The Wason Selection Task: A Meta-Analysis

Marco Ragni (ragni@informatik.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany

Ilir Kola (kola@informatik.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg, 79110 Freiburg, Germany

P. N. Johnson-Laird (phil@princeton.edu)
Princeton University, Princeton NJ 08540 USA and New York University, New York, NY 10003, USA

Abstract

In Wason's selection task, participants select whichever of four cards could provide evidence about the truth or falsity of a conditional rule. As our meta-analysis of hundreds of experiments corroborates, participants tend to overlook one of the cards that could falsify the rule. 15 distinct theories aim to explain this phenomenon and others, but many of them presuppose that cards are selected independently of one another. We show that this assumption is false: Shannon's entropy for selections is reliably redundant in comparison with those of 10,000 simulated experiments using the same four individual probabilities for each real experiment. This result rules out those theories presupposing independent selections. Of the remaining theories, only two predict the frequencies of selections, one (due to Johnson-Laird & Wason, 1970a) provides a better fit to the experimental data than the other (due to Klauer et al., 2007). We discuss the implications of these results.

Keywords: Conditional reasoning; Entropy; Falsity, Selection task; Mental models.

Introduction

Human beings are able to evaluate whether assertions are true, and to select evidence relevant to such evaluations. The late Peter Wason (1968) carried out a paradigmatic study to test whether naive individuals grasped the relevance of falsification. In his original "selection" task, the experimenter explains to the participants that there is a pack of cards in which each card has a letter on one side and a number on the other side. Four cards are taken at random from the pack, and placed in front of the participant (see Fig. 1). The experimenter then presents the rule:

If there is a *D* on one side of a card, then there is a *3* on the other side.

The participants' task is to select just those cards that, if turned over, would show whether or not the rule is true or false of the four cards. The task is a demonstration, not an experiment, because it has no independent variable.

Participants tend to select the *D* card alone, or the *D* and *3* cards, but rarely the *D* and *7* cards. Yet, if the 7 has D on its other side, the rule would be false. This failure to falsify was shocking. Perhaps as a consequence more than 300 experiments investigating the task have been published over the last 50 years.

In order to try to understand performance, psychologists developed various versions of the task. They explored rules of different sorts, such as disjunctions and rules framed

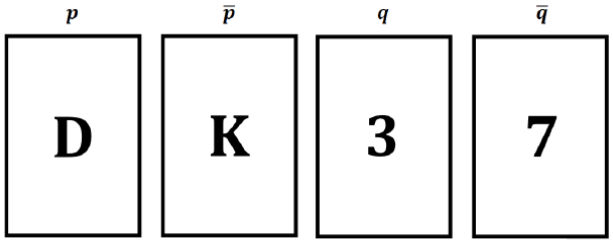

Figure 1. The four cards in Wason's selection task. Each has a number on one side and a letter on the other side. The participants' task is to select just those cards that, if turned over, would show whether or not the rule shown above holds for the four cards. The letters p, q, etc. are added for illustrative purposes as the rule is of the sort, *if p then q*.

with "every" in place of "if" (Wason & Johnson-Laird, 1969; Wason & Shapiro, 1971), cards with all the information on one side but partly masked, choices of just two cards (e.g., Johnson-Laird & Wason, 1970b), or choices of multiple cards, with repetitions of one or more cards (e.g., Oaksford & Chater, 1994). But, two main versions elicited better performance than *abstract* rules, such as the one in Fig. 1. One version used *everyday* rules, such as one about destinations and modes of transport (Wason & Shapiro, 1971). The other version switched the task around so that participants had to select those cards representing individuals who might be violating a *deontic* rule (e.g., Griggs & Cox, 1982), such as:

If a person is drinking beer, then the person must be over 19 years of age.

The efficacy of some deontic rules, such as one about the amount of postage on letters (Johnson-Laird, Legrenzi, & Legrenzi, 1972), depended on the participants' familiarity with them, but not all do so.

As the number of experimental studies grew, so too did the number of theories. By our reckoning, there are at least 15 distinct theories of the selection task including ones based on the meaning of conditionals, on formal rules of inference for them, on heuristics such as "matching" in which participants merely select those cards referred to in the rule (Evans, 1977), on content-specific rules of inference, and on the probabilities with which the various items on the cards occur in reality (Oaksford & Chater, 1994). Given that the selection task has been under investigation

for half a century, the existence of 15 theories about it is embarrassing for cognitive science. Our aim in what follows is therefore to describe meta-analyses of the experiments that aimed to eliminate as many theories as possible.

Meta-analyses

The reliability of the results

We searched the literature for experiments on the selection task with the proviso that they used a conditional rule of the sort: *if p then q*, and that they reported at least the frequencies of the four *canonical* selections of p, pq, pq$\bar{q}$, and p$\bar{q}$, which the early studies had reported. Henceforth, we abbreviate selections in the preceding way, stating which of the 4 cards they included, e.g., pq denotes a selection of the p and q cards (see Fig. 1). We divided the resulting experiments into three categories according to the nature of the rule they used: abstract, everyday, or deontic. We also classified them according to whether they reported the frequencies of only the 4 canonical selections and a category of "other" selections, or the frequencies of all 16 possible selections. The studies can be found at http://www.cc.uni-freiburg.de/data.

Because the first studies were carried out half century ago and subsequent ones in many countries, their results might be too heterogeneous for an informative test of the theories. We assessed the overall homogeneity of the results for the three categories of task from the reliability of the rank orders of the frequencies of their canonical selections. Table 1 reports Kendall's coefficient of concordance, W, which ranges from 0 for no consensus to 1 for perfect consensus, for the three categories of task. The results show a reasonable and robust consensus over the experiments. Table 2 presents the overall percentages of each of the four canonical selections for the three sorts of selection task. It shows why the deontic task yielded a greater concordance, W: the majority of participants selected cards denoting potential violations of the rule.

Table 1. The concordance across different experiments examining the three main sorts of selection tasks as assessed with Kendall's coefficient of concordance, W, and stating its χ^2 and p values.

Three sorts of selection task	Number of experiments	Kendall's	χ^2 and p value
Abstract	104	W = .34	107, p < .001
Everyday	44	W = .25	33, p < .001
Deontic	80	W= .54	29, p < .001

The redundancy of the selections

Many studies of the selection task report only the four separate probabilities with which participants selected each of the cards (e.g., Evans, 1977). These results, however, make sense only if the selection of each card is independent of the others. Some investigations have reported this independence (e.g., Evans, 1977). But, others have refuted it by establishing correlations between the selections (Pollard, 1985; Oaksford & Chater, 1994). Correlations, however, are only among pairs of cards in selections. A better assessment would take into account each selection as a whole

Table 2. The percentages of each of the four canonical selections for the three sorts of selection task

	The canonical selections			
	p	pq	pq$\bar{q}$	p$\bar{q}$
Abstract	36	39	5	19
Everyday	23	37	11	29
Deontic	13	19	4	64

and all the selections made in an experiment. We therefore introduced a new procedure that combines Shannon's measure of entropy (or informativeness) with the computer simulation of thousands of experiments. The underlying intuition is straightforward. Suppose the selections in an experiment are more redundant – more predictable – than a prediction made solely from the frequencies of selecting each of the four individual cards in the experiment. It follows that the cards in selections are, not independent of one another, but interdependent. And some aspect in the process of selecting cards yields the redundancy.

The first step in our procedure is to compute the amount of information in the selections in an experiment, i.e., the difficulty of predicting them. We use Shannon's measure of entropy:

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

for the set of selections, where P_i denotes the probability of the i-th selection, and $\log_2$ denotes a logarithm to the base 2. In general, the greater the number of different selections, and the more evenly distributed the frequencies over them, so the value of H increases, and it is harder to predict the selections. If participants chose each card independently of the others, the value of H for the experiment would not differ reliably from its value for selections derived from sampling according to the four probabilities for selecting each card. 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 reflected in a smaller value of H reflects interdependence in the selections.

As an illustrative example, consider the selections in Experiment 2 of Stahl et al. (2008), which we choose because of its large number of participants: 351. Here are the frequencies of the selections, in which 6 participants selected none of the cards:

p 92, pq 99, pq$\bar{q}$ 2, p$\bar{q}$ 20, p$\bar{p}$q$\bar{q}$ 19, $\bar{p}$q 6, p$\bar{p}$$\bar{q}$ 2,
$\bar{p}$q$\bar{q}$ 2, $\bar{q}$ 18, $\bar{p}$$\bar{q}$ 22, p$\bar{p}$ 7, q$\bar{q}$ 6, $\bar{p}$ 7, q 43, none 6.

They show that the probabilities of selecting each of the four cards were as follows:

p 0.69, q 0.49, $\bar{q}$ 0.26, $\bar{p}$ 0.19.

The value of H for the selections in the experiment is 2.8 bits. Could this value have occurred by chance? We used a resampling procedure to find out its chance probability (see, e.g., Good, 2001). We ran a computer program to carry out 10,000 simulated experiments based both on the

number of participants in the original study and on its probabilities above of selecting the four individual cards. The resulting mean value of H was 3.13, which shows that the observed selections in the experiment have a redundancy of 0.33. More important, however, is that not one of the simulated experiments yielded an entropy as low as 2.8 bits, and so the difference is statistically significant ($p < .0001$). The redundancy in the original experiment did not occur by chance. In summary, a statistically significant degree of redundancy in selections in an experiment is evidence for their interdependence.

We programmed an algorithm based on the same idea. Its key difference from our analysis of Stahl's data above is that it concerns only the four canonical selections. This constraint is necessary because so many experimental reports state the results only for them. Four selections have a maximum entropy of 2 bits if they are each equiprobable. The mean over the 228 experiments (in Table 1) is 1.27 bits (with a standard deviation of 0.48). The input to the program states the number of participants and the frequencies of the four selections for each experiment in the set. Its main steps are as follows. For each experiment:

1. Compute N, the number of participants, and the probabilities with which each of the 4 cards occurred in the experiment's selections.
2. Compute Shannon's entropy H for the experiment.
3. Carry out 10,000 simulated experiments based on the probabilities of selecting each card, assigning a selection to each of the N participants.
4. Return the number of simulated experiments with a higher entropy than the actual experiment and the number of them with the same or a lower entropy.

Table 3. The mean entropies (in bits) of 228 experiments on three sorts of selection task, the mean entropies of sets of 10,000 simulations of each experiment, and Wilcoxon's tests (W, and its p-value) of the difference between them.

The three sorts of selection task	Mean entropy of experiments	Mean entropy of sets of simulations	Wilcoxon's W and p-value
Abstract	1.32	1.42	$W = 469$, $p < .001$
Everyday	1.51	1.66	$W = 28$, $p < .001$
Deontic	1.06	1.21	$W = 68$, $p < .001$

Table 3 presents the mean entropies of the 228 experiments investigating the three sorts of selection task, the mean entropies of each of their 10,000 simulations, and the results of Wilcoxon's W test and its p-value comparing the pairs of means. These results allow us to reject the null hypothesis of independent selections. The redundancy shown in the smaller entropies of real experiments over simulated ones shows that the cards in selections are not selected independently of one another. They are selected in an interdependent way. This result eliminates any theory that predicts that selections are independent.

Theories of the selection task

Some theories of the selection task are informal and make only qualitative predictions about selections (e.g., Wason, 1968). Some predict only whether the correlations between selecting the possible pairs of cards are positive or negative (Oaksford & Chater, 1994). Some predict only the probabilities of selecting each of the four cards (Evans, 1977; Hattori, 2002; Oaksford & Wakefield, 2003). We discount all of these theories as insufficiently powerful to make quantitative predictions about the frequencies of the canonical selections, let alone all 16 possible selections. There remain just two theories, which we now outline.

The insight model

The first algorithms to model the mental processes underlying the selection task were due to Johnson-Laird and Wason (1970a). Their principal algorithm posits three levels of insight into the importance of falsification: no insight, which implies that reasoners select only cards referred to in the rule – an anticipation of "matching" bias (Evans, 1972); partial insight, which implies that reasoners consider all the cards, adding any further cards that verify the rule, or, failing that, that falsify the rule; and complete insight, which implies that reasoners select only cards that can falsify the rule. The algorithm was published as a flow chart, but not implemented, because of a lack of access to a main-frame computer. We recently programmed it, replacing its use of truth tables with mental models and fully explicit models, simplifying its processes, but keeping its original functionality so the program makes the identical predictions to the original version.

Given a rule of the sort *if p then q*, the program begins by compiling a list of cards to select, and its first step is to scan its mental model of the conditional, and as a result to put p on this list. If the program also scans the model in the opposite direction, it adds q to the list. With no insight into the task, these selections verify the rule. However, the program implements two interrelated levels of insight. Partial insight is to assess all the cards, and to add any further card that verifies the rule, or, if none does, to add any that falsifies the rule. So, if q is already in the list, partial insight adds $\bar{q}$, because it can falsify the rule, yielding the selection $pq\bar{q}$. Complete insight is to select only cards that can falsify the rule, and yields the selection $p\bar{q}$. Complete insight occurs only if all the cards are examined. An explicit biconditional as an input yields a selection of all four cards in certain cases, e.g., when it scans its model in both directions with partial insight.

Fig. 2 presents, not the algorithm, but a tree diagram summarizing its parameters and its predictions for conditionals and biconditionals. As it illustrates, the algorithm produces the same selections as a result of different processes, and it is not deterministic, i.e., nothing in the algorithm determines the level of insight (pace Evans, 1977, who took the algorithm to be deterministic). The predictions in Fig. 2 explain why selections should be interdependent, e.g., verifying cards include q only if they include p and falsifying cards include $\bar{q}$ if and only if they include

p. The only exceptions to the algorithm's outputs should be the result of guessing or haphazard errors. In fact, these exceptions occur at a rate less than chance in the 288 experiments.

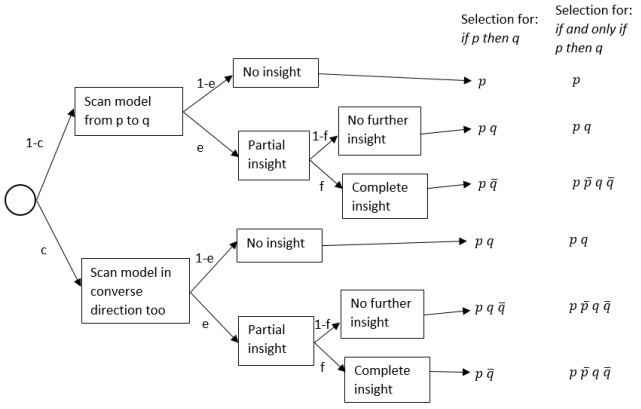

Figure 2. The predictions of the insight model (Johnson-Laird & Wason, 1970a) as a binary decision tree. Each decision is controlled in its recent implementation by a parameter (see text). Participants with no insight select only cards referred to in the rule. Those with partial insight consider all cards, selecting any further card that can verify the rule, or, failing that, that can falsify it. Participants with complete insight select only cards that can falsify the rule.

Our implementation of the algorithm contains three probabilistic parameters in the unit interval from 0 to 1. The first parameter, *c*, is the probability of scanning the model in both directions as opposed to scanning in only one direction. The second parameter, *e*, is the probability of examining all four cards, and if the result fails to add any card that verifies the rule, adding any card that falsifies it. This corresponds to partial insight. The third parameter, *f*, is the probability of complete insight, which makes only a falsifying selection.

The inference-guessing model

Klauer et al. (2007) proposed a set of related theories, including one with a heuristic component allowing for guessing, and an inferential component. There is no algorithm the implements the theory's underlying processes, but its predictions were modeled in a binary tree. This model has 10 parameters, which are each the probability that one sort of process occurs rather than another, and so each is in the unit interval from 0 to 1. The model's first parameter is the probability that the inference governs the selection as opposed to guessing. The guessing component makes independent selections of each of the four cards according to four parameters that are the respective probabilities of selecting each of them independently as a result of guessing or any heuristic factor such as "matching" (Evans, 1977). The theory assumes that selections are governed, not by the meaning of the rule, but by inferences from the rule. The particular inferences depends on five parameters:

1. The probability that the rule, *if p then q*, is interpreted as a conditional as opposed to a biconditional.

2. The probability that the inference is forwards from the *if*-clause: modus ponens (MP) or denial of the antecedent (DA), as opposed to backwards from the *then*-clause: modus tollens (MT) or affirmation of the consequent (AC).

3. Given the biconditional interpretation, the probability that the interpretation is bidirectional, *if p then q & if q then p*, as opposed to a case distinction, *if p then q & if not-p then not-q*. With the bidirectional interpretation, the distinction between forwards and backwards inferences does not apply – both are made, but with a case distinction interpretation, the distinction still applies.

4. The probability that an inference from a conditional or a bidirectional biconditional is a sufficient one as opposed to a necessary one. Normally, *p* is judged sufficient to infer *q* from *if p then q*, but sometimes *p* is judged necessary to infer *q,* as when the conditional is interpreted as stating an enabling condition akin to *only if p then q*. A forward sufficient inference is MP, whereas a forward necessary inference is DA; and a backward sufficient inference is AC, whereas a backward necessary inference is MT.

5. The probability that inferences are made only about the visible sides of cards as opposed to the invisible sides of cards too, i.e., individuals can envisage items on them.

The model contains 10 parameters but the data are the frequencies of the four canonical selections. Hence, to ensure that the process of fitting model to data converges and does not overfit the data, we implemented a *restricted* inference-guessing model that makes the four canonical selections. Fig. 3 summarizes the predictions of this restricted inference-guessing model. The reasoning component in the original model makes no more than two inferences on a trial, and so it cannot make the canonical selection of three cards: pq$\bar{q}$. We therefore changed the original guessing component to make this selection.

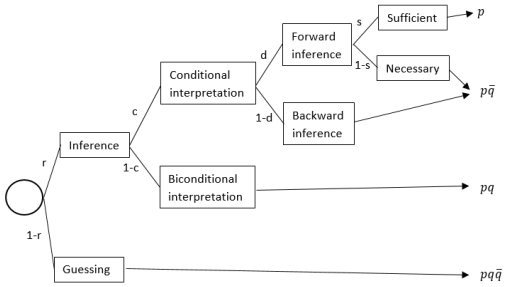

Figure 3. A restricted version of the binary decision tree of the inference-guessing model (Klauer et al., 2007) for the 4 canonical selections. Each decision is controlled by a parameter (see text).

The two models are based on the only theories that we could find in the literature that can be programmed with parameters that fit data about the frequencies of selections.

An evaluation of the two models

We evaluated the insight model with 3 parameters (Johnson-Laird & Wason, 1970a) and the restricted inference-guessing model with 4 parameters (cf. Klauer et al., 2007). Their respective predictions can be represented as trees of

binary decisions (see Fig. 2 and Fig. 3). Both models invoke alternative sequences of processes depending either on three decisions in the insight model or four decisions in the inference-guessing model. Because each model's predictions correspond to a tree of decisions, we evaluated each of them as a multinomial processing tree (MPT) in which the probability of a particular cognitive state is estimated from the observed frequencies of selections (Riefer & Batchelder, 1988). A program fitted each of the two models to the frequencies of the canonical selections of the three sorts of selection task: 104 experiments with the abstract task, 44 experiments with everyday task, and 80 experiments with deontic task (see Tables 1-3 above). We used the maximum-likelihood method from the R-package for multinomial processing trees (the MPTinR of Singmann & Kellen, 2012). We calculated three measures to compare the goodness of fits of the two theories:-

• The root mean square errors (RMSEs) of the fits.

• The Bayesian information criterion (BIC), which indicates how much information is lost when a model represents the process that generates the data, taking into account both its goodness of fit and its number of parameters. It penalizes models according to the number of their parameters, and the smaller its value, the better the fit between a model and the data.

• The Bayes factor (BF; Schwarz, 1978), which is a Bayesian method to compare different models. It uses an approximation of the difference between the BIC value of model 1 and BIC value of model 2 as computed by MPTinR. The higher its value between 30 and 100, the stronger the support for model 1 over model 2 (Wagenmakers et al., 2011).

Table 4 presents the three measures for each of the two models. As it shows, the insight model with three parameters has a closer fits, and lower BIC values, than the restricted inference-guessing model. The Bayesian factor likewise shows stronger evidence for the insight model than for the restricted inference-guessing model. The insight model has the advantage of fewer parameters. As a theory, it is simpler because it relies on the meaning of the rule rather than inferences from it, and because it has no machinery to account for selections that occur at a rate less than chance. But, it is not a paragon, and we explain why below.

General Discussion

Half a century of research and over 300 articles should have led to a single unique theory of a cognitive task rather than to 15 different theories. That was the situation for Wason's selection task. The present research, however, has eliminated all but one theory. And it did so using the following strategy. It established a large but representative set of experiments investigating rules of the sort *if p then q* that had a reliable concordance in their results (Table 1). These results established the rarity of falsifying selections, $p\bar{q}$, except when they violate a deontic rule Table 2). The four canonical selections (p, pq, $pq\bar{q}$, and $p\bar{q}$) are reliably redundant in most experiments in comparisons of each experiment's entropy (informativeness) with the entropy of its

Table 4. The insight model's and the restricted inference-guessing model's goodness of fit with the individual canonical selections for 288 experiments overall and for the three sorts of selection task: the root mean square errors (RMSE) for their predictions, their Bayesian information criteria (BIC), and the Bayes factors for the better-fitting model.

The 3 sorts of the selection task	Cognitive model	RMSE	Bayesian Information Criterion (BIC)	Bayes factor
Overall	Insight	2.69	27.7	
	Inference-guessing	19.35	37.0	99.5
Abstract	Insight	1.97	25.7	
	Inference-guessing	3.28	34.3	73.7
Everyday	Insight	1.7	23.2	
	Inference-guessing	2.18	30.9	47
Deontic	Insight	0.8	23.5	
	Inference-guessing	1.05	31.4	49.4

10,000 simulations based on its four probabilities of selecting each card (Table 3). Not all experiments yield redundant selections, but the vast majority do. This result ruled out theories that imply that selections of cards are independent of one another. Above all, theories therefore need to predict the frequencies of the canonical selections. Perhaps surprisingly, this criterion rules out nearly all the remaining theories. Klauer et al. (2007) had programmed an MPT of their inference-guessing model using 10 parameters to make predictions for the frequencies of all 16 possible selections – most of which do not occur more often than chance. More than twice as many experiments reported the frequencies only of the four canonical selections than reported them for all 16 selections. Hence, we produced an MPT for a restricted version of the model that used four parameters to predict the frequencies of the canonical selections. To do so, we reduced the original parameters for guessing to one, which made a selection of three cards, otherwise impossible for the model to select. For the insight model, we programmed an algorithm that carried out its processes (Johnson-Laird & Wason, 1970a), and we used it to construct an MPT model with three parameters. The insight model yielded a better fit with fewer parameters (Table 4).

The story of the selection task does not end here. But, the success of the insight theory tells us that we have returned to how it was conceived after only a handful of studies. Naive individuals focus on those cards mentioned in the rule, and select them if they can verify the rule. With a little bit of insight, they consider all the cards, and may select additional cards. With complete insight, they select only cards that can falsify the rule (Johnson-Laird & Wason,

1970a). We now know that various factors – the competence of participants, the contents of the rule, and the framing of the task – can all enhance insight. An account along these lines seems to be correct, except perhaps when experiments implicate probabilities in their contents or framing (e.g., Oaksford & Chater, 1994).

The excellent fit of the insight model must be viewed with caution. The number of parameters in a model is a measure of our ignorance. Those for guessing seem to be dispensable. Indeed, some selections are very odd, as we saw earlier in our analysis of the results from Stahl et al. (2008). They are so odd that they must count as irrational on any criterion: the participants erred or guessed. Introducing parameters to model guessing has no theoretical value other than to index the difficulty of a task. The insight theory has three essential parameters, and the original inference-guessing model has five. The difference reflects an crucial distinction: whether people determine the truth value of an assertion based on its meaning (the insight model) or based on inferences from it (the inference-guessing model). Therein may lie the advantage of the insight model. But, we are bound to ask what mechanisms might replace its parameters. We now know that the insight to make falsifying selections depends on various factors, including intellectual ability (e.g., Stanovich & West, 1998). Hence, it may be feasible to replace the parameter for the probability of complete insight with a measure of ability. It is even conceivable that the parameter of partial insight might reflect a lesser but above average intellect. The parameter for scanning a model of the conditional in both directions is more problematic. It may depend on the processing capacity of working memory. These speculations in no way rule out the possibility of some quite different theory of the selection task outperforming the insight model.

If our research has any general moral, it is an old one: cognitive theories should be effective procedures (Johnson-Laird, 1983, p. 6). They should be programmable.

Acknowledgements

This research was supported by a DFG-Heisenberg fellowship DFG RA 1934/3-1, and DFG projects RA 1934/2-1 and 4-1. We thank Linden Ball, Ruth Byrne, Nick Chater, Jonathan Evans, Keith Holyoak, Sangeet Khemlani, Ken Manktelow, Mike Oaksford, Klaus Oberauer, Lance Rips, Carlos Santamaría, Walter Schaeken, Walter Schroyens, Dan Sperber, Christoph Stahl, and Valerie Thompson, for their help. We thank Christoph Klauer for his advice and sending us his algorithm,. We dedicate this paper to the memory of Peter Wason (1924-2003), a unique and most extraordinary researcher.

References

Evans, J. S. B. (1972). Interpretation and matching bias in a reasoning task. *The Quarterly Journal of Experimental Psychology*, *24*, 193-199.

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

Good, P. I. (2001). *Resampling methods: A practical guide to data analysis*. NY: Birkhauser Boston.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in Wason's selection task. *British Journal of Psychology,73(3),* 407-420.

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.

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

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. S. (1972). Reasoning and a sense of reality. *British Journal of Psychology*, 63, 395-400.

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*(1), 49-61.

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.

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

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition*, *31*(1), 143-154.

Pollard, P. (1985). Nonindependence of selections on the Wason selection task. *Bulletin of the Psychonomic Society*, *23*(4), 317-320.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318-339.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Singmann, H., & Kellen, D. (2012). MPTinR: Analysis of multinomial processing tree models in R. *Behavioral Research Methods, 45,* 560–575.

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.

Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking & Reasoning*, *4*, 193-230.

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,* 3, 426–432

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

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

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*, 63-71.