel in the selection task; and if truth allows a probability of less than 1 and falsity a probability of more than 0, then the selection task, as Wason formulated it, is untestable. No matter what items are on the other side of the four cards, the hypothesis could be true. Very intelligent people, as shown in their level of education or performance on the SAT, tend to make counterexample selections (Hoch & Tschirgi, 1985; Stanovich & West, 1998a, b; Newstead, Handley, Harley, Wright, & Farrelly, 2004), and even probabilists allow that these individuals carry out the task's instructions correctly. A plausible conjecture is that people think probabilistically about the selection task—and indeed about any task—only if its contents, instructions, or framing make at least an implicit reference to probabilities. How they do so is rightly the target of probabilistic theories.

After 50 years of research into the testing of hypotheses, we are bound to ask whether it has paid off and what, if anything, might be done to prevent the future breeding of multiplicities of theories of the same cognitive paradigm. Skeptics might argue that multiple theories are inevitable in studies of the mind. Because we have winnowed away most of the present theories, we reject this view. One antidote to overmultiplication is to require theories to have a purview of more than a single experimental paradigm (Newell, 1973). Another antidote is to require them to describe mental processes and to do so in the form of algorithms (Johnson-Laird, 1983, p. 6). Nearly all the theories in Table 5 fail to meet these two criteria. Still another antidote is to carry out more stringent experiments, but that in turn demands more precise and more sorts of theoretical predictions.

How do individuals choose potential evidence to test a hypothesis in daily life? The answer is perhaps simpler than the size of the literature suggests. When people understand a general hypothesis, such as "If a Brit admires Trump, then the Brit voted for Brexit," its mental models lead them to find out whether Brits who admire Trump also voted for Brexit, and perhaps whether Brits who voted for Brexit also admire Trump. So far, so good. But, they have overlooked a potential counterexample to the hypothesis. To prevent this oversight, they need to go beyond intuition, to deliberate, and to flesh out a model of a counterexample—a Brit who admires Trump but who did not vote for Brexit. This counterexample can guide them to some potentially falsifying evidence. Our exploration has indeed led us back to where we started (Johnson-Laird & Wason, 1970b), and perhaps, as the epigraph to this paper suggests, we now know it for the first time. We know its weaknesses. It needs to use parameters to model insight into falsification. And they are a substitute for our ignorance. The next step in the exploration is to replace them with explanations.

References

Adams, E. W. (1998). *A primer of probability logic*. Stanford, CA: Center for the Study of Language and Information.

Almor, A., & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review, 103,* 374–380. http://dx.doi.org/10.1037/0033-295X.103.2.374

Almor, A., & Sloman, S. A. (2000). Reasoning versus text processing in the Wason selection task: A nondeontic perspective on perspective effects. *Memory & Cognition, 28,* 1060–1070. http://dx.doi.org/10.3758/BF03209354

Augustinova, M., Oberlé, D., & Stasser, G. L. (2005). Differential access to information and anticipated group interaction: Impact on individual reasoning. *Journal of Personality and Social Psychology, 88,* 619–631. http://dx.doi.org/10.1037/0022-3514.88.4.619

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance. *Quarterly Journal of Experimental Psychology, 56,* 1053–1077. http://dx.doi.org/10.1080/02724980244000729

Ball, L. J., Lucas, E. J., & Phillips, P. (2005). Eye movements and reasoning: Evidence for relevance effects and rationalization processes in deontic selection tasks. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 186–201). Alpha, NJ: Sheridan.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York, NY: Cambridge University Press.

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *Quarterly Journal of Experimental Psychology, 40,* 269–297. http://dx.doi.org/10.1080/02724988843000122

Beth, E. W., & Piaget, J. (1966). *Mathematical epistemology and psychology*. Dordrecht, the Netherlands: Reidel.

Bracewell, R. J., & Hidi, S. E. (1974). The solution of an inferential problem as a function of stimulus materials. *The Quarterly Journal of Experimental Psychology, 26,* 480–488. http://dx.doi.org/10.1080/14640747408400437

Brown, C., Keats, J. A., Keats, D. M., & Seggie, I. (1980). Reasoning about implication: A comparison of Malaysian and Australian subjects. *Journal of Cross-Cultural Psychology, 11,* 395–410. http://dx.doi.org/10.1177/0022022180114001

Brown, W. M., & Moore, C. (2000). Is prospective altruist-detection an evolved solution to the adaptive problem of subtle cheating in cooperative ventures? Supportive evidence using the Wason selection task. *Evolution and Human Behavior, 21,* 25–37. http://dx.doi.org/10.1016/S1090-5138(99)00018-5

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology, 50,* 159–193. http://dx.doi.org/10.1016/j.cogpsych.2004.08.001

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416. http://dx.doi.org/10.1016/0010-0285(85)90014-3

Cheng, P. W., & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition, 33,* 285–313. http://dx.doi.org/10.1016/0010-0277(89)90031-0

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328. http://dx.doi.org/10.1016/0010-0285(86)90002-2

Chrostowski, J. J., & Griggs, R. A. (1985). The effects of problem content, instructions, and verbalization procedure on Wason's selection task. *Current Psychology, 4,* 99–107. http://dx.doi.org/10.1007/BF02686577

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–370. http://dx.doi.org/10.1017/S0140525X00009092

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. http://dx.doi.org/10.1016/0010-0277(89)90023-1

Dominowski, R. L. (1995). Content effects in Wason's selection task. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 41–65). Hove, England: Lawrence Erlbaum.

Dyson, F. (2004). A meeting with Enrico Fermi. *Nature, 427,* 297–297. http://dx.doi.org/10.1038/427297a

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 624–629). Alpha, NJ: Sheridan.

Evans, J. S. B. T. (1983). Linguistic determinants of bias in conditional reasoning. *The Quarterly Journal of Experimental Psychology, 35,* 635–644. http://dx.doi.org/10.1080/14640748308402151

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *The Quarterly Journal of Experimental Psychology, 24,* 193–199. http://dx.doi.org/10.1080/00335557243000067

Evans, J. St. B. T. (1977). Toward a statistical theory of reasoning. *The Quarterly Journal of Experimental Psychology, 29,* 621–635. http://dx.doi.org/10.1080/14640747708400637

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75,* 451–468. http://dx.doi.org/10.1111/j.2044-8295.1984.tb01915.x

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

Evans, J. St. B. T. (1993). The mental model theory of conditional reasoning: Critical appraisal and revision. *Cognition, 48,* 1–20. http://dx.doi.org/10.1016/0010-0277(93)90056-2

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: Lawrence Erlbaum.

Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87,* 223–240. http://dx.doi.org/10.1111/j.2044-8295.1996.tb02587.x

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning, 4,* 45–82. http://dx.doi.org/10.1080/135467898394247

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13,* 378–395. http://dx.doi.org/10.3758/BF03193858

Evans, J. St. B. T. (2017). A brief history of the Wason selection task. In N. Galbraith, E. Lucas, & D. Over (Eds.), *The thinking mind: A Festschrift for Ken Manktelow*. New York, NY: Routledge.

Evans, J. St. B. T., & Ball, L. J. (2010). Do people reason on the Wason selection task? A new look at the data of Ball et al. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 63,* 434–441. http://dx.doi.org/10.1080/17470210903398147

Evans, J. St. B. T., Clibbens, J., & Rood, B. (1996). The role of implicit and explicit negation in conditional reasoning bias. *Journal of Memory and Language, 35,* 392–409. http://dx.doi.org/10.1006/jmla.1996.0022

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64,* 391–397. http://dx.doi.org/10.1111/j.2044-8295.1973.tb01365.x

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.

Fairley, N., Manktelow, K. I., & Over, D. E. (1999). Necessity, sufficiency and perspective effects in causal conditional reasoning. *Quarterly Journal of Experimental Psychology, 52A,* 771–790. http://dx.doi.org/10.1080/713755829

Finocchiaro, M. A. (1980). *Galileo and the art of reasoning*. Dordrecht, the Netherlands: Reidel. http://dx.doi.org/10.1007/978-94-009-9017-3

Fugard, A. J. B., & Stenning, K. (2013). Statistical models as cognitive models of individual differences in reasoning. *Argument & Computation, 4,* 89–102. http://dx.doi.org/10.1080/19462166.2012.674061

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. http://dx.doi.org/10.1037/0033-295X.102.4.684

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43,* 127–171. http://dx.doi.org/10.1016/0010-0277(92)90060-U

Gilhooly, K. J., & Falconer, W. A. (1974). Concrete and abstract terms and relations in testing a rule. *The Quarterly Journal of Experimental Psychology, 26,* 355–359. http://dx.doi.org/10.1080/14640747408400424

Girotto, V., Gilly, M., Blaye, A., & Light, P. (1989). Children's performance in the selection task: Plausibility and familiarity. *British Journal of Psychology, 80,* 79–95. http://dx.doi.org/10.1111/j.2044-8295.1989.tb02304.x

Girotto, V., Kemmelmeier, M., Sperber, D., & van der Henst, J. B. (2001). Inept reasoners or pragmatic virtuosos? Relevance and the deontic selection task. *Cognition, 81,* B69–B76. http://dx.doi.org/10.1016/S0010-0277(01)00124-X

Girotto, V., Light, P., & Colbourn, C. (1988). Pragmatic schemas and conditional reasoning in children. *Quarterly Journal of Experimental Psychology, 40,* 469–482. http://dx.doi.org/10.1080/02724988843000023

Golding, E. (1981, April). *The effect of past experience on problem solving*. Paper presented at the British Psychological Society Conference, Surrey University, Guildford, England.

Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General, 143,* 1214–1241. http://dx.doi.org/10.1037/a0034232

Goodwin, R. Q., & Wason, P. C. (1972). Degrees of insight. *British Journal of Psychology, 63,* 205–212. http://dx.doi.org/10.1111/j.2044-8295.1972.tb02101.x

Gralinski, J. H., & Kopp, C. B. (1993). Everyday rules for behavior: Mothers' requests to young children. *Developmental Psychology, 29,* 573–584. http://dx.doi.org/10.1037/0012-1649.29.3.573

Green, D. W. (1995). Externalisation, counter-examples and the abstract selection task. *Quarterly Journal of Experimental Psychology, 48A,* 424–446. http://dx.doi.org/10.1080/14640749508401398

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). New York, NY: Academic Press.

Griggs, R. A. (1984). Memory cueing and instructional effects on Wason's selection task. *Current Psychology, 3,* 3–10. http://dx.doi.org/10.1007/BF02686552

Griggs, R. A. (1995). The effects of rule clarification, decision justification, and selection instruction on Wason's abstract selection task. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 17–39). Hillsdale, NJ: Erlbaum.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73,* 407–420. http://dx.doi.org/10.1111/j.2044-8295.1982.tb01823.x

Griggs, R. A., & Cox, J. R. (1983). The effects of problem content and negation on Wason's selection task. *Quarterly Journal of Experimental Psychology, 35,* 519–533. http://dx.doi.org/10.1080/14640748308402486

Griggs, R. A., & Jackson, S. L. (1990). Instructional effects on responses in Wason's selection task. *British Journal of Psychology, 81,* 197–204. http://dx.doi.org/10.1111/j.2044-8295.1990.tb02355.x

Handley, S. J., Feeney, A., & Harper, C. (2002). Alternative antecedents, probabilities, and the suppression of fallacies in Wason's selection task. *Quarterly Journal of Experimental Psychology, 55,* 799–818. http://dx.doi.org/10.1080/02724980143000497

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology: Section A, 55,* 1241–1272. http://dx.doi.org/10.1080/02724980244000053

Hempel, C. G. (1945). Studies in the logic of confirmation, Parts I and II. *Mind, 54,* 1–26, 97–121. http://dx.doi.org/10.1093/mind/LIV.213.1

Hewitt, C. (1971). Procedural embedding of knowledge in Planner. *International Joint Conference on Artificial Intelligence, 1,* 167–184.

Hoch, S. J., & Tschirgi, J. E. (1985). Logical knowledge and cue redundancy in deductive reasoning. *Memory & Cognition, 13,* 453–462. http://dx.doi.org/10.3758/BF03198458

Jeffrey, R. (1981). *Formal logic*. New York, NY: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109,* 646–678. http://dx.doi.org/10.1037/0033-295X.109.4.646

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19,* 201–214. http://dx.doi.org/10.1016/j.tics.2015.02.006

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63,* 395–400. http://dx.doi.org/10.1111/j.2044-8295.1972.tb01287.x

Johnson-Laird, P. N., & Tagart, J. (1969). How implication is understood. *The American Journal of Psychology, 82,* 367–373. http://dx.doi.org/10.2307/1420752

Johnson-Laird, P. N., & Wason, P. C. (1970a). Insight into a logical relation. *The Quarterly Journal of Experimental Psychology, 22,* 49–61. http://dx.doi.org/10.1080/14640747008401901

Johnson-Laird, P. N., & Wason, P. C. (1970b). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology, 1,* 134–148. http://dx.doi.org/10.1016/0010-0285(70)90009-5

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Strauss, Giroux.

Khemlani, S. S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines, 27,* 11–35. http://dx.doi.org/10.1007/s11023-017-9421-x

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51,* 1–28. http://dx.doi.org/10.1016/0010-0277(94)90007-8

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 680–703. http://dx.doi.org/10.1037/0278-7393.33.4.680

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94,* 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211

Kolodner, J. (1993). *Case-based reasoning*. San Mateo, CA: Morgan Kaufmann. http://dx.doi.org/10.1016/B978-1-55860-237-3.50005-4

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives, 27,* 312–365. http://dx.doi.org/10.1111/phpe.12029

Krauth, J. (1982). Formulation and experimental verification of models in propositional reasoning. *Quarterly Journal of Experimental Psychology, 34A,* 285–298. http://dx.doi.org/10.1080/14640748208400842

Kroger, J. K., Cheng, P. W., & Holyoak, K. J. (1993). Evoking the permission schema: The impact of explicit negation and a violation-checking context. *Quarterly Journal of Experimental Psychology, 46,* 615–635. http://dx.doi.org/10.1080/14640749308401030

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford & Chater (1994). *Psychological Review, 103,* 364–373. http://dx.doi.org/10.1037/0033-295X.103.2.364

Leighton, J. P., & Dawson, M. R. W. (2001). A parallel distributed processing model of Wason's selection task. *Journal of Cognitive Systems Research, 2,* 207–231. http://dx.doi.org/10.1016/S1389-0417(01)00035-3

Liberman, N., & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange cheating detection or task understanding. *Cognition, 58,* 127–156. http://dx.doi.org/10.1016/0010-0277(95)00677-X

Light, P. H., Girotto, V., & Legrenzi, P. (1990). Children's reasoning on conditional promises and permissions. *Cognitive Development, 5,* 369–383. http://dx.doi.org/10.1016/0885-2014(90)90002-B

Love, R., & Kessler, C. (1995). Focusing in Wason's selection task: Content and instruction effects. *Thinking & Reasoning, 1,* 153–182. http://dx.doi.org/10.1080/13546789508251502

Lucas, E., & Ball, L. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalisation processes. *Thinking & Reasoning, 11,* 35–66. http://dx.doi.org/10.1080/13546780442000114

Lunzer, E. A., Harrison, C., & Davey, M. (1972). The four-card problem and the generality of formal reasoning. *The Quarterly Journal of Experimental Psychology, 24,* 326–339. http://dx.doi.org/10.1080/14640747208400288

Manktelow, K. I., & Evans, J. St. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology, 70,* 477–488. http://dx.doi.org/10.1111/j.2044-8295.1979.tb01720.x

Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition, 39,* 85–105. http://dx.doi.org/10.1016/0010-0277(91)90039-7

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

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-170150-5.50012-3

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. http://dx.doi.org/10.1080/02724980343000116

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning, 2,* 1–31. http://dx.doi.org/10.1080/135467896394546

Nickerson, R. S. (2015). *Conditional reasoning*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780190202996.001.0001

Oakhill, J. V., & Johnson-Laird, P. N. (1985). Rationality, memory and the search for counterexamples. *Cognition, 20,* 79–94. http://dx.doi.org/10.1016/0010-0277(85)90006-X

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. http://dx.doi.org/10.1037/0033-295X.101.4.608

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103,* 381–391. http://dx.doi.org/10.1037/0033-295X.103.2.381

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10,* 289–318. http://dx.doi.org/10.3758/BF03196492

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198524496.001.0001

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 883–899. http://dx.doi.org/10.1037/0278-7393.26.4.883

Oaksford, M., & Hall, S. (2016). On the source of human irrationality. *Trends in Cognitive Sciences, 20,* 336–344. http://dx.doi.org/10.1016/j.tics.2016.03.002

Oaksford, M., & Moussakowski, M. (2004). Negations and natural sampling in data selection: Ecological versus heuristic explanations of matching bias. *Memory & Cognition, 32,* 570–581. http://dx.doi.org/10.3758/BF03195848

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition, 31,* 143–154. http://dx.doi.org/10.3758/BF03196089

Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology, 53,* 238–283. http://dx.doi.org/10.1016/j.cogpsych.2006.04.001

Oberauer, K., Weidenfeld, A., & Hörnig, R. (2004). Logical reasoning and probabilities: A comprehensive test of Oaksford And Chater (2001). *Psychonomic Bulletin & Review, 11,* 521–527. http://dx.doi.org/10.3758/BF03196605

Oberauer, K., Wilhelm, I. V. O., & Rosas-Diaz, R. R. (1999). Bayesian rationality for the Wason selection task? A test of optimal data selection theory. *Thinking & Reasoning, 5,* 115–144. http://dx.doi.org/10.1080/135467899394020

Osherson, D. N. (1976). *Logical abilities in children* (Vol. 4). Hillsdale, NJ: Lawrence Erlbaum.

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, 46,* 591–613. http://dx.doi.org/10.1080/14640749308401029

Platt, R. D., & Griggs, R. A. (1995). Facilitation and matching bias in the abstract selection task. *Thinking & Reasoning, 1,* 55–70. http://dx.doi.org/10.1080/13546789508256905

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 49,* 447–462. http://dx.doi.org/10.1080/713755628

Politzer, G., & Nguyen-Xuan, A. (1992). Reasoning about promises and warnings: Darwinian algorithms, mental models, relevance judgments or pragmatic schemas? *Quarterly Journal of Experimental Psychology, 44A,* 402–421.

Pollard, P. (1981). The effect of thematic content on the 'Wason selection task'. *Current Psychological Research, 1,* 21–29. http://dx.doi.org/10.1007/BF02684422

Pollard, P. (1985). Nonindependence of selections on the Wason selection task. *Bulletin of the Psychonomic Society, 23,* 317–320. http://dx.doi.org/10.3758/BF03330170

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Quelhas, A. C., Johnson-Laird, P. N., & Juhos, C. (2010). The modulation of conditional assertions and its effects on reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 63,* 1716–1739. http://dx.doi.org/10.1080/17470210903536902

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2017). A priori true and false conditionals. *Cognitive Science, 41*(Suppl. 5), 1003–1030. http://dx.doi.org/10.1111/cogs.12479

Ragni, M., & Johnson-Laird, P. N. (2017). *Reasoning about possibilities*. Manuscript submitted for publication.

Ramsey, F. R. (1990). Philosophy. In D. H. Mellor (Ed.), *F. R. Ramsey, philosophical papers* (pp. 1–7). New York, NY: Cambridge University Press. (Original work published 1931)

R Development Core Team. (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Reich, S. S., & Ruth, P. (1982). Wason's selection task: Verification, falsification and matching. *British Journal of Psychology, 73,* 395–405. http://dx.doi.org/10.1111/j.2044-8295.1982.tb01822.x

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339. http://dx.doi.org/10.1037/0033-295X.95.3.318

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *Quarterly Journal of Experimental Psychology, 54,* 1031–1048. http://dx.doi.org/10.1080/713756016

Roth, E. M. (1979). Facilitating insight in a reasoning task. *British Journal of Psychology, 70,* 265–271. http://dx.doi.org/10.1111/j.2044-8295.1979.tb01684.x

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences, 20,* 883–893. http://dx.doi.org/10.1016/j.tics.2016.10.003

Schroyens, W., & Schaeken, W. (2003). A critique of Oaksford, Chater, and Larkin's (2000). conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 140–149. http://dx.doi.org/10.1037/0278-7393.29.1.140

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning, 7,* 121–172. http://dx.doi.org/10.1080/13546780042000091

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Shannon, C. E. (1948). A mathematical theory of communication, part I. *The Bell System Technical Journal, 27,* 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New NY: McGraw-Hill.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods, 45,* 560–575. http://dx.doi.org/10.3758/s13428-012-0259-0

Smalley, N. S. (1974). Evaluating a rule against possible instances. *British Journal of Psychology, 65,* 293B69–304A. http://dx.doi.org/10.1111/j.2044-8295.1974.tb01404.x

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57,* 31–95. http://dx.doi.org/10.1016/0010-0277(95)00666-M

Sperber, D., & Wilson, D. (1995). *Relevance* (2nd ed.). Oxford, England: Blackwell.

Stahl, C., Klauer, K. C., & Erdfelder, E. (2008). Matching bias in the selection task is not eliminated by explicit negations. *Thinking & Reasoning, 14,* 281–303. http://dx.doi.org/10.1080/13546780802116807

Staller, A., Sloman, S. A., & Ben-Zeev, T. (2000). Perspective effects in nondeontic versions of the Wason selection task. *Memory & Cognition, 28,* 396–405. http://dx.doi.org/10.3758/BF03198555

Stanovich, K. E., & West, R. F. (1998a). Cognitive ability and variation in selection task performance. *Thinking & Reasoning, 4,* 193–230. http://dx.doi.org/10.1080/135467898394139

Stanovich, K. E., & West, R. F. (1998b). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188. http://dx.doi.org/10.1037/0096-3445.127.2.161

Stenning, K., & van Lambalgen, M. S. (2008). *Human reasoning and cognitive science*. Cambridge, MA: MIT Press.

Strawson, P. F. (1952). *Introduction to logical theory*. London, England: Methuen.

Tweney, R. D., & Yachanin, S. A. (1985). Can scientists rationally assess conditional inferences? *Social Studies of Science, 15,* 155–173. http://dx.doi.org/10.1177/030631285015001006

Valentine, E. R. (1985). The effect of instructions on performance in the Wason selection task. *Current Psychology, 4,* 214–223. http://dx.doi.org/10.1007/BF02686572

van Duyne, P. C. (1973). A short note on Evans' criticism of reasoning experiments and his matching response hypothesis. *Cognition, 2,* 239–242. http://dx.doi.org/10.1016/0010-0277(72)90013-3

van Duyne, P. C. (1974). Realism and linguistic complexity in reasoning. *British Journal of Psychology, 65,* 59–67. http://dx.doi.org/10.1111/j.2044-8295.1974.tb02771.x

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. http://dx.doi.org/10.1037/a0022790

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology, 12,* 12–40.

Wason, P. C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4,* 7–11. http://dx.doi.org/10.1016/S0022-5371(65)80060-3

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*. Harmondsworth, England: Penguin Books.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20,* 273–281. http://dx.doi.org/10.1080/14640746808400161

Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3,* 141–154. http://dx.doi.org/10.1016/0010-0277(74)90017-1

Wason, P. C., & Green, D. W. (1984). Reasoning and mental representation. *Quarterly Journal of Experimental Psychology, 36,* 597–610. http://dx.doi.org/10.1080/14640748408402181

Wason, P. C., & Johnson-Laird, P. N. (1969). Proving a disjunctive rule. *The Quarterly Journal of Experimental Psychology, 21,* 14–20. http://dx.doi.org/10.1080/14640746908400189

Wason, P. C., & Johnson-Laird, P. N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology, 61,* 509–515. http://dx.doi.org/10.1111/j.2044-8295.1970.tb01270.x

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology, 23,* 63–71. http://dx.doi.org/10.1080/00335557143000068

Wetherick, N. E. (1970). On the representativeness of some experiments in cognition. *Bulletin of the British Psychological Society, 23,* 213–214.

Wetherick, N. E. (1995). Reasoning and rationality: A critique of some experimental paradigms. *Theory & Psychology, 5,* 429–448. http://dx.doi.org/10.1177/0959354395053009

Yachanin, S. A. (1986). Facilitation in Wason's selection task: Content and instructions. *Current Psychology, 5,* 20–29. http://dx.doi.org/10.1007/BF02686593

Yachanin, S. A., & Tweney, R. D. (1982). The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society, 19,* 87–90. http://dx.doi.org/10.3758/BF03330048

Yama, H. (2001). Matching versus optimal data selection in the Wason selection task. *Thinking & Reasoning, 7,* 295–311. http://dx.doi.org/10.1080/13546780143000053

Received July 4, 2017
Revision received January 3, 2018
Accepted January 6, 2018 ■