Computational Brain & Behavior
https://doi.org/10.1007/s42113-020-00077-9

ORIGINAL PAPER

Explanation or Modeling: a Reply to Kellen and Klauer

Marco Ragni[1] · P. N. Johnson-Laird[2,3]

Abstract
In Wason's "selection" task, individuals often overlook potential counterexamples in selecting evidence to test hypotheses. Our recent meta-analysis of 228 experiments corroborated the main predictions of the task's original theory, which aimed to explain the testing of hypotheses. Our meta-analysis also eliminated all but 1 of the 15 later theories. The one survivor was the inference-guessing theory of Klauer et al., but it uses more free parameters to model the data. Kellen and Klauer (this issue) dissent. They defend the goal of a model of the frequencies of all 16 possible selections in Wason's task, including "guesses" that occur less often than chance, such as not selecting any evidence. But an explanation of hypothesis testing is not much advanced by modeling such guesses with independent free parameters. The task's original theory implies that individuals tend to choose items of evidence that are dependent on one another, and the inference-guessing theory concurs for those selections that are inferred. Kellen and Klauer argue against correlations as a way to assess dependencies. But our meta-analysis did not use them; it used Shannon's measure of information to establish dependencies. Their modeling goal has led them to defend a "purposely vague" theory. Our explanatory goal has led us to defend a "purposely clear" algorithm and to retrieve long-standing evidence that refutes the inference-guessing theory. Individuals can be rational in testing a hypothesis: in repeated tests, they search for some examples of it, and then exhaustively for counterexamples.

Keywords Conditionals · Counterexamples · Hypothesis testing · Modeling data · Selection task · Theory comparison

> With four [free] parameters I can fit an elephant, and with five I can make him wiggle his trunk. (von Neumann cited in Dyson 2004)

Kellen and Klauer (this issue) disagree with our meta-analysis of studies of the selection task (Ragni et al. 2018). We thank them for their critique. It has led us to several discoveries. One root of the disagreement is the difference between the two theories—the *model* theory of Ragni et al. (ibid.) and the *inference-guessing* theory of Klauer et al. (2007). Another root is the difference in our respective goals, which is more a matter of scientific taste than an empirical issue. Our goal is to apply a general explanation of thinking—the model theory—to the mental processes of testing hypotheses. Kellen and Klauer (henceforth K&K) have the goal, following Klauer et al., of a precise model of the frequencies of occurrence of each of the 16 possible selections in tests of abstract conditional hypotheses. Their goal has led them to criticize two of the three main predictions of our theory and our fit of their theory (and ours) to experimental data. In this reply, we defend the model theory, restore its two predictions that they criticize, and show that the inference-guessing theory, despite its premeditated vagueness, is false.

Our goal goes back to Peter Wason's invention of the selection task. Consider this hypothesis:

> If anyone has cholera, then they have had close contact with an infected person.

Doctors in the nineteenth century realized that counterexamples occurred: people caught the disease without personal contact, and so the hypothesis was false. The resulting controversy about how the disease leapt large distances led John Snow and others to found epidemiology (see, e.g., Johnson-

✉ Marco Ragni
ragni@informatik.uni-freiburg.de

P. N. Johnson-Laird
phil@princeton.edu

[1] Cognitive Computation Laboratory, Technical Faculty, University of Freiburg, 79110 Freiburg, Germany

[2] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

[3] New York University, New York, NY, USA

Published online: 15 April 2020

Laird 2006, Chap. 27). The selection task was one of Wason's experimental paradigms devised to find out whether naive individuals—those innocent of science, logic, or philosophy—understood the importance of counterexamples. Later, Johnson-Laird devised another task with the same goal, the *repeated* selection task.

In Wason's (1966, 1968) initial studies, participants had to select potential evidence to test a general but abstract hypothesis, such as *If there is a D on one side of a card, then there is a 3 on the other side*. They were told to select all, and only, those cards—from the four on the table in front of them—that they needed to turn over to find out whether the hypothesis was true or false about the four cards. Different individuals made different selections where by definition a *selection* is the set of cards that a participant chose. The first four studies of the task yielded these percentages of selections for abstract conditional hypotheses, *if p then q*:

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

where pq signifies the selection of the two cards corresponding to the clauses in the conditional (D and 3 in the hypothesis above), and the bar over q signifies the selection of the not-q card (the 2 card in the hypothesis above). We refer to these four different selections as *canonical*. There are 16 possible selections, including selecting all four cards and selecting none of them. And the remaining 10% of miscellaneous selections in the first four studies each occurred less often than chance (1/16 = 6.25%) or not at all, e.g., there were no selections of q alone (Johnson-Laird and Wason 1970a, p. 136). The important result was a negative one: most participants failed to select the $\overline{q}$ card. But with p on its other side, it is a counterexample falsifying the hypothesis.

Wason thought that most people relied only on their intuitions whereas J-L took comfort from the few who deliberated and had insight into the power of potential counterexamples. He therefore devised an algorithm simulating a theory that allowed a switch from intuition to deliberation (Johnson-Laird and Wason 1970a). It was a dual-process theory, one of the first in modern studies, and perhaps unique in having an algorithmic description (cf. Evans 1984). The algorithm represents the meaning of any hypothesis, such as a conditional or a disjunction (as used in Wason and Johnson-Laird's (1969) study). And for a conditional *if p then q*, it yields intuitive selections of pq or p depending on whether or not the conditional is interpreted as implying its converse. If deliberation leads to a partial insight into falsification, it can yield pq$\overline{q}$, and complete insight yields only the correct selection, p$\overline{q}$ (see Johnson-Laird and Wason 1970a, Fig. 2).

The failure of participants to select potential counterexamples was stunning. And defenders of human rationality pounced (see Ragni et al. 2018) for a description of their arguments). Prior to their criticisms, however, Johnson-Laird and Wason (1970b) reported on the repeated selection task. In this paradigm, the participants make a series of choices that can be q or $\overline{q}$ and get immediate feedback that q occurs with p, and $\overline{q}$ occurs with $\overline{p}$. Nearly every participant in the first study began by selecting instances of q, but they soon realized the importance of potential counterexamples, and then they all selected every instance of $\overline{q}$. Critics of the selection task, like its later investigators, often overlooked the rationality of these selections.

The Meta-analysis

The initial investigators aimed to explain the mental processes underlying the testing of hypotheses (Wason and Johnson-Laird 1972, Chap. 13–15). And they thought that they had made enough progress to allow them to take up different topics for research. Others, however, started to investigate the selection task. What shocked them was the robust but irrational neglect of counterexamples. Half a century later, what shocked us was the existence of 16 theories of the task. That was not a sign of scientific progress. So, we decided to try to use evidence to refute as many theories as possible. Back we went to the original theory and implemented its algorithm in the programming language Python. Unlike the original version, our implementation copes only with conditionals because they are the focus of almost all the experiments in the literature, and we replaced the truth table for a conditional with its core meaning in the model theory (see, e.g., Johnson-Laird et al. 2015). This replacement has no effect on the predictions for the selection task, but as we will see, it creates a more sensible normative account of hypothesis testing. The model theory's algorithm for conditionals works in an identical way to the original one.

Some of the 16 theories seemed too sketchy, some seemed too sophisticated, and almost all seemed too narrow—purpose-built for the standard selection task so that if the hypothesis were changed to a disjunction, they would not work. Some could not be formulated in an algorithm, and some could not be used to fit data. What to do? In the end, we asked one decisive question: does experimental evidence refute them? The model theory makes three main predictions. The meta-analysis corroborated them (Ragni et al. 2018) and refuted all but one of the alternative theories. The only surviving alternative was the inference-guessing theory due to Klauer et al. (2007). Their theory concerns the abstract selection task with a conditional hypothesis, and the researchers wrote: "The purpose … is to develop and validate a mathematical model for the 16 possible selection patterns" (ibid., p. 681). Given this goal, their theory opts to have no account of what conditionals mean or of how people make the inferences that it

postulates as yielding selections. As K&K comment, "The model is purposely vague about the nature of the underlying reasoning process." It sufficed for Klauer et al. to formulate a multinomial processing tree with ten free parameters to fit the frequencies of the 16 selections in their large online experiments. So, individuals either guess or infer a selection (one parameter). If they guess, their guesses for each of the four cards are independent of one another (four parameters). If they reason, they make one or two inferences (five parameters) yielding 11 different selections. These ten parameters are *free* in that they can take any values whatsoever in order to yield the best possible fit to the data. Indeed, 10 parameters for 16 outcomes do an extravagantly good job.

We now turn to K&K's critique of the model theory. We deal first with their criticisms of its three main predictions. We introduce three of its other predictions whose corroborations refute the inference-guessing theory. Finally, we address their analysis of fitting theories to data.

The Three Main Predictions of the Model Theory

Prediction 1: Only the four canonical selections should occur more often than chance

The model theory postulates that the meaning of a general conditional, *if p then q*, refers to the possibility of *p and q* in default of knowledge to the contrary and to the possibility of *not-p* whether the conditional is true or false (Johnson-Laird et al. 2015). It follows that a general conditional is true given that there is at least some example of p and q and no counterexamples of p and $\overline{q}$. The theory therefore predicts that four canonical selections should occur more often than chance: pq and p, which are examples of the hypothesis depending on whether or not it is interpreted as implying its converse; $pq\overline{q}$, which reflects partial insight into falsification; and $p\overline{q}$, which reflects complete insight. Any other of the 16 possible selections should not occur more often than chance (6.25%). The theory recognizes that all cognitive experiments may have participants who cannot or will not do the task, who guess, or who make eccentric selections or none at all, e.g., one of Wason's initial participants, alas, went into a catatonic trance. K&K write: "[Guessing processes] play a critical role when evaluating the merit of competing theories, especially when doing so on the basis of aggregate data." But guesses, refusals, and eccentricities, which fall into their category of "guesses", do not elucidate the mental processes underlying the testing of hypotheses. Their analogs occur in other cognitive tasks, such as syllogistic reasoning.

K&K claim that there is no clear rationale for treating certain selections as canonical and others as "idiosyncratic and rare". In fact, a clear rationale existed from the beginning: All and only selections occurring more often than chance should be treated as canonical. Three selections occurred more often than chance in the initial studies (see above), and we added the correct selection, in part because it *is* correct and in part because over 30% of participants made it in another early study (Wason 1969). Subsequent experiments confirmed these four selections as canonical, but they also raised the frequencies of two other selections to a marginal status (of 6% over all contents): q alone, and all four cards: $p\overline{p}$ $q\overline{q}$ (see Ragni et al. 2018, Table 2). As Ragni et al. argued, good theoretical grounds exist for rejecting these selections as idiosyncratic in the standard selection task.

K&K make three arguments for changes to the set of canonical selections. *First, an experimental manipulation could make a selection, such as q alone, occur more often than chance*. It might. But consider again the hypothesis that if anyone has cholera then they have had contact with an infected person. To investigate it by studying only those who have had contact with a cholera victim (q alone) would be stupid because it would be impossible to discover that the hypothesis is false. So, a manipulation that increases the frequency of this hitherto negligible or non-existent selection may show only that experiments can stupidify their participants. *Second, an exclusion of selections such as guesses could distort the interpretation of thoughtful processes.* We agree. Guessing can yield a canonical selection, too. Some way to correct for this problem is needed to assess the fit of a theory to data. But more is at stake than goodness of fit. Experimental results can refute theories. *Third, the partial insight pattern, $pq\overline{q}$, which passes the criterion for a canonical selection, is no more frequent than q alone or of all four cards, and so either the prediction of canonical selections is false or else these two selections should be included in the set.* What this claim overlooks is that the partial insight selection became the most frequent of all in Wason's (1969) study. And the model theory allows that an explicit biconditional, *if and only if p then q*, should lead to the selection of all four cards (see, e.g., Ragni et al., Fig. 2), but none of the studies in our meta-analysis used explicit biconditionals. Some of the everyday contents in selection tasks may have suggested this interpretation, which may explain the 9% selections of them (see ibid., Table 2).

The case for adding q alone and all four cards to the canonical set rests on the frequencies of their selections in later experiments. So, what happened to increase these frequencies? Some studies may no longer have instructed participants to be economical, i.e., to select *only* those cards relevant to truth or falsity. When economy matters, no one selects all four cards. A major change, however, was to test participants in online studies. It led to a greater diversity of selections than in face-to-face experiments. The median amount of statistical information—a measure that we explain below—in the selections from the 89 face-to-face experiments with 4230 participants was 2.15 bits, whereas from the 10 online experiments

with 3787 participants, the median was 2.67 bits (a highly reliable difference, Mann-Whitney test, $W = 172$, $p < .0002$). So, the selections in online studies are more idiosyncratic than those in earlier studies, e.g., q alone now occurs whereas it never occurred in the four initial experiments (see above). No need exists to change the canonical set. Moreover, the addition of new selections to it cannot rehabilitate any theories that fail to predict the four original ones. The set is viable, and its corroboration refutes 12 of the alternative theories.

Prediction 2: Choices in selections should be dependent

Both the model theory and the inference-guessing theory imply that some selections should be based on choices of cards that are dependent on one another. In statistics, two events, such as A and B, are *dependent* on one another if the probability of A does not equal the conditional probability of A given B (see Feller 1957, p. 115). Hence, some studies used correlations between different pairs of cards to confirm dependency (see Ragni et al. 2018, for a review).

K&K argue that aggregate data of selections do not permit inferences about subject-level dependencies. We agree. An individual participant who makes a selection may have thought or guessed. K&K illustrate their claim with a significant correlation between the selections of two cards but from data that can be split into two subsets that neither yield a reliable correlation. We agree again: it can be dangerous to evaluate dependency from correlations. But we have never used them to assess dependency. We relied on a different method (see Ragni et al. 2018).

Suppose you believe that people always select each card independently of the others. You notice that two cards often occur together in a selection so that, given their independence, they should each occur without the other in appropriate frequencies. Yet, they do not. And so you wonder about their independence, and you are right to do so. Is there a way to transform this thought experiment into a test that applies to all the selections in an experiment?

There is. The test considers the probability in an experiment of each possible selection, P_i, where i denotes the *i*th selection in the set of 16. It multiplies this *i*th probability by its logarithm to the base 2: $P_i \log_2 P_i$. The result is Shannon's H, the *informativeness* of each of the 16 sorts of selection. The measure is additive, and their sum yields the amount of information in the experiment as a whole:

$$H = -\Sigma P_i \log_2 P_i$$

where the minus sign switches the sign of the sum because $\log_2$ of a probability between 0 and 1 is a negative number. A simple Monte Carlo procedure can make a wholly independent selection: it decides whether or not to choose a card for a selection solely from its overall probability of occurrence over all the selections in an actual experiment, and it uses this procedure on all four cards in order to create the selection. It does not generate just one such selection; it generates the same number of selections that occurred in the actual experiment. The informativeness (H) of this simulated experiment will tend to be quite high because nothing constrains selections other than the individual probabilities of the four cards. There are many possible such simulations. So, the test generates 10,000 simulations of the experiment. Their overall mean informativeness approximates to the H for independent selections. In contrast, if the actual experiment embodied dependent choices of cards, it will be less informative. In fact, the informativeness of the 99 experiments in the literature—those that recorded the frequencies of all 16 selections—was often smaller than any of the 10,000 values in their simulations. Some exceptions occurred, but overall, the actual experiments were reliably less informative than their simulations embodying independence. Their reduced informativeness shows that the actual selections of cards were often dependent on one another.

Can we divide an experiment yielding a lower H than the mean of its simulations into two subsets that both yield higher H's than the means of their respective simulations? The additive nature of H makes such a division unlikely. But we tested the possibility using an experiment yielding a lower value of H than its simulations. A program split the experimental data into two arbitrary subsets at random 10,000 times. In none of these splits did both subsets yield H's larger than their respective simulations based on independent selections. Hence, such a split is unlikely to occur often enough to refute Ragni et al.'s (2018) corroboration of dependent selections. It occurred in experiments with the abstract selection task, with everyday conditionals, and with the deontic version of the task (ibid., Table 3). The results corroborate both the model theory and the inference-guessing theory and refute nine theories that do not predict dependent selections.

Prediction 3: Salience of counterexamples should increase the selection of $p\bar{q}$ for conditionals

The model theory predicts that manipulations that increase the saliency of potential counterexamples should increase the likelihood that participants make the correct selection of $p\bar{q}$ for conditionals. K&K do not criticize this prediction, which experimental results corroborate (see Ragni et al. 2018). But they do claim: "there are many more reliable findings in the Wason selection-task paradigm". In fact, no findings from the 228 experiments appear to be either so large or so robust (see Ragni et al. 2018, Table 4). Yet, ten theories of the selection task fail to make the prediction.

In summary, the meta-analysis corroborated the model theory's three main predictions. Of the remaining 15 theories of the selection task, 6 make none of the three predictions, 5 make

one of the predictions, and 3 make two of the predictions. Strictly speaking, the inference-guessing theory does not *make* the predictions, but it can tweak its free parameters to model them.

Evidence Against the Inference-Guessing Theory

The model theory makes other predictions apart from its three principal ones, and it is time to consider a further three of them. Prompted by K&K's critique, we have discovered existing evidence that corroborates them, and that is contrary to the inference-guessing theory.

Individuals Use the Meaning of a Hypothesis to Select Evidence to Test It The alternative according to the inference-guessing theory is that individuals make inferences from the hypothesis. What is at stake is a participant selecting an item of evidence, $\overline{q}$, because its occurrence with p falsifies the hypothesis, as opposed to selecting $\overline{q}$ because together with the hypothesis *if p then q* it implies $\overline{p}$. The two processes are different: one depends on a counterexample to a meaning, and the other depends on a modus tollens deduction, which could be carried out—as Klauer et al. allow—using, not meaning, but a formal rule of inference. No experiments seem to have addressed which of the two processes occurs. But, when participants had to justify their selections, their reasons for selecting p tended to be to determine whether the hypothesis was true, and their reasons for selecting $\overline{q}$ tended to be to determine whether the hypothesis was false (Goodwin and Wason 1972). They referred to truth or falsity rather than to ponens and tollens. The phenomenon corroborates the model theory but is contrary to the inference-guessing theory.

The Meaning of a Hypothesis Determines the* Correct *Selection The correct selection for a conditional hypothesis is $p\overline{q}$. Why? According to the model theory, it follows from the meaning of the conditional hypothesis (see Prediction 1 above). The inferences in the inference-guessing theory, however, neither identify the correct selection nor explain why it is correct (see Fig. 1 below). Its authors presuppose the "logically correct solution" (Klauer et al., p. 680). But orthodox logic has no need for examples of conditional hypotheses in order to verify them, and so it leads to the "paradox" of confirmation (Hempel 1945). A conditional hypothesis, such as:

If anything is a black hole then it has a massive gravity

is equivalent in orthodox logic to:

If anything does not have a massive gravity then it is not a black hole.

Koala bears do not have a massive gravity and are not black holes, and so they corroborate the hypothesis about black holes. Philosophers and others have sought to eradicate this paradox. The model theory postulates a different meaning for conditionals (see Prediction 1 above). Confirmation of a general conditional hypothesis calls for at least one example of it to exist—a black hole with a massive gravity, and for no counterexamples to exist—black holes without a massive gravity. Koalas are irrelevant. The selection task cannot discriminate between this meaning for a conditional and its meaning in orthodox logic. But the repeated selection task can and does: people look for examples of a hypothesis, and once they have found some, they switch to an exhaustive search for potential counterexamples (Johnson-Laird and Wason 1970b). The inference-guessing theory cannot predict this exhaustive search. Likewise, it cannot predict the ability of participants to explain the correctness of the selection $p\overline{q}$ (Wason and Johnson-Laird 1972, p. 173–4).

***A Partial Insight into Falsification of* if p then q *Can Lead to the Selection of* $p\overline{q}$** This selection is a crucial test between the two theories—a result of deliberation according to the model theory, but a result only of "guessing" according to the inference-guessing theory. Suppose an experimental manipulation made the partial insight selection the most popular of all with a frequency so high, and with such a paucity of other selections, that its lowly informativeness could not be a result of independent guessing. The finding would bear out the model theory, but refute the inference-guessing theory. Wason (1969) reported such a result nearly 40 years before the publication of the latter theory. Each of a series of experiences, such as imagining what was on the other side of a card, made participants in an experiment more likely to think of counterexamples. By the end of the experiment, the most frequent selection was the partial insight one (53% of participants), and the next most frequent was the correct selection (31% of participants). The increase in the partial insight selection could not have resulted from independent guesses of each card. And the final selections were far too dependent to result from independent guessing.

Subjective reports, such as justifications of selections, are indicative rather than decisive. The repeated selection task is not the standard selection task. But the experimental manipulation of the partial insight selection refutes the inference-guessing theory, and the other results support its refutation.

Some Technical Nuggets in Fitting the Theories to Data

Another way to assess theories is to determine how well they fit data. The inference-guessing theory has ten free parameters, but if a difference in fit between two theories depends

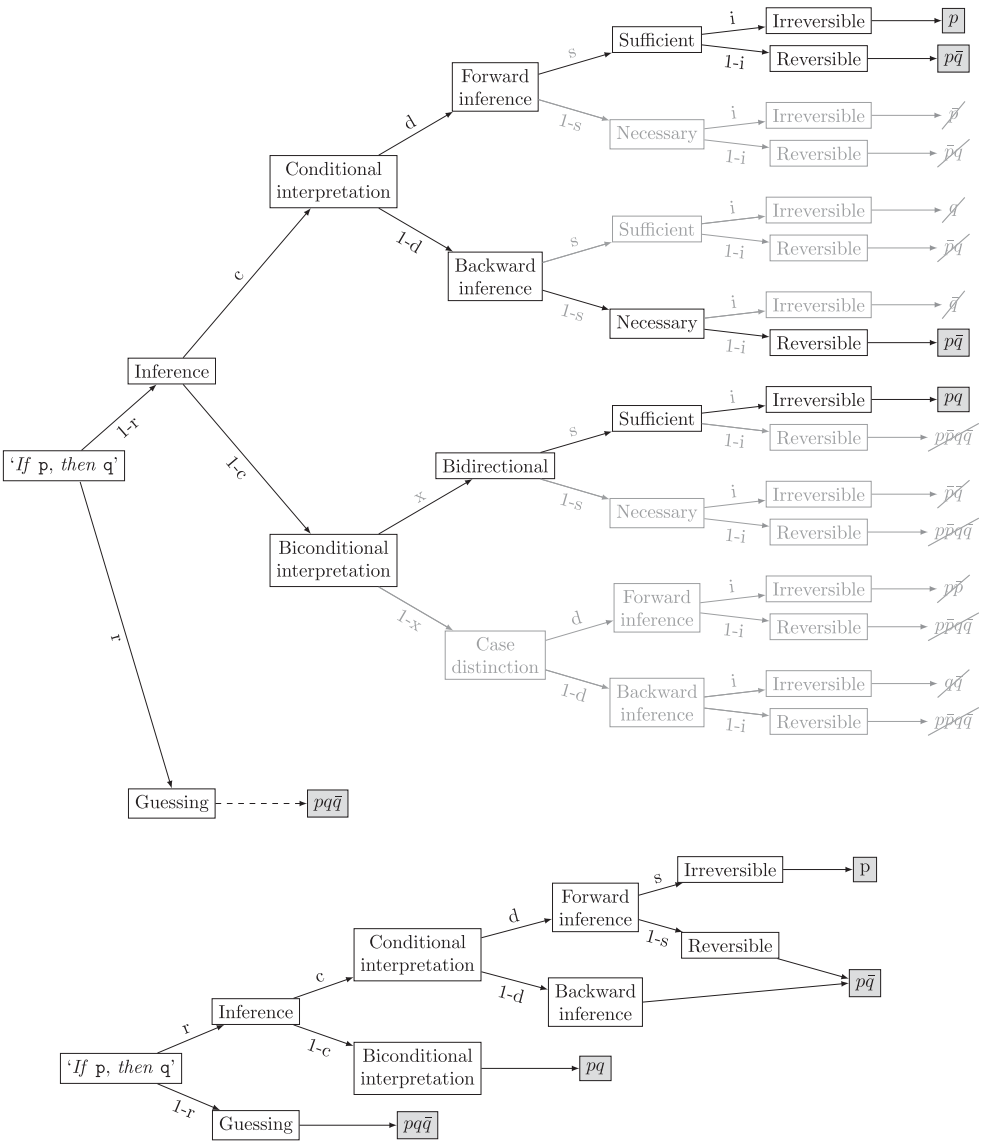

Fig. 1 The tree for the inference-guessing theory in Klauer et al. (2007) in which we have crossed out those branches we pruned, and beneath it is the resulting pruned tree in Ragni et al. (2018) that requires four parameters (in which we have corrected two erroneous labels, which appear to have misled K&K—our fault, for which we apologize)

only on four independent parameters for guessing each card, then its results would hardly be conclusive. Indeed, when K&K made such a fit by adding four parameters for guessing to the model theory, they report that the two theories were "succeeding and failing together". In contrast, Ragni et al. (2018) simplified the inference-guessing theory in order to compare its fit with the model theory's for all 228 experiments, and the model theory showed signs of a slightly better fit. But K&K made some cogent criticisms of our procedure, and here we reply to their main points.

1. Our pruning of their multinomial processing tree to address only canonical selections "risks distorting and/or limiting [its] ability to account for data at large". There is no need to reduce the number of its parameters to fit it to the four canonical selections. The only issue "[is] the inability to obtain a unique set of best fitting parameters …". But had we neither pruned its tree nor introduced a new path for the selection of pq$\bar{\text{q}}$, we could have been accused of not enabling the theory to make the canonical selections. Figure 1 presents both Klauer et al.'s original tree and the result of our pruning. It shows that the pruning itself did not err. The pruned tree yielded an excellent fit to the data, and its only major problem was that it had one more free parameter than the model theory. Given that the original inference-guessing theory uses five parameters to infer selections, no simple way exists to fit the theory to just the four canonical selections, one of which it can only guess.

2. In our fittings of the two theories, they both have too many parameters for four canonical selections (i.e., they are "over-saturated"). The model theory has two sets of best-fitting parameter values (i.e., it is not "identifiable"), and the simplified inference-guessing theory's "range of

predictions completely covers the space of possible outcomes" *(i.e., it is not "testable").* Despite their criticism, K&K went ahead to fit the model theory to the results of each experiment (see point 5 below). And one of our aims was to use the pruned tree as a baseline for comparison with the model theory.

3. Our goodness-of-fit results reported in terms of root mean square errors are "simply impossible", and must favor the inference-guessing theory rather than, as we reported, the model theory. The results are in Table 6 of Ragni et al. (2018), and as its caption states, they are the output of the L-BFGS-B algorithm. So, they are not impossible, but the outcome of a standard procedure. The values of the root mean square errors are miniscule, but not zero, and do not favor the inference-guessing theory.

4. Our use of the Bayesian Information Criterion, BIC, to select between the two theories is invalid because the model theory has more than one set of optimal parameter values and the inference-guessing theory completely covers the possible outcomes for the canonical selections. Nonetheless, it is clear that the model theory is more parsimonious than the information-guessing theory, even with a guessing set of parameters added to it.

5. K&K made a new fit of the model theory to the canonical selections in each experiment separately, and "in the cases where fit is not perfect, [the model theory] always underestimates the probability of responses p and pq$\overline{q}$ and overestimates pq". We are grateful to K&K for showing that the model theory's algorithm may not yield the best three-parameter fit of the data. But perhaps they have corroborated that the partial insight pattern, pq$\overline{q}$, is a consequence of deliberation.

The Resolution of the Controversy

We agree with K&K on some matters, such as the need to winnow theories, and the importance of empirical results in refuting theories. One nub of the disagreement is how to treat guessing. It is one of the "defining characteristics" of the inference-guessing theory, which fits it using independent free parameters. The theory emphasizes the goodness of fit of theories to data. But, oddly, it does not allow that one card in a selection is inferred while another is guessed—a contingency that any theory of guessing in cognitive tasks should consider. By contrast, the model theory discounts any selection that occurs less often than chance because it is unlikely to illuminate the mental processes of selecting potential evidence. The theory instead aims to explain the remaining selections, even though some instances of them may result from guessing. K&K argue that to focus only on four canonical explanations is to exclude 26% of data from theoretical explanation, and so it is risky. But Klauer et al. (ibid. p. 691) consigned about 25% of their data to independent "guessing". The difference is that the model theory drops these selections from any further analysis: there is no need to try to explain noise whereas the inference-guessing theory does not try to explain them. It fits their frequencies with four free parameters. In the initial face-to-face studies of the task, only 10% of selections occurred less often than chance. So, a pertinent factor is whether or not an experiment was online, which yielded a greater percentage of such miscellaneous selections.

Three theoretical differences separate the two theories. First, like some other accounts, the inference-guessing theory bases thoughtful selections on inferences from a card and the hypothesis whereas the model theory uses the meaning of the hypothesis to search for examples and counterexamples. The reasons that participants gave for their selections favor the model theory. Second, the inference-guessing theory cannot predict that participants make an exhaustive search for counterexamples in the repeated selection task. And it cannot predict individuals' ability to explain what counts as a correct selection—it does not distinguish correct selections in any way. The model theory's use of the meanings of hypotheses predicts both phenomena. Third, the inference-guessing theory treats the selection of pq$\overline{q}$ as a result of guessing; the model theory treats it as a result of thought. The evidence corroborates the model theory: in one study, it was the participants' most frequent selection.

To model data, both theories use free parameters, which can take any value in order to yield an optimal fit. And free parameters are a telltale sign of theorists' ignorance. Imagine, say, that all and only those individuals with no training in logic interpret *if p then q* as implying its converse, and so as a result, they choose p and q according to both the inference-guessing theory and the model theory. We could then replace the corresponding free parameter in both theories with a decision based on evidence about an individual's training. In our collective ignorance, however, both theories suffer for using a free parameter instead. So, the number of free parameters in a model is an index of ignorance: theorists do not know which course of thought participants will take. With many free parameters, it becomes all too easy to fit erroneous theories to data. The inference-guessing theory fits the data well, and yet it is wrong. In the epigraph to our paper, the great polymath von Neumann embodied his skepticism about free parameters. With no theory at all, they can take values to fit an elephant. It was not empty boasting (see Mayer et al. 2010).

K&K write: "we are still far from a much-needed theoretical winnowing." But, if all the past experiments—with enough participants to populate a small town—still leave us with 16 theories of the selection task, then we are doing something wrong. We are engaged, not in science, but in a scholasticism impervious to empirical refutation. A warning sign, perhaps, is when experimental paradigms replace cognition as the target of theorizing. Theories of the abstract selection

task alone do not comprehend the testing of hypotheses. They need reformulations for new sorts of hypotheses, for the effects of feedback, for the reframing of hypotheses, and so on. In another half century, thanks to their "purposely vague" and "charitable" natures, they will have metastasized still further. Three dozen theories will not add up to a science of hypothesis testing.

Conclusions

Ragni et al. (2018) eliminated all but one alternative to the model theory. K&K's critique has led us to a refutation of that one alternative, the inference-guessing theory. So, after 50 years of research, do cognitive scientists at last understand the mental processes underlying the selection of potential evidence to test hypotheses? We hope so, and with one caveat, we believe so. The repeated selection task shows that naive individuals grasp what is at stake in testing a hypothesis. They first check for examples of the hypothesis, but once they have some, they switch to an exhaustive search for counterexamples—an exact corollary of the goals for induction (see Nicod 2000/1924, p. 219). The role of intuition in yielding examples and deliberation in yielding counterexamples is a theme of the model theory that runs through its accounts of many sorts of reasoning (e.g., Johnson-Laird et al. 2015). A typical abstract version of the selection task bamboozles people. They have not thought about testing an arbitrary hypothesis before, and they have only one chance to get it right. They rely on their intuitions, or else on rarer occasions, they opt out—they guess or balk at the task. Yet, depending on their cognitive ability and the contents or framing of the task, they may go beyond intuition to deliberate about possible counterexamples. For the caveat, we are indebted to K&K: the algorithm implementing the model theory may need correction. For K&K, the next step may be to show that an alternative multinomial processing tree yields a better fit. For the model theory, the next step may be a better algorithm. Any computable theory can be implemented in many different algorithms—a denumerable infinity of them, and so it is important not to confuse theory with algorithm (Johnson-Laird 1983, p. 7). The two research goals need not be at odds; they are compatible. Theories for precise fits of data could be better if they embodied precise accounts of mental processes.

Funding Information Open Access funding provided by Projekt DEAL.

References

Dyson, F. (2004). A meeting with Enrico Fermi. *Nature, 427*, 297–297.

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

Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1, 2nd ed.). New York: Wiley.

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

Hempel, C. G. (1945). Studies in the logic of confirmation, Parts I & II. *Mind, 54*(1–26), 97–121.

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

Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.

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

Johnson-Laird, P. N., & Wason, P. C. (1970b). Insight into a logical relation. *Quarterly Journal of Experimental Psychology, 22*, 49–61.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences, 19*, 201–214.

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.

Mayer, J., Khairy, K., & Howard, J. (2010). Drawing an elephant with four complex parameters. *American Journal of Physics, 78*, 648–649.

Nicod, J. (2000). *Foundations of geometry and induction*. London: Routledge (Originally published in 1924).

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses. *Psychological Bulletin, 144*, 779–796.

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

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273–281.

Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology, 60*, 471–480.

Wason, P. C., & Johnson-Laird, P. N. (1969). Proving a disjunctive rule. *Quarterly Journal of Experimental Psychology, 21*, 14–20.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning*. London: Batsford. Cambridge: Harvard University Press.