has made some progress applying probabilistic models to individual differences (e.g., category learning; Navarro et al. 2006) and cognitive development (e.g., causal reasoning; Sobel et al. 2004). This work represents a step in the right direction; however, we expect that no single model can predict reasoning performance equally well across age groups and levels of experience. Indeed, systematic variations in peoples' behavior suggest that several different models (or modifications of a given model) may be required to explain developing behavior (Shultz 2007). Nevertheless, investigating differences between the models across age groups and skill levels may help us to understand more exactly "what differs" between and "what develops" within individuals.

In closing, we must emphasize O&C's comment that probabilistic models are often only functional level theories that should not be confused with algorithmic level theories (process models). Brighton and Gigerenzer (2008) have pointed out in their discussion of the limits of Bayesian models of cognition that the question of why the human mind does what it does (functional level) cannot be separated from the question of how the human mind does it (algorithmic level). Therefore, it is crucial that future Bayesian rational analyses specify how exactly their functional level models constrain theorizing about cognitive processes. This issue is especially relevant as the data connecting development, expertise, working memory, and reasoning imply that multiple strategies (and therefore processes) are at play. Though Bayesian rationality seems to provide a functional level account of prototypical adult reasoning, the development of cognitive capacities and expertise remains underappreciated.

ACKNOWLEDGMENTS
Preparation of this commentary was supported in part by a stipend from the International Max Planck Research School LIFE to NS. We are grateful to Henrik Olsson for comments on a previous draft.

How do individuals reason in the Wason card selection task?

doi:10.1017/S0140525X09000508

Eric-Jan Wagenmakers
Department of Psychological Methods, University of Amsterdam, Amsterdam 1018 WB, The Netherlands.
EJ.Wagenmakers@gmail.com
http://users.fmg.uva.nl/ewagenmakers/

Abstract: The probabilistic approach to human reasoning is exemplified by the information gain model for the Wason card selection task. Although the model is elegant and original, several key aspects of the model warrant further discussion, particularly those concerning the scope of the task and the choice process of individuals.

In the book *Bayesian Rationality* (Oaksford & Chater 2007, henceforth *BR*), Oaksford & Chater (O&C) present a summary and a synthesis of their work on human reasoning. The authors argue that formal logic and deduction do not explain how people reason in everyday situations. The deficiencies of the most simple forms of logic are obvious when one considers that they may assign "true" to absurd statements such as "if the moon is blue, than cows eat fish" (*BR*, p. 70). More importantly, the authors propose that, in contrast to formal logic, probability calculus does provide the right tools for an analysis of human reasoning. Thus, the authors argue that people solve deductive tasks by inductive methods. From this perspective, human reasoning can be characterized as Bayesian or rational.

Consider the Wason card selection task discussed in Chapter 6. Participants are confronted with four cards, showing an A, a K, a 2, and a 7. Participants are told that each card has a number on one side and a letter on the other. They are given a rule, "if there is an A on one side, then there is a 2 on the other side," and subsequently, have to select those cards that need to be turned over to assess whether the rule holds true or not. A moment's thought reveals that the cards that need to be turned over are the A card and the 7 card. Yet, the majority of participants do not choose the 7 card, but tend to choose the 2 card instead.

O&C propose an elegant Bayesian model – the information gain model – to account for people's performance in the Wason task. According to the model, people select the cards that reduce their expected uncertainty the most. Specific assumptions about the rarity of the information on the cards lead to the conclusion that selection of the 2 card might be rational after all.

The information gain model has been subjected to intense scrutiny (e.g., Oberauer et al. 1999). For non-experts, the details of this discussion are somewhat difficult to follow. A useful guideline is that a model should only be abandoned when it can be replaced with something better. And – criticisms raised against the information gain model notwithstanding – I have not come across a model that does a better job explaining how people make their card selections.

Despite its simplicity and elegance, some important details of the information gain model were not clear to me. First, O&C argue, on page 210, that their account only holds if participants regard the cards as a sample from a larger population. Perhaps the authors could spell out this argument in a bit more detail. Taking probability as a reflection of degree of belief, I did not immediately see what calculations are in need of adjustment. Second, the authors mention that participants who realize that the cards are *not* sampled from a larger population would always choose the A card and the 7 card. I do not know whether this prediction has been tested empirically, but I find it only slightly more plausible than cows eating fish. Note that in the Wason task a substantial proportion of participants do not even select the A card.

Another issue that warrants closer examination is the way the model's predictions relate to the data. In the information gain model, each card reduces the expected uncertainty to some extent. Why then does an individual participant not select all four cards, but generally only selects one or two? In other words, it was unclear to me how the model, from a consideration of expected uncertainty reduction, can predict card selections for an individual participant.

A fourth point concerns the role of individual differences. As the authors discuss on page 211, a subgroup of undergraduate students with high intelligence (about 10%) do select the A card and the 7 card. This result strengthened my initial belief that a motivated, intelligent person would always choose the A and 7 cards, when given sufficient time. In the spirit of falsification, I then tested this assumption on a colleague, who of course immediately selected the A and 2 cards. Perhaps she was not sufficiently motivated to think the problem through carefully – would incentives of time or money increase the selection of the 7 card?

O&C are to be admired for their principled approach to quantitative modeling, and for their courage to take on the unassailable dogma of human irrationality. It is unfortunate that much of the material in the book was already available elsewhere (e.g., Oaksford & Chater 2001; 2003b); therefore, it was not entirely clear to me what the book adds to our current knowledge base.

One final comment. It strikes me as paradoxical that researchers who argue for a coherent, rational approach to human reasoning then proceed to apply an incoherent, irrational approach to the statistical analysis of their experimental data. Throughout the book, the authors renounce Popper's stance on the importance of falsification, arguing that this is not how science works, nor how people reason. But then, in the very same work, the authors measure the validity of their models by means of *p*-values, and include statements such as "the model could not be rejected." Why?

Oaksford and Chater (2008) found that the revised model presented in *BR* may provide better fits to Oberauer's data. Second, Oberauer argues that the most relevant empirical evidence comes from studies where probabilities were directly manipulated, of which he mentions two, Oaksford et al. (2000) and Oberauer et al. (2004). Moreover, he argues that their results are equivocal. However, several other studies have manipulated probabilities in conditional inference and found evidence in line with a probabilistic account (George 1997; Liu 2003; Liu et al. 1996; Stevenson & Over 1995). Oberauer also leaves aside the many studies on data selection showing probabilistic effects (see *BR*, Ch. 6).

Liu's arguments about second-order conditionalization point, we think, to an important factor that we have yet to consider in reasoning, that is, the effects of context. **Liu** has found that people often endorse the conclusion that, for example, *Tweety flies* on being told that *Tweety is a bird* in the absence of the conditional premise (reduced problems). This occurs because they fill in this information from world knowledge. However, **Liu** also found that endorsements increase when the conditional premise is added (complete problems). In *BR*, we argued that this occurs because people take the conditional premise as evidence that the conditional probability is higher (an inference that may arise from conversational pragmatics). **Liu** argues that our account implies that manipulations affecting reduced problems should also affect complete problems and provides evidence against this. Yet context, both cognitive and physical, may explain these differences in a way similar to recent studies of decision-making (Stewart et al. 2006). For example, suppose one is told about two swanneries, both containing the same number of swans. In one, 90% of swans are black ($P(\text{black}|\text{swan}) = .9$); in the other, 90% of swans are white ($P(\text{white}|\text{swan}) = .9$). On being told that *Tweety is a swan*, presumably one would only endorse *Tweety is white* at .5. This is because conversational pragmatics and world knowledge indicate that Tweety is in one of the just mentioned swanneries, but the dialogue up to this point does not indicate which one.[5] However, the addition of the conditional premise *if a bird is a swan it is white* immediately indicates which swannery is being talked about, that is, the one in which $P(\text{white}|\text{swan})$ is high, and now endorsements should increase to .9. Clearly, although manipulations of the relative number of swans in each swannery might affect the reduced problem, they should not affect the complete problem. So if the swannery in which most swans are black were one tenth of the size of the other swannery, then, given natural sampling assumptions, endorsements for the reduced problem should increase to .83, but endorsements of the complete problem should remain the same.

R5.2. Data selection

Wagenmakers raises a variety of concerns about our optimal data selection model. First, why do we concede that people should select the standard "logical" A card and 7 card choices, *if* the rule only applies to the four cards? In *BR* (p. 210), we argue that people rarely use conditionals to describe just four objects – they assume that the cards are drawn from a larger population. Consequently, we quite explicitly do not make the counterintuitive prediction that Wagenmakers ascribes to us. Second, Wagenmakers wonders why – when all cards carry some information – do participants not select all the cards, if they are maximizing information gain? We assume that the pragmatics of the task suggests to participants that they should select some cards, but not others (*BR*, pp. 200–201). Third, Wagenmakers suggests that incentivized individuals with more time might make the logical response. Work on individual differences (e.g., Stanovich & West 2000) is consistent with the view that logical competence is learned, either directly (e.g., studying logic or math) or indirectly (e.g., learning to program or learning conventional, non-Bayesian statistics); such logical competence is a prerequisite for "logical" responses, and covaries with IQ as measured in University populations. Wagenmakers also remarks that, as Bayesians, we should avoid null hypothesis testing in statistically assessing our models. This choice is purely pragmatic: it conforms to the current demands of most journals.

R5.3. Syllogisms and development

Halford argues that mental models theory and a relational complexity measure fit the data as well as the probability heuristics model (PHM), conceding, however, that only PHM generalizes to *most* and *few*. Copeland (2006) has also recently shown that PHM provides better fits than mental models and mental logic for extended syllogisms involving three quantified premises. Halford also suggests that basing confidence in the conclusion on the least probable premise, as in our *max*-heuristic, is counterintuitive. He proposes that confidence should instead be based on relational complexity, which covaries with the least probable premise. But perhaps Halford's intuition goes the wrong way: the least probable premise is the most informative; and surely the more information you are given, the stronger the conclusions you can draw?

De Neys and **Straubinger, Cokely, & Stevens** (**Straubinger et al.**) both argue that there are important classes of evidence that we do not address. De Neys argues that attention to latency data and imaging studies provides a greater role for logic, a claim we disputed earlier. Note, also, that the algorithmic theory in PHM has been applied to latency data and accounts for the data, as well as mental models (Copeland & Radvansky 2004). Straubinger et al. are concerned that we ignore developmental data. In particular, they view the findings on the development of working memory as providing a particular challenge to a Bayesian approach. They do, however, acknowledge that in different areas (e.g., causal reasoning), Bayesian ideas are being successfully applied to developmental data (Navarro et al. 2006; Sobel et al. 2004). Straubinger et al.'s emphasis on working memory provides good reason to believe that our particular approach to deductive reasoning may extend to development. Copeland and Radvansky (2004) explicitly related working-memory limitations to PHM, finding that it provided as good an explanation as mental models theory of the relationship between working-memory capacity and reasoning performance. This result provides some indication that, at least for syllogistic reasoning, developmental trajectories explicable by mental models may be similarly amenable