Bayesians Too Should Follow Wason

A Comprehensive Accuracy-Based Analysis of the Selection Task

Filippo VINDROLA and Vincenzo CRUPI

Abstract

Wason's selection task is a paramount experimental problem in the study of human reasoning, often connected with the celebrated ravens paradox in the philosophical literature. Various normative accounts of the selection task rely on a Bayesian approach. Some claim vindication of participants' rationality. Others don't, thus following Wason's original intuition that observed responses are mistaken. In this paper we argue that, despite claims to the contrary, all these accounts actually speak to the same effect: Wason was right. First, we provide a new accuracy-based analysis of the selection task, that includes the existing proposals as special cases. We then show on this basis that none can actually vindicate participants' rationality. We conclude that all normative renditions considered eventually concur: all in all, Bayesians should follow Wason in the selection task.

1. Introduction

No experimental paradigm has generated more psychological research on rationality than the Wason selection task (Wason [1966], [1968]). Content with Wason's original interpretation, many psychologists and philosophers have thought of the selection task as a textbook example of how humans can systematically fall short of compelling norms of reasoning. Others have protested, however, providing a number of arguments to the effect that people's behaviour in the task is actually rational, given alternative and allegedly appropriate normative accounts.[1] The result: more than 50 years after Wason's original experiment, we are left with a plurality of different normative analyses of the task. Most of them are explicitly Bayesian, all are implicitly based on various auxiliary assumptions and theoretical choices. Some claim vindication of participants' rationality, others don't, and no consensus is in sight.

In this paper we argue that and explain why, despite prevailing views, all these accounts actually speak to the same effect: Wason's original intuition was correct. First, we provide a novel accuracy-based framework for the selection task that includes the existing proposals (including Wason's) as special cases. We then show on this basis that none of these proposals can vindicate participants' rationality in the task without relying on highly debatable auxiliary assumptions that are quite independent of the Bayesian framework. We conclude that all normative renditions considered converge: Bayesians should follow Wason in the selection task.

Here is an outline of the structure of our argument. First, we give a preliminary review of the main accounts that have been proposed to assess participants' rationality in the task (section 2). Since these proposals reflect different approaches and theoretical choices, our second step is to make them

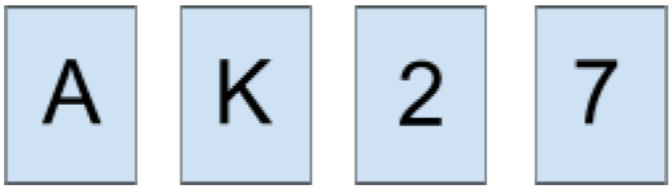
[1] For occurrences in the philosophical literature, see, e.g. (Stich [1990], pp. 4-6, Stein [1996], pp. 79-93, Humberstone [1994], and Bradley [2015], pp. 118-119).

comparable, providing a unified model of the task (section 3). Third, we present our unified normative framework based on the notion of the minimization of epistemic inaccuracy (section 4). In the fourth and last step we discuss the implications (sections 5-7).

We provide three main reasons why our analysis improves on current knowledge. One is that, in the light of our analysis, the existing accounts can be recovered as specifications of a unified view of rational inquiry as aiming at inaccuracy reduction. A further reason is that the resulting framework refines the existing accounts with crucial amendments and integrations – in particular, it provides additional justificatory grounds for their conclusions. The third key point of our contribution is the outcome of our analysis. Despite appearances to the contrary, all normative approaches considered actually concur: prevalent responses in the selection task are best seen as a systematic departure from compelling normative benchmarks of rational thinking.

2. The Persistent Puzzle of the Wason Task

First things first: let us review key episodes in the long history of the normative renditions of the Wason task. The task is as follows.[2] Participants are presented with four cards (see example below), and they are told that each card has a letter on one side and a number on the other.



Participants are asked to say which cards they would turn over in order to find out whether the following conditional statement is true: 'if a card has a vowel on one side, then it has an even number on the other side'. In Wason's original experiments, almost all subjects selected the first card (A), a majority also

[2] We are only concerned with so-called abstract version of the task here.

selected the third card (2), only a few selected the fourth one (7), and almost no one selected the second one (K).

Wason considered these results a clear indication of biased reasoning. This is because most participants tended to select a card that is apparently useless for discovering the truth or falsity of the conditional statement (the third one), and they failed to select a card that is equally accessible and useful to that effect (the fourth one). According to Wason's original story, turning the first and the fourth card is useful because, logically, these cards can potentially falsify the hypothesis at issue (by possibly revealing an odd number and a vowel, respectively, on the other side), whereas the other two cards cannot provide any refuting evidence for that hypothesis. Rational agents should then select the first and the last card, while the other two are useless: 'The correct response is to choose cards displaying vowels and cards displaying numbers which are not even, i.e., odd numbers, since only this combination of letters and numbers on the same card would prove the statement false' (Wason [1966], p. 146). Given this 'quasi-Popperian' analysis, the conclusion seems straightforward: since most participants selected the third card and did not select the fourth card, they behaved irrationally. However, when it comes to explaining *why* this conclusion has normative force, and more generally how the quasi-Popperian story really works, Wason's informal remarks won't help much (more on this below).

More than 20 years after Wason's original experiment, Oaksford and Chater ([1994]) did provide a detailed formal story. But their story subverted Wason's conclusions. According to Oaksford and Chater, participants' performance actually complies with compelling normative principles – simply, these principles are not the ones Wason was relying on: '[T]he psychological data […] has appeared to show human reasoning performance to be hopelessly flawed, [but] *when appropriate rational theories are applied*, reasoning performance may, on the contrary, be rational' (Oaksford and Chater [2007], p. 31, emphasis added). Oaksford and Chater pointed out that participants in the task were asked for a judgment of epistemic utility: which

cards is most useful to turn over in order to discover whether the conditional statement is correct? And the right way to address a question like this — they submitted — is to assess to what extent turning each card can be expected to reduce subjects' uncertainty about the truth of the conditional statement. According to Oaksford and Chater, the expected reduction of uncertainty for one card amounts to the weighted average of the difference between prior and posterior Shannon entropy for each possible outcome of turning that card. Importantly, Oaksford and Chater's analysis implied that such expected uncertainty reduction is higher for choosing the third rather than the fourth card. If this is correct then, against the traditional view, the apparently useless selection of the third card is actually more rational than the choice of the fourth card, and the participants' behaviour seems to be vindicated.

Sixteen years after Oaksford and Chater's key contribution we have the last main episode of our (condensed, yet winding enough) story. Fitelson and Hawthorne ([2010]) thoroughly investigated the connections between Wason's selection task and Hempel's paradox of the ravens (but see Humberstone 1994 for an important precedent). Unlike Oaksford and Chater, Fitelson and Hawthorne quantify the epistemic utility of turning a certain card in the task as given by the 'expected confirmational power' of an evidence search option (e.g., turning over a specific card), thereby extending an earlier probabilistic analysis by Nickerson ([1996]) to a more general assessment. However, while Nickerson's analysis had concluded that 'people's typical performance in the selection task can be explained by consideration of what constitutes an effective strategy for seeking evidence' ([1996], p. 1), Fitelson and Hawthorne refrained from reassuring conclusions: "it is more difficult to *rationalize* the behaviour / performance of actual subjects on the Wason selection task than one might have thought" ([2010], p. 235, original emphasis).

We thus have three major normative approaches to the Wason selection task. One (Wason's original story) ascribed irrationality to the participants, but its justificatory basis was largely informal and incomplete. Another one (Oaksford and Chater's) relies on a sophisticated probabilistic machinery and

an influential piece of formalism from information theory to allegedly vindicate participants' performance as optimal decision making. The last one (Nickerson / Fitelson and Hawthorne) draws on classical philosophical concerns in confirmation theory, but apparently delivers diverging implications about human rationality.

	interpretation of the conditional		*performance measure addressed*		*normative approach adopted*		
	material	probabilistic	response frequencies	selection propensities	quasi-falsificationism	expected entropy reduction	expected confirmational power
Wason	✓	✗	✓	✗	✓	✗	✗
Oaksford and Chater	✗	✓	✗	✓	✗	✓	✗
Fitelson and Hawthorne	✓	✗	✗	✓	✗	✗	✓

Table 1. A variety of theoretical choices in major normative analysis of Wason's selection task.

What these diverse proposals have in common is that they all had to make a number of theoretical choices and background assumptions along the way to establish their conclusions. Table 1 provides a schematic preview, to be further discussed later on.

As Table 1 shows, the theoretical disunity in the normative accounts of the task is quite striking. Are then participants violating compelling normative prescriptions or not? Without a unified framework, it's hard to assess the impact that each assumption has on the outcomes of a given normative analysis. This is especially true given that the parties in play did not always provide explicit independent motivations for their choices. As a result, it's largely unclear what determines the relevant conclusions in each case: is it

really the application of diverse normative models or rather some of the auxiliary assumptions made?

To answer this question, in the following we'll first provide a unified model of the task (section 3) and a unified normative framework (section 4), allowing for a unified assessment of the previous accounts mentioned from a novel perspective (sections 5-7).

3. Modelling the Task

In this section we provide a unified and comprehensive model of the Wason task. Participants in the task are shown four cards, which we'll call c_1, c_2, c_3, c_4 respectively (from left to right as displayed above). Since each card can be turned over to gain new evidence about the truth of the conditional statement, there are exactly four elementary options available to assess whether such statement is true, as explicitly requested in the task. Each elementary search option is represented by the possible outcomes of turning the single card at issue. We thus denote these search options with upper case *C*s, and treat them as binary variables, positing $C_1 = \{even(c_1), \sim even(c_1)\}$; $C_2 = \{even(c_2), \sim even(c_2)\}$; $C_3 = \{vowel(c_3), \sim vowel(c_3)\}$; $C_4 = \{vowel(c_4), \sim vowel(c_4)\}$. Combinations of elementary options are search options too, e.g., $C_1 \times C_2 = \{even(c_1) \wedge even(c_2), even(c_1) \wedge \sim even(c_2), \sim even(c_1) \wedge even(c_2), \sim even(c_1) \wedge \sim even(c_2)\}$, and so on (a combination of n distinct elementary options is modeled as a variable with 2^n values denoting mutually inconsistent and jointly exhaustive possibilities). Since participants are indeed allowed to turn multiple cards, the overall set of response options, call it R, includes all elementary search options and any combination of them (sixteen options in total). So, for instance: $C_1 \in R$, $C_1 \times C_3 \in R$, and $C_1 \times C_2 \times C_3 \times C_4 \in R$.

We want a generalized way to model the selection task to accommodate a variety of assumptions made in the literature. We thus have to choose how to best characterize the epistemic state of a rational agent addressing the task, represented by a probability distribution P. To this aim, we treat the four cards as randomly and independently sampled from a background population (a

large deck). Two possible statistical compositions of the large deck will be considered, and denoted as *d* and ~*d*. As they are assumed to specify the proportions of four kinds of objects (cards with a vowel vs. consonant and an even vs. odd number), the content of each of *d* and ~*d* can be determined by three parameters. We'll call such parameters α, β, ε, and α^*, β^*, ε^*, respectively.

- α and α^* are the probabilities of a card c_i having a vowel on one side given *d* and given ~*d*, respectively. So $\alpha = P[vowel(c_i)|d]$ and $\alpha^* = P[vowel(c_i)|\sim d]$. We also assume $0 < \alpha, \alpha^* < 1$.
- β and β^* are the probabilities of a card c_i having an even number on one side given *d* and given ~*d*, respectively. So $\beta = P[even(c_i)|d]$ and $\beta^* = P[even(c_i)|\sim d]$. We also assume $0 < \beta, \beta^* < 1$.
- ε and ε^* are the probabilities of a card c_i having a non-even (odd) number on one side given a vowel on the other side and given *d* or given ~*d*, respectively. So $\varepsilon = P[\sim even(c_i) \mid vowel(c_i) \wedge d]$ and $\varepsilon^* = P[\sim even(c_i) \mid vowel(c_i) \wedge \sim d]$. We also assume $0 \leq \varepsilon < 1$ and $0 < \varepsilon^* < 1$.

To start making some intuitive sense of this setting, let us first briefly comment on ε and ε^*. If $\varepsilon = 0$, then *d* implies that the universally quantified material conditional $\forall x[vowel(x) \supset even(x)]$ is true in the larger deck from which the four cards are meant to have been drawn. And given $\varepsilon^* > 0$, ~*d* implies that the same quantified material conditional is false. In fact, it is useful to take $D = \{d,\sim d\}$ as another relevant binary variable in our model, because then (for $\varepsilon = 0 < \varepsilon^*$) *d* can represent "if a card has a vowel on one side, then it has an even number on the other side" as referred to the whole population (the whole deck of cards), and ~*d* its plain logical negation. Following a typical assumption of earlier Bayesian analyses of the Wason task, we also posit a flat prior distribution on *D*, i.e., $P(d) = P(\sim d) = 0.5$. As a consequence of this and the random independent sampling assumption about the four cards, a full probability distribution over $C_1 \times C_2 \times C_3 \times C_4 \times D$ can be determined through our six parameters, as illustrated by Table 2 and Appendix 3 [as in the

corresponding models from the literature, probabilistic coherence is enforced positing $\alpha(1-\varepsilon) \leq \beta \leq 1-\alpha\varepsilon$ and $\alpha^*(1-\varepsilon^*) \leq \beta^* \leq 1-\alpha^*\varepsilon^*$].[3]

As a final piece of formalism, we need to label the conjunction (*vowel*(c_1) ⊃ *even*(c_1)) ∧ … ∧ (*vowel*(c_4) ⊃ *even*(c_4)) or, more concisely, $\bigwedge_{1\leq i\leq 4}[vowel(c_i) \supset even(c_i)]$. *This* statement, call it *f*, represents 'if a card has a vowel on one side, then it has an even number on the other side' as referred to the sample (of the four cards available). Of course, *f* is a straightforward consequence of *d* provided that $\varepsilon = 0$ (intuitively meaning that, there are 'no exceptions' in the whole deck). Also note that all probabilities involving *f* and its negation, thus propositional variable $F = \{f, \sim f\}$, are also fully determined by α, α^*, β, β^*, ε, and ε^* (given the other background assumptions we made).

given *d*

	even(c_i)	~*even*(c_i)	
vowel(c_i)	$\alpha(1-\varepsilon)$	$\alpha\varepsilon$	α
~*vowel*(c_i)	$\beta - \alpha(1-\varepsilon)$	$(1-\beta) - \alpha\varepsilon$	$1-\alpha$
	β	$1-\beta$	

given ~*d*

	even(c_i)	~*even*(c_i)	
vowel(c_i)	$\alpha^*(1-\varepsilon^*)$	$\alpha^*\varepsilon^*$	α^*
~*vowel*(c_i)	$\beta^* - \alpha^*(1-\varepsilon^*)$	$(1-\beta^*) - \alpha^*\varepsilon^*$	$1-\alpha^*$
	β^*	$1-\beta^*$	

Table 2. Probability distribution concerning a given card c_i as determined by the parameters α, α^*, β, β^*, ε and ε^*.

Let us briefly comment on the motivations for our modelling framework. An important point for the analysis of the Wason task is: what is the actual

[3] More precisely, one should say that the probability distribution *P* over $C_1 \times C_2 \times C_3 \times C_4 \times D$ arises by taking the ur-prior distribution determined through Table 2, and then conditionalizing on the evidence that is already given in the experimental scenario through the visible sides of the cards, namely (in our notation), *vowel*(c_1) ∧ ~*vowel*(c_2), ∧ *even*(c_3) ∧ ~*even*(c_4). This elucidation is technically appropriate but immaterial for our purposes. (See Appendix 3 for two specific examples.) For the probabilistic coherence clauses – $\alpha(1-\varepsilon) \leq \beta \leq 1-\alpha\varepsilon$ and $\alpha^*(1-\varepsilon^*) \leq \beta^* \leq 1-\alpha^*\varepsilon^*$ – note that, given the background model, these are necessary and sufficient for all cells in Table 2 to embed values in [0,1].

epistemic target, namely, the partition of hypotheses with regards to which the usefulness of the available information search options should be rationally assessed? If we label such target as T, the choice is between positing $T = D$ or $T = F$. Bayesians have almost exclusively discussed the former case, while Wason clearly had in mind the second. As we will see, both can be explicitly included (and compared) in our treatment – something that no earlier analysis pursued, to the best of our knowledge. The choice to posit $T = D$ enables the parallelism between the Wason task and the ravens paradox (see Humberstone [1994]), a move adopted by Fitelson and Hawthorne ([2010]), with the additional assumptions that $\varepsilon = 0$ and $\varepsilon^* > 0$, as expressed in our formalism. $T = D$ is also a key assumption in (Oaksford and Chater [1994], [2003]), but their interpretation of d is that the probability of a card c_i having an even number of one side given that it has a vowel on the other side is high, allowing for ε to be small but positive ($0 \leq \varepsilon \leq 0.1$). In this vein, Oaksford and Chater have construed their foil hypothesis, $\sim d$, as a probabilistic independence claim concerning $vowel(c_i)$ and $even(c_i)$, which implies $\beta^* = 1 - \varepsilon^*$ in our framework (check Table 2 on the right).

Oaksford and Chater have also found it natural to posit two further independence constraints, to wit, $P[vowel(c_i)|d] = P[vowel(c_i)|\sim d]$ and $P[even(c_i)|d] = P[even(c_i)|\sim d]$, so that $\alpha = \alpha^*$, and $\beta = \beta^*$ (compare Table 2 above and Oaksford and Chater [2003], p. 291). That's why Oaksford and Chater's model ends up requiring only three parameters: α, β, and ε, in our notation. Nickerson's ([1996]) model, in turn, retains the equalities $\alpha = \alpha^*$, and $\beta = \beta^*$ (that Fitelson and Hawthorne [2010] challenge, see pp. 220-223, and also Vranas [2004]), but agrees with Fitelson and Hawthorne in that (unlike Oaksford and Chater) one has $\varepsilon = 0$ and ε^* is independently set (see Nickerson [1996], p. 16).

The integrated representation provided here will be of much help in our later discussion. But before getting there, we still have to articulate a similar unifying move concerning the normative foundations of an analysis of the selection task.

4. A Principled Normative Framework

In this section we provide a general normative framework for the task. The key idea is simple and the relevant technical machinery is well understood in statistics and decision theory (see, e.g., Savage [1971], Dawid [1998], and Gneiting and Raftery [2007]). According to our proposal, epistemic utility is analyzed in terms of accuracy — roughly: closeness of probabilistic credences to actual truth-value assignments. In turn, our normative basis for an analysis of the selection task will be a measure of inaccuracy (to be minimized), to wit, a scoring rule.[4] Let's see in more detail how this works.

Given a probability distribution P defined over $C_1 \times C_2 \times C_3 \times C_4 \times D$ and a specific scoring rule as a measure of epistemic inaccuracy, each element in the set of the response options R can be assessed by the expected reduction of inaccuracy that it yields concerning D or F. In our approach, such expected reduction in inaccuracy will determine how much a given option is epistemically useful for an agent whose aim is to find out the truth about D or F.

For our current purposes, a scoring rule is a function $s : \{H \times \boldsymbol{P}\} \rightarrow \Re$, where H is a finite partition of hypotheses,[5] $H = \{h_1, \ldots, h_n\}$, and $\boldsymbol{P}$ the set of possible probability distributions over H, representing possible epistemic states of an agent. Then, $s(h_i,P)$ (with $h_i \in H$ and $P \in \boldsymbol{P}$) will be a measure of the (actual) inaccuracy of P with respect to H assuming the h_i is true. As a rule, of

[4] Following Joyce's ([1998], [2009]) seminal work, an extensive literature has developed in formal epistemology where scoring rules are investigated (e.g., D'Agostino and Sinigaglia [2010], Dunn [2019], Fallis and Lewis [2016], Leitgeb and Pettigrew [2010], Pettigrew [2013], Predd *et al*. [2009], Shoenfield [2017]). Our discussion is of course closely connected to this strand of research by the reference to the key notion of accuracy. Notice, however, that while we do take probabilism (and conditionalization) as normatively compelling (much as Oaksford and Chater, Nickerson, and Fitelson and Hawthorne), our argument is not committed to the prospects of the specific project of motivating probabilism itself (and conditionalization) through so-called accuracy-based approach. As a consequence, criticism of the latter, however effective, does not generally apply to the former (see, for instance, Carr [2017], Greaves [2013], Konek and Levinstein [2019], Oddie [2019]).

[5] This means that one assumes $h_1 \vee \ldots \vee h_n$ and also $\sim(h_j \wedge h_k)$ for each $j \neq k$.

course, an epistemic agent will not initially have access to the truth in *H*. However, the expected inaccuracy of a given probability distribution *Q* over *H* can be assessed relative to a distribution *P* over *H* that provides the expectation weights, as follows:

$$S(P, Q) = \sum_{h_i \in H} P(h_i) \cdot s(h_i, Q)$$

A scoring rule *s* is said to be *proper* if $S(P,P) \leq S(P,Q)$ for all *P*,*Q*. For a strictly proper score, moreover, it holds that, if $S(P,P) = S(P,Q)$, then $P = Q$. That is, a score *s* will be strictly proper if and only if it is proper and any distribution *Q* other than *P* has an expected score given *P* that is strictly higher (indicating more inaccuracy) than *P* itself. There is wide consensus that rational agents measure epistemic inaccuracy through strictly proper scoring rules (see Campbell-Moore and Levinstein [2020] for a recent discussion).

How does all this relate to the assessment of an information search option such as, say, $C_1 \times C_3$ in the Wason selection task? To address this point in general, we have to consider a partition of hypotheses $H = \{h_1, \ldots, h_n\}$, an evidence partition $E = \{e_1, \ldots, e_m\}$, their combination $H \times E = \{h_1 \wedge e_1, h_1 \wedge e_2, \ldots, h_n \wedge e_{m-1}, h_n \wedge e_m\}$, and a probability distribution *P* on $H \times E$ such that: $P(h_i) > 0$ for any *i*; the conditional probability of h_i given e_j, $P_{e_j}(h_i)$, is defined for each *i* and *j*; and *P* represents the epistemic state of an agent. Then, given distribution *P*, a piece of evidence *e* can itself be assigned (indirectly, as it were) a certain amount of epistemic utility to the extent that it decreases the agent's expected inaccuracy with respect to the target hypothesis partition *H*. Such epistemic utility, denoted $u(H, e)$, will thus correspond to the extent to which the expected inaccuracy given *e* is lower than the expected inaccuracy of the initial probability distribution, *P*, on the basis of the updated distribution, P_e, where the new information *e* is taken into account (see Roche and Shogenji [2018] for a neat recent discussion of this idea), namely:

$$u(H, e) = S(P_e, P) - S(P_e, P_e)$$

This allows us to define, eventually, the epistemic utility of a *test E* with respect to *H*, which is simply the expected utility of its possible outcomes:

$$U(H,E) = \sum_{e_j \in E} P(e_j) \cdot u\big(H, e_j\big)$$

It is important to emphasize that a measure $U(H,E)$ here is not simply motivated as a matter of convenience, popularity, or intuitive appeal. It arises in a principled way from exactly three antecedent assumptions: (i) that the key epistemic utility is accuracy; (ii) that inaccuracy is measured by a (proper) scoring rule; and (iii) that an improvement in accuracy (decrease in inaccuracy) after updating on evidence *e* is appropriately assessed on the basis of the posterior (and more informed) distribution, P_e.

This fundamental approach still leaves room for the choice of the scoring rule(s) to be employed as a basic building block in our setting (see, e.g., Douven [2020] for a recent discussion), but major specifications can be recovered as special cases of the comprehensive parametric family of the Tsallis scores:[6]

$$s_\tau(h_i, P) = \tau ln_\tau\left(\frac{1}{p_i}\right) - \left(1 - \sum_{h_j \in H} p_j^\tau\right)$$

where $\tau \geq 0$. The function ln_τ is a generalized version of the natural logarithm found in Tsallis's (1988) work: $ln_\tau(x) = \frac{x^{(1-\tau)}-1}{1-\tau}$. The ordinary logarithm is recovered in the limit for $\tau \to 1$, so that one can safely equate $ln_\tau(x) = ln(x)$ for $\tau = 1$, and make the parametric family s_τ continuous in τ (see Appendix 1).

Not only are Tsallis scores proper (in fact strictly proper as long as $\tau > 0$). Most importantly for our purposes, they also provide exactly the unified

[6] Tsallis's name is mostly associated with a parametric family of *entropies* (see Tsallis [2011], and Crupi *et al*. [2018]). What we here call Tsallis *scores* can be derived working back towards s_τ so that Tsallis entropies amount to $S_\tau(P,P)$. Savage ([1971]) and Dawid ([1998]) spell out the details of this connection.

normative framework that we look for. As we will see in the next sections, with $\tau = 0$ we obtain a well-behaved variant of Wason's original quasi-Popperian approach to the selection task (Section 5); with $\tau = 1$ we recover Oaksford and Chater's formal machinery, now equipped with a thorough motivation (Section 6); and with $\tau = 2$ we achieve a similar result with respect to Nickerson and Fitelson and Hawthorne (Section 7).

In principle, one could provide a motivation for a particular choice of τ rather than others. For example, as concerns the search for evidence, values of τ close to 0 represent the attitude of an agent who is especially eager to prune down the list of the elements in *H*, whereas for very high values of τ the agent is narrowly focused on the prospects of getting to near-certainty about the true item in *H*, and largely insensitive to anything else (see Crupi *et al*. [2018] for a related discussion). But we are not committed to such choice here: what matters for the present purposes is that prominent options yield the same result, as we will see later on.

This approach has other advantages too. Intuitive desiderata may be valuable resources in support of specific normative choices, and indeed both Wason and Oaksford and Chater have appealed to intuitive considerations in support of their conclusions (see Sections 5-6, below). But regardless of whether one thinks that intuitive appeal is enough to justify particular normative choices, our approach also has a more thorough motivation. Unlike previous normative renditions of the Wason task, our proposal explicitly embeds the idea that rational inquiry has a specific epistemic goal: reducing inaccuracy. This implies, for instance, that authors who have challenged bits of Bayesian epistemology as 'means with no end' (Brössel and Huber [2014]) should find our line of thought particularly appealing here (see also Schurz [2011], [2015]).

5. Wason Vindicated

We are back to the first episode of our story: Wason's original account. Wason unequivocally interpreted 'if a *c*ard has a vowel on one side, then it has an even number on the other side' as a material conditional, thus ruling out the

possibility of a card with a vowel and an odd number. Setting $\varepsilon = 0$ will be enough to model such assumption in our framework.

Wason's analysis of the task is best seen as addressing response frequencies as performance measure for participants' behaviour (see Table 1). Response frequencies are simply the proportions of participants who selected a given element in R, including all combinations of C_1-C_4. Observed behaviour shows that a majority of participants (60% to 80%) choose either C_1 or $C_1 \times C_3$.[7] According to Wason, such participants are actively choosing a dominated option, for a strictly better one is available, namely, $C_1 \times C_4$. In our notation, Wason's diagnosis of irrationality is committed to the implication that $U(T, C_1 \times C_4) > U(T,C_1)$, $U(T, C_1 \times C_3)$ (where T is a rational agent's epistemic target in the task), and surely this is fully in line with Wason's ([1966], [1968]) remarks. But how is this conclusion supported?

Wason seemed to endorse the principle that $U(T,C_4) > U(T,C_3)$ on the basis of an informal 'quasi-Popperian' line of reasoning.[8] The general idea would be something like the following: given two options $X,Y \in R$, if any member x_i of X can falsify an element in T (for example, it's logically incompatible with h, so that $P(h|x_i) = 0$), whereas no element in Y can falsify any element in T, then $U(T,X) > U(T,Y)$. As plausible as it may sound, this claim still remains starkly insufficient for two reasons. First, it does not offer any insight as to why should the ranking at issue hold in general for a rational agent. Second, and no less important, the quasi-Popperian principle above is too weak: given the material interpretation of the conditional (thus $\varepsilon = 0$), it justifies the ranking $U(T,C_4) > U(T,C_3)$, but it is completely silent for just those comparisons that seem crucial to sustain Wason's diagnosis of mistaken reasoning, namely, $U(T, C_1 \times C_4)$ *vs*. $U(T,C_1)$ and $U(T, C_1 \times C_4)$ *vs*. $U(T, C_1 \times C_3)$.

[7] There results remained robust across countless further replications – see e.g. (Stenning and van Lambalgen [2008], p. 46; Evans *et al*. [1993]; and Ragni, Kola, and Johnson-Laird [2018]).

[8] For further discussions of this point see, e.g., (Mercier and Sperber [2017], p. 212, and Humberstone [1994], p. 396).

Our approach outlined above provides a simple and satisfactory solution, filling both gaps in the traditional Wasonian approach to the task. The move required is to pick up s_0 from the Tsallis score formalism (see Appendix 1):

$$s_0(t_i, P) = \sum_{t_j \in T} P(t_j)^0 - 1$$

Given the convenient convention that $0^0 = 0$ (which is standard, in particular, in information theory), s_0 will simply correspond to the number of false hypotheses in T that P does not rule out. s_0 may well seem a poor measure of inaccuracy, as it actually ignores all the quantitative information conveyed by P, yet the measure $u_0(T,e)$ thus generated is not without interest: it yields the number of elements in T that become falsified by updating on the evidence e. In turn, we have that the corresponding expected reduction of inaccuracy $U_0(T,E)$ computes the expected number of hypotheses in T that will be falsified by performing evidence search E – a motivated form of 'quasi-Popperianism'.[9] Given only our basic assumptions (see Section 3. above), one can then prove that $U_0(T, C_1 \times C_4) > U_0(T,C_1) = U_0(T, C_1 \times C_3)$ for both $T = F$ (Wason's choice) and $T = D$ (the typical Bayesian choice) (see Appendix 2).

What's the upshot? For $\tau = 0$ our framework provides the basis for the justification of Wason's orderings that Wason himself lacked. This fills the relevant gaps in this first part of the story, clarifying the normative ground required to support Wason's original argument for the diagnosis of irrational behaviour.

6. Oaksford and Chater Revised

We now turn to the second episode in our story, Oaksford and Chater's analysis, that allegedly subvert the implications of Wason's. As compared to Wason, Oaksford and Chater take a different target for the assessment of reasoning performance in the task: the percentage of participants selecting

[9] See (Baron *et al*. [1988], p. 106) for an earlier occurrence.

each single card. These values can be interpreted as selection propensities (see Table 1): how much one is willing to turn each single card. This plausibly indicates how much, on average, a participant considers each elementary search option epistemically useful. Notably, this choice amounts to a different representation of the possible experimental outcomes as one of the twenty-four possible (strict) rankings of the four elementary evidence search options C_1-C_4. Actual figures for C_1, C_2, C_3, C_4 are 89%, 16%, 62%, and 25%, respectively, implying an aggregated judgment that $U(T,C_1) > U(T,C_3) > U(T,C_4) > U(T,C_2)$.[10]

We already know that Oaksford and Chater's model of the task can be obtained by our Table 2, given some basic background assumptions – random independent sampling, and flat prior on D (Section 3 above) —, setting $\alpha = \alpha^*$, $\beta = \beta^* = 1 - \varepsilon^*$, and allowing for, e.g., $\varepsilon = 0.1$. But what about the normative benchmark? Oaksford and Chater's recurrent approach has been to characterize epistemic utility as complementary to uncertainty, and to rely on Shannon entropy as a measure of uncertainty (Shannon [1948]). As a consequence, the epistemic utility of, say, turning card C_1 is given as the expected reduction of the initial uncertainty about T in view of the possible outcomes of the evidence search as determined by a probability distribution P (representing the agent's credal state). As a partial justification for the adoption of this approach, Oaksford and Chater have pointed out that Shannon entropy 'captures our intuitions about a measure of uncertainty' ([2003], p. 291). Evans and Over ([1996]), however, have labelled the choice of this formalism 'a clear mistake'. The source of Evans and Over's dissatisfaction is that a good measure for the epistemic utility of evidence search 'must be positive whenever data are diagnostic, that is, lead one to revise one's belief' ([1996], p. 362), a property that entropy reduction demonstrably lacks. Evans and Over have thus suggested an alternative measure — absolute log likelihood ratio

[10] See (Oaksford and Chater [1994]) for a summary of the influential meta-analysis from which relevant values have been derived.

([1996], p. 358) — against which, however, a powerful criticism was mounted by Nelson ([2005]).

Our accuracy-based framework neatly solves this quandary. As it happens, the Tsallis score for $\tau = 1$ is $s_1(h_i, P) = ln\left(\frac{1}{P(h_i)}\right)$, is the popular logarithmic score.[11] If inaccuracy is measured by the logarithmic score, then the expected reduction in inaccuracy $U_1(H,E)$ is demonstrably equivalent to the expected value of the Kullback-Leibler divergence, which in turn is always numerically identical to Oaksford and Chater's expected reduction of uncertainty as quantified by Shannon entropy.[12] Given that the underlying actual values of the Kullback-Leibler divergence are always non-negative, this can be taken to address Evans and Over's ([1996]) challenge, as noted by Oaksford and Chater themselves ([1996], pp. 381-2).

Our accuracy-based framework also goes beyond both sides of this controversy as it explains why it makes sense for a measure of the actual epistemic utility of a piece of evidence to be non-negative. Following a key insight recently emphasized by Roche and Shogenji ([2018]), the reason is this: the assessment of how the current credal state is inaccurate in expectation as compared to the earlier one should be made on the common ground of the current credal state itself, represented by P_e, for it is by definition better informed (after all, unlike the prior P, the posterior P_e embeds the truthful assumption that e holds). But then the non-negativity of $u(H,e) = S(P_e, P) - S(P_e, P_e)$ follows straight away, provided that the underlying score s is proper, which is taken to be an independently compelling constraint (of course satisfied, in particular, by the logarithmic score s_1). So our account recovers all

[11] See (Good [1952], and Gneiting and Raftery [2007]).

[12] See (Kullback and Leibler [1951], and Cover and Thomas [1991]). The choice of the logarithm base is a largely immaterial matter of convention. For $s_1(h_i, P)$, one can switch to the popular choice of base 2 by a multiplicative constant. Such simple transformation carries over to expected values (thus to expected inaccuracy as measured by $S_1(P,Q)$), to differences of expected values (thus to expected inaccuracy reduction $u_{\mathbf{1}}(H,e)$, KL divergence in this case), and to expected values of such differences (thus to $U_1(H,E)$). Also see Appendix 1.

implications of Oaksford and Chater's favourite machinery, while at the same time displaying principled normative foundations that are crucially lacking in their discussion.

Oaksford and Chater's analysis is still often portraited as a vindication of human reasoning in the Wason task through a 'paradigm shift' towards a probabilistic (rather than 'logical') interpretation of rationality. This idea is quite misleading, however. Oaksford and Chater's analysis accommodates the observed response pattern on the basis of so-called 'rarity assumption', namely, $P[vowel(c_i)] < P[even(c_i)] << P[\sim even(c_i)]$ (also see Fitelson and Hawthorne [2010], p. 233). In Oaksford and Chater's favourite parameter setting, in particular, one has $\alpha = 0.22$ and $\beta = 0.27$, so that $U_1(D,C_1) = 0.22 > U_1(D,C_3) = 0.14 > U_1(D,C_4) = 0.05 > U_1(D,C_2) = 0.03$. (This also works for plain response frequencies, at least to the extent that $U_1(D, C_1 \times C_3) = 0.31 > 0.25 = U_0(D, C_1 \times C_4$. See Appendix 3 for specification of the full joint probability distribution on $C_1 \times C_2 \times C_3 \times C_4 \times D$.) And yet, however valid rarity may be as a default assumption in many settings (including the ravens paradox), it has no plausible normative justification for a probabilistic representation of the Wason abstract selection task as it is. In fact, we submit that in the Wason task a plausible argument can be made against rarity, especially with regards to β, i.e., $P[even(c_i)]$.[13]

On reflection, why should one ever assume finding an even number on a card in the Wason task as significantly less probable than finding a non-even (odd) number? No clear suggestion in this direction arises from the hypothetical sampling procedures of the four cards available, and none has been put forward to the best of our knowledge. Quite on the contrary, for a Bayesian agent, the (second-order) levels of confidence in support of $P[even(c_i)] = x$ *vs*. $P[even(c_i)] = 1 - x$ must be indistinguishable for any x (with $0 < x < 1$) on the basis of symmetry considerations that seem very compelling in

[13] To what extent rarity is independently supported as a descriptive, psychological hypothesis about participants' attitudes is yet another potentially relevant issue which we leave aside here, but see, e.g., (Oberauer *et al*. [1999]).

the context, and an estimate of $\beta = ½$ is of course the most natural consequence of this premise.[14]

We have run a simple exhaustive grid search (at interval 0.01) of the parameter space of Oaksford and Chater's model under the plausible assumption that $\beta = ½$, and found that *all* settings tested converge on Wason's original diagnosis in two crucial respects: first, $U_1(D,C_4) > U_1(D,C_3)$, against measured single-card selection propensities (see Figure 1); and second, $U_1(D, C_1 \times C_4) > U_1(D, C_1 \times C_3)$, against plain observed response frequencies.

The crucial role of the rarity assumption in Oaksford and Chater's account is not a new topic, and its normative weakness was also noted before.[15] Still, a key implication seems to have been underappreciated: as fascinating as it is, Oaksford and Chater's analysis of the selection task does not imply a revision of Wason's original normative assessment in a compelling way. In fact, the apparent rationalization of the participants' behaviour does not arise from the application of Bayesian principles as 'the appropriate normative theory' (Oaksford and Chater [2007], p. 31), but from normatively questionable auxiliary assumptions.

[14] Oaksford and Chater's discussion ([1994], see pp. 627-28) does not dispel this criticism, in our view. In fact, they indirectly argue for the claim that rarity is typically factually adequate 'in our environment', by which they clearly mean in ordinary language and reasoning outside the lab, thus not in the abstract Wason task itself. This reveals that their notion of 'adaptive rationality', even if correct, is consistent with systematic local departures from sound reasoning, which is the only relevant point for our current purposes. The 'extrapolation' of rarity 'from prior experience' to the 'novel' Wason task (were it indeed the case) is no more and no less 'reasonable', we submit, than the analogue 'extrapolation' by which our perceptual system becomes liable to illusions such as Müller-Lyer. Illusion remains illusion, and mistake remains mistake. Indeed, in the abstract Wason task, such tendency to extrapolate (again, if real) unduly overrides very plausible motivations for a Bayesian agent to rely on $P[even(c_i)] = ½$. And such specific and guarded motivations also elude Oaksford and Chater's ([1994], p. 627) quick dismissal of the principle of indifference as an alternative to their own favourite assumptions.

[15] See, e.g., (Laming [1996], and Sloman and Fernbach [2008]).

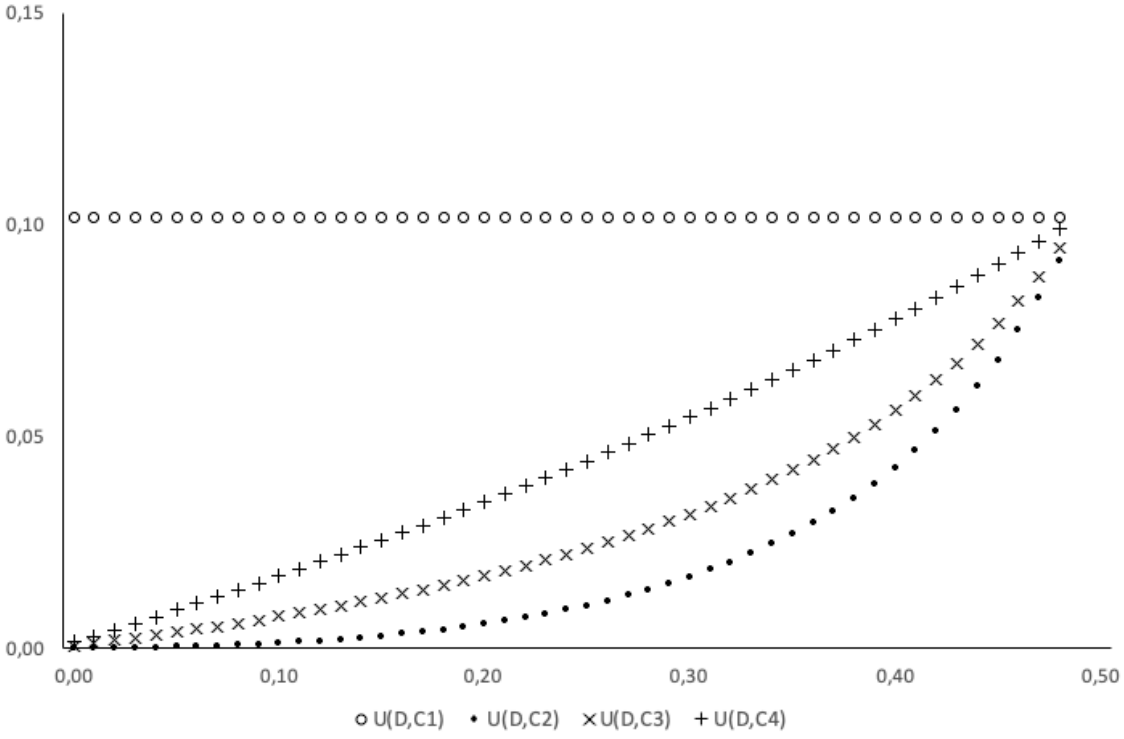


Figure 1. Expected utility of single-card selections in Wason's task in Oaksford and Chater's ([1994], [2003]) analysis. Parameter β (the probability of an even number on the card) is set at 0.5, whereas α (the probability of a vowel on the card) varies on the *x* axis while still satisfying rarity (< 0.5).

The upshot: for $\tau = 1$ our generalized framework provides a coherent normative justification converging with the formalism chosen by Oaksford and Chater ([1994], [2003]). This fills a theoretical gap in their story. At the same time, it shows that, despite claims to the contrary, participants in the Wason selection task do not act as Bayesian rational agents plausibly would. Without key but doubtful assumptions about rarity, a sound reconstruction of Oaksford and Chater's favourite model neatly follows Wason's conclusions.

7. Nickerson and Fitelson and Hawthorne Amended

Both Nickerson ([1996]) and Fitelson and Hawthorne ([2010]) pursue the idea of an explicit connection between the Wason task and Bayesian confirmation theory — a project that we support. Still, we find their proposals defective in

some important respects. In this section we'll explain why, but our constructive criticism and friendly amendments require careful preparation first.

Drawing on (Nickerson [1996]) and contemporary Bayesian confirmation theory,[16] Fitelson and Hawthorne ([2010]) define the confirmational power of evidence *e* for hypothesis *h* as the absolute value of the difference between posterior and prior probability: $D(h,e) = |P(h|e) - P(h)|$. So far so good: the absolute value here allows one to appropriately bracket whether *e*'s contribution increases or decreases the credibility of *h*, the idea being that *e* is equally useful either way. But here is how Fitelson and Hawthorne— again following Nickerson— extend this idea to a test or evidence search option $E = \{e_1, \ldots, e_m\}$:

$$\mathcal{P}(h, E) = \sum_{e_j \in E} \left|P\left(h|e_j\right) - P(h)\right| P\left(e_j\right)$$

This quantity, however, cannot represent the epistemic utility of *E* for an agent who aims at finding out the truth about a target hypothesis space to which *h* belongs. To illustrate, suppose we have: $H = \{h_1, h_2, h_3, h_4\}$ and $E = \{e, \sim e\}$; a prior assignment on *H* of 40%, 30%, 20%, and 10%, respectively; 55%, 20%, 15%, 10% as a posterior distribution given *e*; and 25%, 40%, 25% and 10% as a posterior distribution given ~*e* (all this implies that $P(e) = 50\%$). We then have $\mathcal{P}(h_1,E) = 0.15$, $\mathcal{P}(h_2,E) = 0.10$, $\mathcal{P}(h_3,E) = 0.05$, and $\mathcal{P}(h_4,E) = 0$. What is then the overall epistemic utility of *E* to find out about *H?* We should have one number, but we have four — something is not quite working right.

What one could (and should) have done instead with confirmational power as absolute probability difference is to posit:

$$u_D(H, e) = \frac{1}{|H|} \sum_{h_i \in H} |P(h_i|e) - P(h_i)|$$

(where $|H|$ denotes the cardinality of *H*) and then, as usual, $U_D(H, E) = \sum_{e_j \in E} P(e_j) \cdot u\left(H, e_j\right)$. Now, this measure does yield a definite assessment of test *E* for target hypothesis set *H* in our illustrative example above (i.e., 0.075).

[16] See, e.g., (Crupi [2015], and Crupi and Tentori [2016]).

And predictably this measure also occurs in important earlier discussions in the literature (especially Nelson [2005], under the label of 'impact').

It is true that $\mathcal{P}(h,E) = U_D(H,E)$ for any E as long as H is a binary hypothesis set, i.e., $H = \{h,\sim h\}$, which happens to be the case in the Wason task as usually understood. But this should not obscure the fact that $\mathcal{P}(h,E)$ is just not the right kind of model, as it relates a single hypothesis and a possible experiment. Our favourite suggestion to make sense of $\mathcal{P}(h,E)$ is as a measure of the testability of the specific hypothesis h through experiment E. Quite plausibly, the more the hypotheses in H that are testable through E, and the more testable they are, the higher the epistemic value of E with respect to H overall. Yet the two notions are, and should be kept, conceptually and formally distinct. Conflating the testability of one single hypothesis h in H by E and the epistemic utility of E for H is a mistake.

This is our first issue with Nickerson and Fitelson and Hawthorne. If it were the only issue, it could be solved by replacing $\mathcal{P}(h,E)$ with $U_D(H,E)$. That's surely an improvement, but still limited, we believe. As far as we can see, the sole motivation to be found in favour of $U_D(H,E)$ is the plausible idea that the epistemic utility of test E for H should arise as a weighted average of the confirmational power of the elements of E relative to the elements of H. By construction, this approach will imply that $u_D(H,e)$ is strictly positive as long as the posterior probability departs from the prior — a property that at least some authors (such as Evans and Over [1996], see above) find very attractive. We tend to concur, and yet there are still some potential problems. First, these remarks do not go much deeper than Oaksford and Chater's alternative idea that the epistemic utility of a test should amount to a reduction of uncertainty (which, as we know, can well be negative). Secondly, $U_D(H,E)$ lacks other important formal properties, like the following:

$$\textit{Additivity}\text{: } U(H,E\times F) = U(H,E) + U(H,F|E)$$

The above statement means that the epistemic utility of a combined test $E \times F$ amounts to the sum of the plain utility of E and the utility of F that is expected considering all possible outcomes of E (where $U(H,F|E) = \sum_{e_j \in E} U\left(H,F|e_j\right) \cdot$

$P(e_j)$ and $U(H, F|e_j)$ denotes the expected utility of *F* for *H* computed when all probabilities are conditionalized on e_j). Roughly, this additivity principle represents the idea that, for an agent assessing in advance the utility of *E* and *F* combined, it does not matter in which order the outcomes of those tests are expected to be revealed. Additivity is important and highly desirable as concerns the analysis of the rational assessment of tests. Yet it is demonstrably *violated* by $U_D(H,E)$.[17]

Our generalized framework addresses all difficulties above by choosing $\tau = 2$. The corresponding Tsallis score s_2 amounts to the Brier score (or 'squared Euclidean distance' – see Selten [1998], and Appendix 1 for a proof of the connection). Expected reduction in inaccuracy will then be measured as follows:

$$U_2(H,E) = \sum_{e_j \in E} \sum_{h_i \in H} \left[P(h_i|e_j) - P(h_i)\right]^2 P(e_j)$$

This measure, we urge, retains the spirit and the attractive features of Nickerson's and Fitelson and Hawthorne's theoretical approaches while overcoming all limitations outlined above. Here, the actual epistemic utility $u_2(H,e)$ is indeed the sum of the confirmational power of *e* for every hypothesis in *H* as quantified by a confirmation measure that is ordinally equivalent to the traditional probability difference and also recently discussed by van Enk ([2014]). The additivity property above is also happily satisfied by $U_2(H,E)$.[18] Finally, and importantly, this arrangement does not have to be accepted on purely intuitive grounds (as nice as they can be), but arises once again as a consequence of the general accuracy-based approach given the specific choice of the Brier score.

We are left, of course, with one final point to be discussed: the implications for observed behaviour in the Wason task. As is typical in the tradition of the Wason / ravens parallelism, both Nickerson ([1996]) and Fitelson and Hawthorne ([2010]) presuppose that $T = D$ and $\varepsilon = 0$, so that $d = \forall x[vowel(x) \supset even(x)]$ (quantifying over all cards in the allegedly large sampled

[17] See (Nelson [2008]) for discussion.

[18] See (Crupi and Tentori [2014]).

deck) and the foil hypothesis, ~*d*, is the plain logical negation. Nickerson's specific numerical model ([1996], p. 16) also embeds the assumption that $\varepsilon^* = ½$. Given the material implication interpretation of *d*, this assumption can be supported again by a plausible symmetry argument. In fact, on the supposition that *d* is false, the probability for a card with a vowel to also have an even number will have to be strictly lower than 1, but for a Bayesian agent the (second-order) levels of confidence in support of $\varepsilon^* = P[\sim even(c_i) \mid vowel(c_i) \wedge \sim h] = x$ *vs*. $P[\sim even(c_i) \mid vowel(c_i) \wedge \sim h] = 1 - x$ must be indistinguishable for any *x* (with $0 < x < 1$). If this consideration is applied, then the agreement between Nickerson and Oaksford and Chater in defense of the rationality of participants' behaviour rests on the same shaky ground. Here again, a grid search (at interval 0.01) of the parameter range for α shows that, if rarity is rejected at least for $P[even(c_i)]$ (so that $\beta = ½$), then Wason's original diagnosis remains valid, namely: $U_2(D,C_4) > U_2(D,C_3)$, against measured single-card selection propensities; and $U_2(D, C_1 \times C_4) > U_2(D, C_1 \times C_3)$, against plain observed response frequencies.[19]

Fitelson and Hawthorne have proven cautious about the rationalization of behaviour in the Wason selection task, and on this we definitely concur (although for quite different motivations, compare [2010], pp. 234-5). Still, one specific point in their extensive discussion deserves comment, as it contributes to the generality of our conclusions. In our understanding, Fitelson and Hawthorne's parallelism between the ravens paradox and the Wason task suggests that the traditional setting $\alpha = \alpha^*$ and $\beta = \beta^*$ would better be relaxed to allow for $\alpha \leq \alpha^*$, and $\beta \geq \beta^*$ (see [2010], pp. 221-3). Here is the idea, informally. If the universally quantified conditional *d* is false, then one may well expect a comparably higher probability for a vowel card (so that $\alpha < \alpha^*$) or for

[19] It should be pointed that, while Nickerson's ([1996]) 'illustrative' numerical example does embed rarity (with $\alpha = 0{,}05$ and $\beta = 0{,}10$, in our notation) as well as Oaksford and Chater's independence settings (with $\alpha = \alpha^*$ and $\beta = \beta^*$), his discussion does not provide arguments in support of such assumptions as normatively sound for the abstract Wason task.

an odd number card (so that $\beta > \beta^*$), for the instances in the population which would make *d* false must be of that kind.

What is interesting for our purposes is that all the plausible conditions which Nickerson ([1996]) and Fitelson and Hawthorne ([2010]) favour can be jointly satisfied. And again, to the extent that rarity is rejected at least for $P[even(c_i)]$, the rationalization of observed behaviour typically fails. For an illustrative numerical example, consider the parameter setting $\alpha = 0.15 < \alpha^* = 0.25$, $\beta = 0.55 > \beta^* = 0.45$, and $\varepsilon^* = 0.50$, whereby one has $U_2(D,C_4) = 0.064 > 0.00002 = U_2(D,C_3)$, against measured single-card selection propensities, and $U_2(D, C_1 \times C_4) = 0.221 > 0.167 = U_2(D, C_1 \times C_3)$, against plain observed response frequencies. (See Appendix 3 for specification of the full joint probability distribution on $C_1 \times C_2 \times C_3 \times C_4 \times D$.)

The upshot: For $\tau = 2$ our generalized framework recovers an amended version of Nickerson's ([1996]) and Fitelson and Hawthorne's ([2010]) analyses. At the same time, it clearly shows that, without theoretical analogues of the rarity assumption, Bayesian models of this kind neatly follow Wason as well.

Conclusion

Wason's original intuition was correct. Meanwhile, claims have been made that controversies on human rationality in experimental reasoning tasks are pointless as they are bound to drown in the intractable problem of the 'arbitration' between competing norms (e.g., Elqayam and Evans [2011]). At least in the paramount case of the selection task, our discussion suggests a very different picture. Three major approaches as different as a 'logical' (Wason), an information-theoretic (Oaksford and Chater), and a confirmation-theoretic (Nickerson / Fitelson and Hawthorne) analysis have been all recovered as specifications of a unified view of rational inquiry as aiming at inaccuracy reduction. And as it turns out, the verdict of defective reasoning has remained unscathed across these variations.

The Wason selection task has been called 'a mysterious beast' (Manktelow [2012], p. 113), bewitching three generations of reasoning scholars – and not without reasons. Despite extensive investigation, a complete

psychological account of observed behaviour is still an open scientific challenge (see, e.g., Ragni, Kola, and Johnson-Laird [2018]). Yet the normative interpretation of the task is no mystery, we believe: the routes of the Bayesians lead back to where it all started, with Wason.

Acknowledgments

We would like to thank Peter Brössel, Martina Calderisi, Matthis Hesse, Edouard Machery, Wim Mol, Jan Sprenger, Corina Stroessner, Nina Poth, and two anonymous reviewers for their helpful comments on previous versions of this paper.

Filippo Vindrola
Ruhr University Bochum
Bochum, Germany
filippo.vindrola@gmail.com

Vincenzo Crupi
University of Turin
Torino, Italy
vincenzo.crupi@unito.it

REFERENCES

Baron J., Beattie J., and Hershey J.C. [1988]: 'Heuristics and Biases in Diagnostic Reasoning II: Congruence, Information, and Certainty', *Organizational Behavior and Human Decision Processes*, **42**, pp. 88–110.

Bradley D. [2015]: *A Critical Introduction to Formal Epistemology*, Bloomsbury.

Brössel P. and Huber F. [2014]: 'Bayesian Confirmation: A Means with No End', *British Journal for the Philosophy of Science*, **66**, pp. 737–49.

Campbell-Moore C. and Levinstein B. [2020]: 'Strict Propriety Is Weak'. *Analysis*.

Carr J. [2017]: 'Epistemic Utility Theory and the Aim of Belief', *Philosophy and Phenomenological Research*, **95**, pp. 511–34.

Cover T.M. and Thomas J.A. [1991]: *Elements of Information Theory*, John Wiley & Sons.

Crupi, V. [2015]: 'Inductive Logic', *Journal of Philosophical Logic*, **44**, pp. 641-50.

Crupi V. and Tentori K. [2016]: 'Confirmation Theory', in A. Hájek and C. Hitchcock (eds.), *Oxford Handbook of Philosophy and Probability* (pp. 650-65), Oxford University Press.

Crupi V. and Tentori, K. [2014]: 'Measuring Information and Confirmation', *Studies in the History and Philosophy of Science*, **47**, pp. 81-90.

Crupi V., Nelson J., Meder B., Cevolani G., and Tentori K. [2018]: 'Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search', *Cognitive Science*, **42**, pp. 1410-56.

D'Agostino M. and Sinigaglia C. [2010]: 'Epistemic Accuracy and Subjective Probability', in M. Suárez, M. Dorato, and M. Rèdei (eds.), *Epistemology and Methodology of Science* (pp. 95–105), Springer.

Dawid A.P. [1998]: 'Coherent Measures of Discrepancy, Uncertainty, and Dependence, with Applications to Bayesian Predictive Experimental Design', Research Report 139, Dept. of Statistical Science, UCL.

Douven, I. [2020]: 'Scoring in Context', *Synthese*, **197**, pp. 1565-80.

Dunn J. [2019]: 'Accuracy, Verisimilitude, and Scoring Rules', *Australasian Journal of Philosophy*, **97**, pp. 151-66.

Elqayam S. and Evans J.St.B.T. [2011]: 'Subtracting *Ought* from *Is*', *Behavioral and Brain Sciences*, **34**, pp. 233–48.

Evans J., Newstead S., and Byme R.M.J. [1993]: *Human Reasoning: The Psychology of Deduction*, Lawrence Erlbaum Associates.

Evans, J.St.B.T. and Over, D.E. [1996]: 'Rationality in the Selection Task: Epistemic Utility versus Uncertainty Reduction', *Psychological Review*, **103**, pp. 356–63.

Fallis D. and Lewis P.J. [2016]: 'The Brier Rule Is not a Good Measure of Epistemic Utility (and Other Useful Facts about Epistemic Betterness)', *Australasian Journal of Philosophy*, **94**, pp. 576-90.

Fitelson B. and Hawthorne J. [2010]: 'The Wason Task(s) and the Paradox of Confirmation', *Philosophical Perspectives*, **24** (Epistemology), pp. 207-41.

Gneiting T. and Raftery A. [2007]: 'Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, **102**, pp. 359-78.

Good I.J. [1952]: 'Rational Decisions', *Journal of the Royal Statistical Society B*, **14**, pp. 107-14.

Greaves H. [2013]: 'Epistemic Decision Theory', *Mind*, **122**, pp. 915–52.

Humberstone C.G. [1994]: 'Hempel Meets Wason', *Erkenntnis*, **41**, pp. 391-402.

Konek J. and Levinstein B.A. [2019]: 'The Foundations of Epistemic Decision Theory', *Mind*, **128**, pp. 69-107.

Kullback S. and Leibler R.A. [1951]: 'On Information and Sufficiency', *Annals of Mathematical Statistics*, **22**, pp. 79-86.

Laming D. [1996]: 'On the Analysis of Irrational Data Selection: A Critique of Oaksford and Chater (1994)', *Psychological Review,* **103***,* pp. 364-73.

Leitgeb H. and Pettigrew R. [2010]: 'An Objective Justification of Bayesianism I: Measuring Inaccuracy', *Philosophy of Science*, **77**, pp. 201-35.

Mercier H. and Sperber D. [2017]: *The Enigma of Reason*, Harvard University Press.

Nelson J.D. [2005]: 'Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain', *Psychological Review*, **112**, pp. 979-99.

Nelson J.D. [2008]: 'Towards a Rational Theory of Human Information Acquisition', in M. Oaksford, and N. Chater (eds.), *The Probabilistic Mind: Prospects for Rational Models of Cognition* (pp. 143-63), Oxford University Press.

Nickerson R. [1996]: 'Hempel's Paradox and Wason's Selection Task: Logical and Psychological Puzzles of Confirmation', *Thinking and Reasoning*, **2**, pp. 1-31.

Oaksford M. and Chater N. [1994]: ‘A Rational Analysis of the Selection Task as Optimal Data Selection, *Psychological Review*, **101**, pp. 608-31.

Oaksford M. and Chater N. [2003]: ‘Optimal Data Selection: Revision, Review, and Reevaluation’, *Psychonomic Bulletin & Review*, **10**, pp. 289-318.

Oberauer K., Wilhelm O., and Diaz R.R. [1999]: ‘Bayesian Rationality for the Wason Selection Task? A Test of Optimal Data Selection Theory’, *Thinking and Reasoning*, **5**, pp. 115-44.

Oddie G. [2019]: ‘What Accuracy Could Not Be’, *British Journal for the Philosophy of Science*, **70**, pp. 551-80.

Pettigrew R. [2013]: ‘Epistemic Utility and Norms for Credences’, *Philosophy Compass*, **8**, pp. 897–908.

Predd J.B., Seiringer R., Lieb E.J., Osherson D., Poor H.V., and Kulkarni S.R. [2009]: ‘Probabilistic Coherence and Proper Scoring Rules’, *IEEE Transactions on Information Theory*, **55**, pp. 4786-92.

Ragni, M., Kola I., and Johnson-Laird P. N. [2018]: ‘On Selecting Evidence to Test Hypotheses: A Theory of Selection Tasks’, *Psychological Bulletin*, **144**, pp. 779–96.

Roche W. and Shogenji T. [2018]: ‘Information and Inaccuracy’, *British Journal for the Philosophy of Science*, **69**, pp. 577-604.

Savage L.J. [1971]: ‘Elicitation of Personal Probabilities and Expectations’, *Journal of the American Statistical Association*, **66**, pp. 783-801.

Schoenfield M. [2017]: ‘The Accuracy and Rationality of Imprecise Credences’, *Noûs*, **51**, pp. 667-85.

Schurz G. [2011]: ‘Truth-Conduciveness as the Primary Epistemic Justification of Normative Systems of Reasoning’, *Behavioral and Brain Sciences*, **34**, pp. 226-7.

Schurz G. [2015]: ‘Cognitive Success: Instrumental Justification of Normative Systems of Reasoning’, *Frontiers in Psychology*, **5**, 625.

Selten R. [1998]: ‘Axiomatic Characterization of the Quadratic Scoring Rule’, *Experimental Economics*, **1**, pp. 43-61.

Shannon C.E. [1948]: ‘A Mathematical Theory of Communication’, *Bell System Technical Journal*, **27**, pp. 379-423 and 623-56.

Sloman S.A. and Fernbach P.M. [2008]: ‘The Value of Rational Analysis: An Assessment of Causal Reasoning and Learning’, in N. Chater and M. Oaksford (eds.). *The probabilistic mind: Prospects for rational models of cognition* (pp. 486-500), Oxford University Press.

Stein E. [1996]: *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*, Clarendon Press.

Stenning K. and van Lambalgen M. [2008]: *Human Reasoning and Cognitive Science*, MIT Press.

Stich S. [1990]: *The Fragmentation of Reason*, MIT Press.

Tsallis C. [1988]: 'Possible Generalization of Boltzmann-Gibbs Statistics', *Journal of Statistical Physics*, **52**, pp. 479–87.

Tsallis C. [2011]: 'The Nonadditive Sntropy S_q and Its Applications in Physics and Elsewhere: Some Remarks', *Entropy*, **13**, pp. 1765-804.

van Enk S.J. [2014]: 'Bayesian Measures of Confirmation from Scoring Rules', *Philosophy of Science*, **81**, pp. 101-13.

Wason P. [1966]: 'Reasoning', in B. Foss (ed.), *New Horizons in Psychology* (pp. 135-51), Penguin.

Wason P. [1968]: 'Reasoning about a Rule', *Quarterly Journal of Experimental Psychology*, **20**, pp. 273-81.

TECHNICAL APPENDICES

1. Special Cases of the Tsallis Scores

The case $\tau = 0$ is straightforward. We have

$$s_\tau(h_i, P) = \tau ln_\tau\left(\frac{1}{p_i}\right) - \left(1 - \sum_{h_j \in H} p_j^\tau\right) = \sum_{h_j \in H} p_j^0 - 1$$

For $\tau = 1$, we consider the Tsallis logarithm, $ln_\tau(x) = \frac{1}{1-\tau}\left[x^{(1-\tau)} - 1\right]$. To show that the ordinary natural logarithm is recovered from $ln_\tau(x)$ ($x > 0$) in the limit for $\tau \to 1$, posit $x = 1 - y$ and first consider $x \leq 1$, so that $|-y| < 1$. Then we have

$$\lim_{\tau \to 1}\{ln_\tau(x)\} = \lim_{\tau \to 1}\{ln_\tau(1-y)\} = \lim_{\tau \to 1}\left\{\frac{1}{1-\tau}\left[(1-y)^{(1-\tau)} - 1\right]\right\}$$

and, by the binomial expansion of $(1-y)^{(1-\tau)}$:

$$\begin{aligned}
&\lim_{\tau \to 1}\left\{\frac{1}{1-\tau}\left[(1-y)^{(1-\tau)} - 1\right]\right\} = \\
&\quad = \lim_{\tau \to 1}\left\{\frac{1}{1-\tau}\left[-1 + \left(1 + (1-\tau)(-y) + \frac{(1-\tau)(1-\tau-1)(-y)^2}{2!} + \frac{(1-\tau)(1-\tau-1)(1-\tau-2)(-y)^3}{3!} + \cdots\right)\right]\right\} \\
&\quad = \lim_{\tau \to 1}\left\{(-y) + \frac{(-\tau)(-y)^2}{2!} + \frac{(-\tau)(-\tau-1)(-y)^3}{3!} + \cdots\right\} \\
&\quad = \lim_{\tau \to 1}\left\{(-y) - \frac{\tau(-y)^2}{2!} + \frac{(\tau)(\tau+1)(-y)^3}{3!} - \cdots\right\} \\
&\quad = (-y) - \frac{(-y)^2}{2!} + \frac{2!(-y)^3}{3!} - \cdots \\
&\quad = (-y) - \frac{(-y)^2}{2} + \frac{(-y)^3}{3} - \cdots
\end{aligned}$$

which is the series expansion of $ln(1-y) = ln(x)$ (recall that $|-y| < 1$). For the case $x > 1$, one can posit $x = 1/(1-y)$, so that again $|-y| < 1$ and compute $\lim_{\tau \to 1}\left\{\frac{\left(\frac{1}{1-y}\right)^{(1-\tau)} - 1}{1-\tau}\right\} = \lim_{\tau \to 1}\left\{-\frac{1}{\tau-1}\left[(1-y)^{(\tau-1)} - 1\right]\right\}$, thus getting the same result from a similar derivation. This justifies positing $ln_1(x) = ln(x)$. As a consequence, $s_1(h_i, P)$ indeed amounts to the logarithmic score:

$$s_1(h_i, P) = ln_1\left(\frac{1}{p_i}\right) - \left(1 - \sum_{h_j \in H} p_j\right) = ln\left(\frac{1}{p_i}\right)$$

Finally, for the case $\tau = 2$, we have:

$$\begin{aligned}
s_\tau(h_i, P) &= \tau ln_\tau\left(\frac{1}{p_i}\right) - \left(1 - \sum_{h_j \in H} p_j^\tau\right) = 2ln_2\left(\frac{1}{p_i}\right) - \left(1 - \sum_{h_j \in H} p_j^2\right) \\
&= 2(1-p_i) - \left(1 - \sum_{h_j \in H} p_j^2\right) = 1 - 2p_i + \sum_{h_j \in H} p_j^2 \\
&= \left(1 - 2p_i + p_i^2\right) + \sum_{h_j \in H - \{h_i\}} p_j^2 = (1-p_i)^2 + \sum_{h_j \in H - \{h_i\}} p_j^2
\end{aligned}$$

and the latter just is the Brier score.

2. Accuracy in the Wason Task with $\tau = 0$

First we show that $U_0(D,C_1\times C_3) = U_0(D,C_1)$. For simplicity, we start denoting $even(c_1) \wedge vowel(c_3)$, $even(c_1) \wedge \sim vowel(c_3)$, $\sim even(c_1) \wedge vowel(c_3)$, $\sim even(c_1) \wedge \sim vowel(c_3)$ as x, y, w, and z, respectively.

$$
\begin{aligned}
U_0(D, C_1 \times C_3) &= P(x)\, u_0(D,x) + P(y)\, u_0(D,y) + P(w)\, u_0(D,w) + P(z)\, u_0(D,z) \\
&= P(x)\ [S_0(P_x,P) - S_0(P_x,P_x)] + P(y)\ \left[S_0\left(P_y,P\right) - S_0\left(P_y,P_y\right)\right] + P(w)\ [S_0(P_w,P) - S_0(P_w,P_w)] \\
&\qquad + P(z)\ [S_0(P_z,P) - S_0(P_z,P_z)] \\
&= P(x)\{[P_x(d)s_0(d,P) + P_x(\sim d)s_0(\sim d,P)] - [P_x(d)s_0(d,P_x) + P_x(\sim d)s_0(\sim d,P_x)]\} \\
&\qquad + P(y)\{\left[P_y(d)s_0(d,P) + P_y(\sim d)s_0(\sim d,P)\right] - \left[P_y(d)s_0\left(d,P_y\right) + P_y(\sim d)s_0\left(\sim d,P_y\right)\right]\} \\
&\qquad + P(w)\{[P_w(d)s_0(d,P) + P_w(\sim d)s_0(\sim d,P)] - [P_w(d)s_0(d,P_w) + P_w(\sim d)s_0(\sim d,P_w)]\} \\
&\qquad + P(z)\{[P_z(d)s_0(d,P) + P_z(\sim d)s_0(\sim d,P)] - [P_z(d)s_0(d,P_z) + P_z(\sim d)s_0(\sim d,P_z)]\} \\
&= P(x)\{1-1\} + P(y)\{1-1\} + P(w)\{1-0\} + P(z)\{1-0\} \\
&= P(w \vee z) \\
&= P[\sim even(c_1)] \\
&= P[even(c_1)]\{1-1\} + P[\sim even(c_1)]\{1-0\} \\
&= P[even(c_1)]\{\left[P_{even(c_1)}(d)s_0(d,P) + P_{even(c_1)}(\sim d)s_0(\sim d,P)\right] \\
&\qquad - \left[P_{even(c_1)}(d)s_0\left(d,P_{even(c_1)}\right) + P_{even(c_1)}(\sim d)s_0\left(\sim d,P_{even(c_1)}\right)\right]\} \\
&\qquad + P[\sim even(c_1)]\{\left[P_{\sim even(c_1)}(d)s_0(d,P) + P_{\sim even(c_1)}(\sim d)s_0(\sim d,P)\right] \\
&\qquad - \left[P_{\sim even(c_1)}(d)s_0\left(d,P_{\sim even(c_1)}\right) + P_{\sim even(c_1)}(\sim d)s_0\left(\sim d,P_{\sim even(c_1)}\right)\right]\} \\
&= P[even(c_1)]\ \left[S_0\left(P_{even(c_1)},P\right) - S_0\left(P_{even(c_1)},P_{even(c_1)}\right)\right] \\
&\qquad + P[\sim even(c_1)]\left[S_0\left(P_{\sim even(c_1)},P\right) - S_0\left(P_{\sim even(c_1)},P_{\sim even(c_1)}\right)\right] \\
&= P[even(c_1)]\, u_0\left(D, even(c_1)\right) + P[\sim even(c_1)]\, u_0\left(D, \sim even(c_1)\right) = U_0(D,C_1)
\end{aligned}
$$

Now we compute $U_0(D,C_1\times C_4)$, here denoting $even(c_1) \wedge vowel(c_4)$, $even(c_1) \wedge \sim vowel(c_4)$, $\sim even(c_1) \wedge vowel(c_4)$, $\sim even(c_1) \wedge \sim vowel(c_4)$ as x, y, w, and z, respectively. In this case, we have:

$$
\begin{aligned}
U_0(D, C_1 \times C_4) &= P(x)\, u_0(D,x) + P(y)\, u_0(D,y) + P(w)\, u_0(D,w) + P(z)\, u_0(D,z) \\
&= P(x)\{[P_x(d)s_0(d,P) + P_x(\sim d)s_0(\sim d,P)] - [P_x(d)s_0(d,P_x) + P_x(\sim d)s_0(\sim d,P_x)]\} \\
&\qquad + P(y)\{\left[P_y(d)s_0(d,P) + P_y(\sim d)s_0(\sim d,P)\right] - \left[P_y(d)s_0\left(d,P_y\right) + P_y(\sim d)s_0\left(\sim d,P_y\right)\right]\} \\
&\qquad + P(w)\{[P_w(d)s_0(d,P) + P_w(\sim d)s_0(\sim d,P)] - [P_w(d)s_0(d,P_w) + P_w(\sim d)s_0(\sim d,P_w)]\} \\
&\qquad + P(z)\{[P_z(d)s_0(d,P) + P_z(\sim d)s_0(\sim d,P)] - [P_z(d)s_0(d,P_z) + P_z(\sim d)s_0(\sim d,P_z)]\} \\
&= P(x)\{1-0\} + P(y)\{1-1\} + P(w)\{1-0\} + P(z)\{1-0\} \\
&= P(x) + P(w \vee z) = P[even(c_1) \wedge vowel(c_4)] + P[\sim even(c_1)]
\end{aligned}
$$

Then clearly, $U_0(D,C_1\times C_4) > U_0(D,C_1\times C_3) = U_0(D,C_1)$, because

$$
\begin{aligned}
U_0(D, C_1 \times C_4) - U_0(D, C_1 \times C_3) &= P[even(c_1) \wedge vowel(c_4)] \\
&= P[even(c_1) \wedge vowel(c_4)|d]P(d) + P[even(c_1) \wedge vowel(c_4)|\sim d]P(\sim d) \\
&= P[even(c_1) \wedge vowel(c_4)|\sim d]P(\sim d) \\
&= P[even(c_1)|\sim d]P[vowel(c_4)|\sim d]P(\sim d) = \frac{1}{2}\alpha^*\beta^*
\end{aligned}
$$

which is strictly positive, as $\alpha^*,\beta^* > 0$.

The above derivation applies with essentially no variation by replacing D with F.

3. Full Joint Probability Distributions for the Wason Task

Below is the full joint probability distribution over $C_1 \times C_2 \times C_3 \times C_4 \times D$ arising from Oaksford and Chater's analysis (see Section 6 above). Figures are obtained from *Table 2* given a flat prior over D, the independent sampling assumption for the four cards, and the following parameter setting (including rarity): $\alpha = \alpha^* = 0.22$; $\beta = \beta^* = 1 - \varepsilon^* = 0.27$; $\varepsilon = 0.1$. All probabilities are conditionalized on information about the visible sides of the four cards in the abstract Wason task, namely, on $vowel(c_1) \wedge \sim vowel(c_2) \wedge even(c_3) \wedge \sim even(c_4)$.

$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 918018967334036 \cdot 10^{-18}$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 176418 \cdot 10^{-8}$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 295435194942044 \cdot 10^{-16}$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 625482 \cdot 10^{-8}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 333825079030559 \cdot 10^{-18}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 625482 \cdot 10^{-8}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 107430979978925 \cdot 10^{-16}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 2217618 \cdot 10^{-8}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 902718651211802 \cdot 10^{-17}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 476982 \cdot 10^{-8}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 290511275026343 \cdot 10^{-15}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 1691118 \cdot 10^{-8}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 328261327713382 \cdot 10^{-17}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 1691118 \cdot 10^{-8}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 105640463645943 \cdot 10^{-15}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 5995782 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 10200210748156 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 476982 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 328261327713382 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 1691118 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 370916754478398 \cdot 10^{-19}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 1691118 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 119367755532139 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 5995782 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 100302072356867 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 1289618 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 322790305584826 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 4572282 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 364734808570425 \cdot 10^{-18}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 4572282 \cdot 10^{-8}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 117378292939937 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 16210818 \cdot 10^{-8}$

Below is the illustrative full joint probability distribution over $C_1 \times C_2 \times C_3 \times C_4 \times D$ employed in Section 7 above, dismissing rarity for $P[even(c_i)]$ but otherwise in line with conditions favoured by Nickerson (1996) and Fitelson and Hawthorne (2010). Figures are obtained from *Table 2* given a flat prior over D, the independent sampling assumption for the four cards, and the following parameter setting: $\alpha = 0.15$; $\alpha^* = 0.25$; $\beta = 0.55$; $\beta^* = 0.45$; $\varepsilon = 0$; and $\varepsilon^* = 0.5$. All probabilities are conditionalized on information about the visible sides of the four cards in the abstract Wason task, namely, on $vowel(c_1) \wedge \sim vowel(c_2) \wedge even(c_3) \wedge \sim even(c_4)$.

$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 683922558922559 \cdot 10^{-17}$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 641711229946524 \cdot 10^{-16}$
$P[even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 23253367003367 \cdot 10^{-15}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 177819865319865 \cdot 10^{-16}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 171122994652406 \cdot 10^{-15}$
$P[even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 604587542087542 \cdot 10^{-16}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 894360269360269 \cdot 10^{-17}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 72192513368984 \cdot 10^{-15}$
$P[even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 304082491582492 \cdot 10^{-16}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 23253367003367 \cdot 10^{-15}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 192513368983957 \cdot 10^{-15}$
$P[even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 790614478114478 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 683922558922559 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 23253367003367 \cdot 10^{-15}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 177819865319865 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 604587542087542 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 894360269360269 \cdot 10^{-17}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 304082491582492 \cdot 10^{-16}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge vowel(c_4) \wedge \sim d] = 23253367003367 \cdot 10^{-15}$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge d] = 0$
$P[\sim even(c_1) \wedge \sim even(c_2) \wedge \sim vowel(c_3) \wedge \sim vowel(c_4) \wedge \sim d] = 790614478114478 \cdot 10^{-16}$