Journal of Experimental Psychology:
Learning, Memory, and Cognition
2016, Vol. 42, No. 1, 158–166

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000165

COMMENTARY

Modeling Valuations From Experience: A Comment on Ashby and Rakow (2014)

Dirk U. Wulff and Thorsten Pachur
Max Planck Institute for Human Development, Berlin, Germany

What are the cognitive mechanisms underlying subjective valuations formed on the basis of sequential experiences of an option's possible outcomes? Ashby and Rakow (2014) have proposed a sliding window model (SWIM), according to which people's valuations represent the average of a limited sample of recent experiences (the size of which is estimated by the model) formed after sampling has been terminated (i.e., an end-of-sequence process). Ashby and Rakow presented results from which they concluded, on the basis of model-selection criteria, that the SWIM performs well compared with alternative models (e.g., value-updating model, summary model). Further, they reported that the individual window sizes estimated by the SWIM correlated with a measure of working-memory capacity. In a reanalysis of the Ashby and Rakow data, we find no clear evidence in support of any of the models tested, and a slight advantage for the summary model. Further, we demonstrate that individual differences in the window-size estimated by the SWIM can reflect differences in noise. In computer simulations, we examine the more general question of how well the models tested by Ashby and Rakow can actually be discriminated. The results reveal that the models' ability to fit data depends on a complex interplay of noise and the sample size of outcomes on which a valuation response is based. This can critically influence model performance and conclusions regarding the underlying cognitive mechanisms. We discuss the implications of these findings and suggest ways of improving model comparisons in valuations from experience.

Keywords: valuations from experience, active sampling, cognitive modeling, model complexity, monetary gambles

When a valuation of an object is formed on the basis of sequential experiences, not all experiences with the object necessarily contribute equally to the valuation. Instead, more recent experiences tend to have a stronger impact than do less recent ones (e.g., Hogarth & Einhorn, 1992). For instance, imagine that five draws from a lottery with an initially unknown payoff distribution have yielded the following sequence of outcomes: € 2.00, € 2.00, € 0.50, € 2.00, € 0.50. When asked to judge the value of the lottery, respondents often seem to give disproportionate weight to outcomes sampled at the end of the sequence. How can experience-based valuations, in particular such potential *recency* effects, best be modeled? A common assumption is that more distant experiences have a gradually decreasing influence, in line with the idea that memory traces decay with time (Ebbinghaus, 1885/1913). A prominent instantiation of this notion in decision research is the value-updating model (VUM; Hertwig, Barron, Weber, & Erev, 2006), according to which a valuation v after experiencing n outcomes (with x_n being the most recent one) is determined as follows.

$$v_n = \left[1 - \left(\frac{1}{n}\right)^{\varphi}\right] v_{n-1} + \left(\frac{1}{n}\right)^{\varphi} x_n. \quad (1)$$

The parameter φ either gives more weight to earlier samples ($\varphi > 1$, *primacy*) or to later samples ($\varphi < 1$, recency), or weights all samples equally ($\varphi = 1$).

Ashby and Rakow (2014) recently proposed an intriguing alternative way of modeling a stronger influence of more recent experiences. Rather than assuming a gradually decreasing impact, they postulated an all-or-nothing mechanism that considers all experiences in a recent window and excludes more distant experiences. Specifically, the sliding window model (SWIM) proposes that a valuation v_n is formed by averaging ζ out of n total experiences x_i:

$$v_n = \frac{1}{\zeta} \sum_{i=1+n-\zeta}^{n} x_i \quad (2)$$

If the size of the window, ζ, is smaller than the total number of experiences n, the SWIM implements recency effects by assuming that some (specifically, $n - \zeta$) of the earliest experiences are

Dirk U. Wulff and Thorsten Pachur, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

We are grateful to Susannah Goss for editing the manuscript. We also thank Nathaniel Ashby and Tim Rakow for providing us with the raw data of their experiments.

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: wulff@mpib-berlin.mpg.de

excluded from consideration; within the window, all experiences contribute equally to the valuation. This account of recency is structurally consistent with models of working memory that posit a fixed, limited storage capacity, and in which an item is currently either activated in memory or not (e.g., Cowan, 2001).

Ashby and Rakow (2014) argued that the SWIM should be interpreted as an *end-of-sequence mechanism* the evaluation of which is formed only after the sampling process has been terminated. This deviates from the assumption in the VUM (and other models; e.g., Hogarth & Einhorn, 1992; March, 1996) of a *step-by-step mechanism*, the valuation of which is formed and continually updated online during the sampling process.

In two empirical studies, Ashby and Rakow (2014) pitted the SWIM against the VUM as well as the summary model (SUM; Hills & Hertwig, 2010; Wulff, Hills, & Hertwig, 2012), which calculates an average across all experiences (thus assuming no recency). In both studies, participants were presented with a total of 40 lotteries, each of which had two randomly generated outcomes, one high and one low, ranging between £ 0.39 and £ 4.00. The possible outcomes of each lottery and their probabilities were initially unknown to the participants, but they could sample (i.e., take random draws) from the payoff distribution. Participants could draw as many samples as they wanted up to 100. After terminating sampling each round, they provided a valuation for the lottery. The valuation was incentivized using the Becker–DeGroot–Marschak procedure (Becker, DeGroot, & Marschak, 1964).[1]

From their analyses of participants' valuation responses, Ashby and Rakow (2014) concluded "that for many individuals not all information is used and that the amount of information integrated is, in part, related to individual differences in cognitive abilities such as memory span" (p. 1160). This conclusion was mainly based on the findings that (a) the SWIM showed a better average fit on the Akaike information criterion (AIC; Akaike, 1973), (b) the window-size estimated for individual participants using the SWIM was consistently smaller than the average number of samples the participant had drawn, and (c) there was a higher correlation between sample size and response time for people better fit by the SWIM than for those better fit by the VUM or SUM, in line with the assumption that the SWIM implements an end-of-sequence process (which predicts that the larger the number of experiences that can be retrieved, the longer the response should take). Further, to validate that the SWIM's window size reflects the number of experiences processed, the authors analyzed the relationship between each participant's window size as estimated by the SWIM and a measure of working memory capacity.[2]

The SWIM represents an attractive addition to the growing literature on models of experienced-based judgment and decision making (e.g., Hertwig & Erev, 2009), and the empirical findings presented by Ashby and Rakow (2014) are intriguing. Moreover, the proposal that sampled outcomes are processed in an all-or-nothing fashion has elegant conceptual similarities with the assumption of limited capacity in prominent conceptions of working memory (Baddeley, 2012). In this article, however, we argue that the current evidence may not warrant the conclusion that experience-based valuations are based on an all-or-nothing, end-of-sequence evaluation process, as embodied in the SWIM. In fact, our analyses show that the common setup used by Ashby and Rakow—and in investigations into decisions from experience in general (e.g., Hertwig & Erev, 2009)—is problematic for distinguishing between candidate mechanisms and needs to be improved.

The article is structured as follows. In the first part, we critically reevaluate the model comparison conducted by Ashby and Rakow (2014) and find that if the data do support one particular model, it is the SUM (which assumes no recency effect) rather than the SWIM. We then illustrate that recency as estimated by the SWIM may be confounded with the amount of noise in the valuation process. It is therefore unclear whether individual differences in window-size estimated by the SWIM indeed specifically reflect the amount of information considered. This result complicates the interpretation of Ashby and Rakow that correlations between the SWIM estimates of window size and working memory capacity would support the specific processes assumed by the model. We also examine the assumption of an end-of-sequence process on a conceptual level, arguing that the requirement to conduct and terminate search actively in itself necessitates some form of step-by-step process, which is at odds with the assumption of a pure end-of-sequence process (as embodied in the SWIM).

In the second part, we turn to the more general question of how well the different models of valuations from experience tested by Ashby and Rakow (2014)—the SUM, VUM, and SWIM—can actually be recovered from data. Specifically, we examine the models' relative ability to fit data for different levels of sample size and noise. The analysis reveals that the models' performance depends on a complex interplay of these factors, which in many cases can impair the recovery of the mechanism that actually generated the data. We end by discussing implications of these results for the study of valuations from experience and proposing ways to improve investigations of the cognitive mechanisms underlying experience-based decision making.

Reanalysis of Ashby and Rakow (2014)

Model Comparison

A key basis for Ashby and Rakow's (2014) conclusion regarding the viability of the SWIM was its performance in a model comparison that pitted it against the VUM and the SUM. The authors used two popular measures to evaluate the three models: the Bayesian information criterion (BIC; Schwarz, 1978) and the AIC. Both indices penalize for model complexity based on the number of free parameters.[3] Two aspects of model performance were considered: the number of participants best accounted for by each model (according to BIC and AIC), and each model's median

[1] Specifically, participants were informed that their valuation of the lottery would be compared with a randomly drawn value between £0.00 and £4.00. If the value drawn was greater than or equal to their valuation, they would receive that value; otherwise the gamble would be played out and they would receive the resulting outcome.

[2] Note that evidence for a positive relationship was claimed in the initial report, but that in a later correction ("Correction to Ashby & Rakow," 2014), the authors clarified that the correlation was in fact negative.

[3] The BIC approximates the marginal likelihood of the data, given a specific model, and the AIC approximates the Kullback–Leibler divergence between the true and the evaluated model. Which of the two measures is to be preferred is debated (for overviews, see Burnham & Anderson, 2002; Lewandowsky & Farrell, 2010; Vrieze, 2012).

and mean (across participants) BIC and AIC values. According to BIC, the VUM and the SUM performed best in Studies 1 and 2, respectively, in terms of the number of participants they best accounted for. According to AIC, the SWIM accounted for the largest number of participants (in Study 2; see also "Correction to Ashby and Rakow," 2014). In terms of the average BIC and AIC values, the SWIM emerged as the best model.

For several reasons, however, the model-testing approach used by Ashby and Rakow (2014) has limited power to support conclusions about the viability of the mechanism assumed in the SWIM: First, the formal specifications of the tested models differed in aspects beyond the implementation of recency, making it difficult to unequivocally identify the source of differences in model performance. Specifically, a linear value function was assumed for outcomes in the SWIM, but a nonlinear value function for outcomes in the VUM (i.e., x^{α} with $\alpha = .88$). Moreover, the VUM was allowed to accommodate both recency and *primacy* effects, whereas the SWIM could accommodate only recency effects. Second, the models were not fit with equal precision. For the VUM and the SWIM (which have two free parameters), Ashby and Rakow used a two-stage fitting procedure. In a first step, they estimated a noise parameter (implemented as the standard deviation of a normal distribution of errors; see Appendix A) with the SUM as the underlying model; they then estimated the recency parameter and determined the overall maximum likelihoods of the models based on the estimate obtained in the first step. This procedure ignores the interplay between parameters (e.g., Scheibehenne & Pachur, 2015) and may lead to nonoptimal estimates, thus giving the SUM (which has only one parameter) an edge. Third, although the possible outcomes of the lotteries used by Ashby and Rakow as well as participants' valuations were confined to the interval between 0 and 4, the models were allowed to produce valuations going beyond this range, as an untruncated (rather than a truncated) error distribution was used in the models. As a consequence, parameter values giving predictions close to the boundaries (i.e., close to 0 or 4) are given a low likelihood (as greater portions of the probability mass are cut off by the boundaries); this might distort the parameter estimates and model comparisons.

We reanalyzed the data of Ashby and Rakow (2014) to see how the different models fared when a more appropriate methodological approach is used and, in addition, the models are equated with respect to all aspects except for how recency is implemented. Specifically, (a) we used a linear value function for all models, (b) we fitted the VUM such that, for greater comparability with the SWIM, it could accommodate only recency by constraining $0 \leq \varphi \geq 1$; we refer to this version of the model as VUMr, (c) for the VUMr and the SWIM, we estimated both model parameters simultaneously, and (d) we used a truncated normal distribution to model noise (see Appendix A). Further, the parameter estimation was based on a combination of grid search and subsequent optimization using quasi-Newton minimization, and we used the AIC_c (AIC corrected) rather than the AIC for model comparison (Burnham & Anderson, 2002).[4]

Figures 1 a–c plot the performance of one model (in terms of AIC_c) against the other (aggregated across both studies in Ashby & Rakow, 2014), separately for all three pairwise comparisons of the SUM, the VUMr, and the SWIM (the results for BIC values are qualitatively the same). Each point represents a participant. As can be seen, model fits for three participants were far better than those for most other participants (specifically, the fits for the VUMr and the SWIM). As these three participants might have displayed response patterns diverging from those of the other participants, we excluded them for all following analyses (unless indicated otherwise).[5] Overall, the figure shows that the three models fitted the data similarly well. Accordingly, the median and mean BIC and AIC values reported in Table 1 barely differ across the SUM, VUMr, and SWIM. Nevertheless, as Table 1 shows, the percentage of participants best fit by the SUM was considerably higher than that best fit by the VUMr or the SWIM, in terms of both BIC and AIC_c. Figure 1 d–f plot the data against the predictions of the models separately for the SUM, the VUMr, and the SWIM. It can be seen that although the models capture general trends in the data (indicated by the fact that many data points cluster around the diagonal), for all three models there is also considerable misfit. Table 1 shows that according to the median (across participants) value of the noise parameter under each model (which corresponds to the root mean squared deviation of the model prediction from the data), the expected deviation of the model prediction from the data equaled about a sixth of the entire range of possible values. We will return to this issue of absolute model fit and the rather high estimates of noise in Ashby and Rakow's (2014) data.

In sum, a reanalysis of Ashby and Rakow (2014) using a more appropriate model-comparison approach yielded little evidence that one model consistently outperforms another. On the individual level the majority of participants was best captured by the SUM. Inconsistent with the conclusions drawn by Ashby and Rakow, therefore, if the data do support one particular model, it is the SUM, which considers all experiences in the sample and assumes no recency, rather than the SWIM. The relative superiority of the SUM must, however, be reevaluated in light of the analyses reported below.

Interpreting Noise as Forgetting?

Ashby and Rakow's (2014) argument that valuations from experience (sometimes) follow an all-or-nothing, end-of-sequence process that considers only a limited number of the outcomes experienced was also based on considerations beyond the results of the model comparison. First, the window sizes estimated by the SWIM were, on average, smaller than the number of samples drawn. Second, Ashby and Rakow proposed that the window sizes estimated for respondents classified as following the SWIM should be related to their working-memory capacity. Similarly, respondents following the SWIM should show a stronger positive rela-

[4] It has been shown that the AIC penalizes insufficiently for model flexibility with small sample sizes. For this reason, Burnham and Anderson (2002) have recommended the use of the AIC_c, which corrects for this bias (for large samples, AIC_c approaches AIC) and is defined as $AIC_c = -2ln(L) + 2k + \frac{2k(k+1)}{n-k-1}$, with L being the likelihood, k the number of parameters, and n the number of data points used to calculate the likelihood.

[5] Further analyses indicated that these three participants gave the last observed outcome as their final valuation in at least 95% of the lotteries (although they, on average, drew more than one outcome before making a valuation). Thus, these participants are likely to have applied a qualitatively different strategy than assumed by the SUM, VUM, or SWIM, which provides grounds for their exclusion.

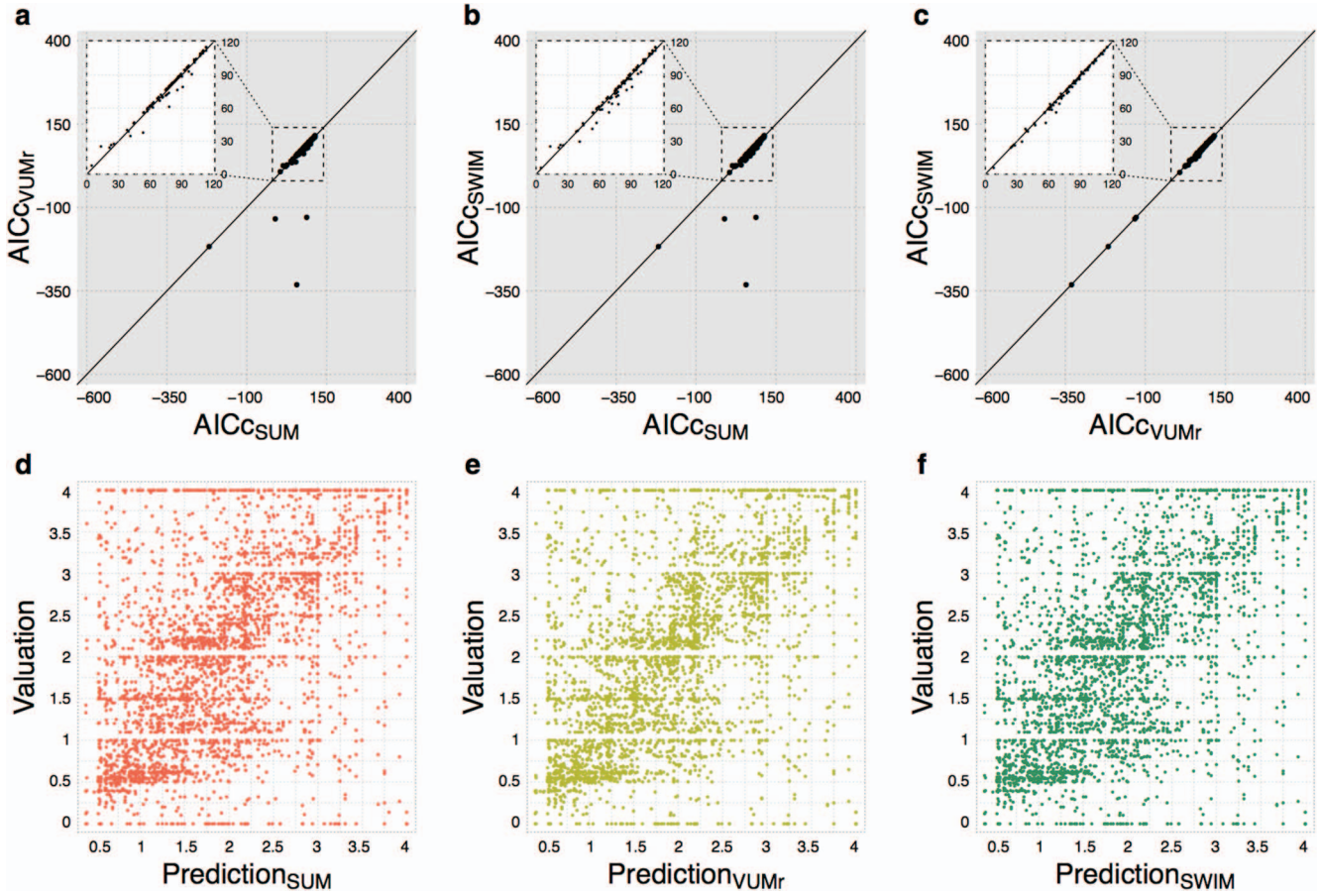

Figure 1. Top row: Scatter plots show pairwise model comparisons of individual participants' AIC_c values, a–c. The white square in the upper left shows the majority of participants on a more fine-grained scale. Bottom row: Scatter plots show the data as a function of the models' predictions, d–f. AIC_c = Akaike information criteria (corrected); SWIM = sliding window mode; SUM = summary model; VUM = value-updating model. See the online article for the color version of this figure.

tionship between window size and response time than respondents following another mechanism (e.g., the SUM). Together, these results would support the interpretation that SWIM respondents process as many samples as their working memory size permits at the end of the sequence. But note that a key assumption inherent in these considerations is that the window-size estimate of the SWIM veridically reflects the number of experiences on which the valuation is based.

To demonstrate the limitations of this assumption, let us consider a special case of the SWIM, where the window size of considered experiences matches the total number of samples drawn (i.e., $\zeta = n$). Under these circumstances, the SWIM necessarily underestimates the window size (on average), because the estimates (which in the presence of noise will always err to some extent) can err only in the direction of smaller window sizes, not in the direction of larger window sizes. This underestimation due to noise severely complicates tests of Ashby and Rakow's (2014) hypotheses. Specifically, to demonstrate that people relied on only a limited sample of recent experiences, it is necessary to show that window-size estimates fall outside the range that could result from noise alone. This was not the case for Ashby and Rakow's setup, however. In a simulation reported in Appendix B, in which the window size was set to be equal to the sample size under realistic levels of noise (matched to those observed by Ashby and Rakow), the resulting window-size estimates suggested use of 87% of the

Table 1
Model Fits Aggregated Over Studies 1 and 2 of Ashby and Rakow (2014)

	BIC			AIC_c			Noise
Model	% best	Mdn	*M*	% best	Mdn	*M*	Mdn
SUM	75	83.1	76.5	62	82.2	75.9	.79
VUMr	7	85.3	78.3	10	82.7	76.5	.76
SWIM	18	83	77.3	27	81.2	75.6	.76

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; SUM = summary model; VUMr = value-updating model–recency; SWIM = sliding window model. Shown are the number of participants best fit by each model (separately for BIC and AIC_c), the mean and median criterion value for each model, and the noise (σ) estimated from the models (for the truncated range between 0 and 4).

sampled outcomes (although all outcomes were in fact used). The estimated window sizes for the empirical data of Ashby and Rakow suggest an only minimally smaller window size, namely 86% of the samples. In sum, the finding that the estimated window size is smaller than the sample size does not warrant the conclusion that only part of the information was used (as assumed by the SWIM). Reduced window sizes (relative to the sample size) may simply result from noise.

Conceptual Issues With End-of-Sequence Processing in Self-Terminated Search

It is also instructive to consider the assumption of an end-of-sequence evaluation process from a conceptual perspective. In the sampling paradigm used by Ashby and Rakow (2014), participants sequentially drew samples from an initially unknown payoff distribution, and it was up to them to decide when to stop sampling. Should one expect respondents to construct a valuation only after sampling has been terminated—as predicted by a strict interpretation of an end-of-sequence process? Note that this would imply that the decision of how many samples to draw is unrelated to (and unaffected by) the outcomes sampled. Based on the evidence presented by Ashby and Rakow, however, this does not seem very plausible. In particular, sample size was found to be (positively) correlated with the variance of the lotteries (for similar findings in the sampling paradigm, see Lejarraga, Hertwig, & Gonzalez, 2012; Pachur & Scheibehenne, 2012). Clearly, for sampling effort to be sensitive to the characteristics of individual lotteries (e.g., their variances), some form of online processing has to occur.

Another reason that speaks against end-of-sequence processing is Ashby and Rakow's (2014) finding that, overall, the data in the sampling paradigm display a recency effect (i.e., that more weight is given to more recent experiences; see also Pachur & Scheibehenne, 2012; Wulff, Hills, & Hertwig, 2014). Reviewing studies with paradigms that encouraged end-of-sequence processing (i.e., people were presented with a sequence of evidence and asked for an evaluation at the end of the sequence), Hogarth and Einhorn (1992) reported that 34 of 54 studies (63%) showed a primacy effect, not a recency effect. Recency effects, by contrast, were found predominantly in studies explicitly enforcing step-by-step processing (in 20 of 22 studies; 91%). If the participants in Ashby and Rakow's studies had indeed relied on an end-of-sequence process, one would—on the basis of these findings—therefore have expected a primacy effect to occur. However, Ashby and Rakow's estimates of the VUM's φ parameter (which allows serial-position effects to be measured) yielded evidence for recency, not for primacy.

Given that, in the sampling paradigm people were not explicitly instructed to conduct step-by-step processing, one might ask what led them to rely on such a process. One possibility is that step-by-step processing is triggered by the requirement in the sampling paradigm for the respondent to actively decide when to stop sampling (note that in typical end-of-sequence studies, participants are presented with a sequence of outcomes of fixed length—that is, they do not have to decide when to terminate search; Hogarth & Einhorn, 1992).

Interim Summary

First, a more appropriate approach to the model evaluation conducted by Ashby and Rakow (2014) provided no clear evidence in support of any of the models tested. Second, under noise, the window size estimated by the SWIM substantially undershot the actual number of samples drawn, casting doubt on the conclusion that reduced window sizes necessarily reflect forgetting or constraints of working-memory capacity. Third, the notion of a strict end-of-sequence process advocated by Ashby and Rakow is inconsistent with the finding of variance-sensitive sampling in both the authors' data and in other investigations (e.g., Pachur & Scheibehenne, 2012), and with a large body of research on belief updating (Hogarth & Einhorn, 1992).

Modeling Valuations From Experience

How Well Can the Models Be Recovered?

To shed light on the cognitive processes underlying valuations from experience, Ashby and Rakow (2014) fit three models to their data: the SUM, the VUM, and the SWIM. One important condition for this approach to be useful is that the models can actually be recovered (if they match the data-generating mechanism). In the following, we examine the degree to which the SUM, VUM, and SWIM can actually be recovered based on the methodological setup used in Ashby and Rakow (2014). Accurate model recovery and discrimination between models are possible only if the true process is not clouded by an excessive amount of noise. Further, the models often have to make different predictions for the data at hand. The latter may actually be quite a challenge in valuations from experience. For instance, people's active search in the sampling paradigm can lead to nondiagnostic data when search is terminated before more than two different outcomes have been observed. If search is stopped after only one outcome has been sampled, then, no matter how models are compared, it is impossible to distinguish them. In Ashby and Rakow (2014), who used two-outcome lotteries, only one type of outcome was observed in 34% of all trials (across all participants; note that these trials were included in Ashby and Rakow's and our analyses). But even if a larger number of different outcomes are observed, the sequence of samples can still make it impossible to discriminate between models—for instance, when outcomes are relatively evenly distributed across the sequence of sampled outcomes.

Moreover, model recovery depends on the relative flexibility of the models under consideration. Importantly, in valuations from experience, flexibility can be a function of the sample size on which the valuations are based; note that in the sampling paradigm the number of samples drawn is determined by the respondent. For instance, in the simple case of just two observed values (e.g., 1, 3), the VUMr can perfectly fit any valuation between 2 and 3 by shifting the φ parameter between 0 and 1. The SWIM, in contrast, can only predict exactly two different valuations, namely 2 ($\zeta = 2$) and 3 ($\zeta = 1$). For very small sample sizes, the VUMr can thus be expected to perform much better in fitting empirical data than the SWIM. For large sample sizes, however, the reverse may be true. The VUMr always considers all samples, which may place a large restriction on the range of valuations it can fit. Because it can ignore entire subsets of the data, the SWIM may then prove to be more flexible. Note that, in contrast to

nondiagnostic data, differences in flexibility may lead to the systematic recovery of a wrong model.

To examine the extent to which the SUM, VUMr, and SWIM can nonetheless be correctly recovered, we conducted a model-recovery analysis based on the setup of Ashby and Rakow (2014). Specifically, we simulated 1,000 agents for every combination of 20 levels of sample size and 20 levels of noise ($2 \leq$ sample size ≥ 40, $.2 \leq \sigma \geq 1.1$, both covering 95% of the values observed by Ashby and Rakow), separately for each of the models as the generating process. Each simulated agent completed the valuation task consisting of 40 lotteries. Given the empirical finding that window and sample sizes are strongly correlated (Ashby & Rakow, 2014), for the SWIM, the window size producing the valuation was determined as a fixed fraction of the agent's sample size (namely ζ/n). For the VUM and the SWIM, we assumed levels of recency that matched the median parameter estimates obtained from the Ashby and Rakow data (VUMr: $\varphi = .77$; SWIM: $\zeta/n = .73$). We then fitted the SUM, VUMr, and SWIM to each simulated agent's data, and determined their relative performances in terms of model weights based on AIC_c and BIC (e.g., Wagenmakers & Farrell, 2004).[6]

Figure 2 shows median model weights for the three models when the SUM (upper panel), the VUMr (middle panel), or the SWIM (lower panel) was the generating mechanism, separately for the different levels of noise (represented on the x axis). The line types and transparent shapes illustrate the effect of the sample size on which a valuation is based. The dashed and solid lines represent mean sample sizes of 2 and 40, respectively. The transparent shapes illustrate the range (minimum to maximum) of median model weights for mean sample sizes between 4 and 38. The white background highlights the range that covers 90% of the reported noise levels for the participants in the two studies by Ashby and Rakow (2014).

As can be seen from Figure 2, there are many misclassifications—that is, situations in which a model obtained the highest model weight even though it had not generated the data. Generally, and not surprisingly, model recovery for all three data sets becomes less accurate at higher levels of noise. However, there is a systematic trend in the misclassifications such that it is particularly the SUM that emerges as the best-fitting mechanism when noise is high. This is because the SUM has fewer parameters than the VUMr or the SWIM. In light of this result, the (slight) advantage of the SUM in our reanalyses of Ashby and Rakow's (2014) empirical data reported above should be treated with caution.

Moreover, model recovery seems to depend on sampling behavior. For the vast majority of conditions in our simulation, the SWIM appears to be more flexible than the VUMr, particularly when the sample size of observations is large. This is evidenced by the fact that valuations generated by the VUMr are more likely to be incorrectly attributed to the SWIM than vice versa.[7]

Finally, as indicated by the relative position of the dashed and solid lines (and the width of the transparent shapes) in Figure 2, for all models, recovery proves to be considerably better for small than for large sample sizes. This suggests that inaccurate model recovery is not primarily due to having nondiagnostic data (which are more likely to occur for small samples). One explanation for the perhaps counterintuitive finding that model recovery is worse with larger sample sizes is that all models converge, with larger sample size, toward predicting the expected value of the lottery (because all three models under consideration differ only in terms of the number or relative weighting of observations considered).

Overall, the model-recovery analysis shows that—unless the error level is very low—correct model recovery and hence discrimination between the SUM, VUMr, and SWIM is difficult, and that the relative performance of each model depends to a considerable degree on the amount of noise and the sample size. These results suggest that the relative performance of a model in an empirical study (in which the respondent plays an active role in determining the amount of sampling) also depends on factors other than the underlying mechanism. This can severely complicate the interpretation of model-comparison analyses as conducted by Ashby and Rakow (2014).

Implications and Suggestions

To the extent that high levels of noise, as estimated by the models (see Table 1), reflect haphazard responses on the part of participants, one approach to improve model discriminability could be to encourage more systematic behavior. For instance, it is important to ensure that participants fully understand the Becker-DeGroot-Marschak procedure (Becker et al., 1964), which is commonly used to establish incentive compatibility in valuation studies (e.g., by Ashby & Rakow, 2014). Several authors have argued that respondents do not always fully comprehend the procedure, which may lead to noisy behavior (James, 2007; Plott & Zeiler, 2005; Safra, Segal, & Spivak, 1990). Another avenue would be to improve ways in which noise is modeled. For instance, future studies might consider alternative noise distributions (e.g., a t or β distribution) that are more robust against outliers, or formalizations of noise where the response error is sensitive to the characteristics of a lottery (e.g., its variance), rather than assuming a constant error (e.g., Carbone & Hey, 2000).

However, a high level of noise may also indicate that the models fail to capture substantial aspects of the data. Future work might thus also consider alternative models of valuations from experience. One approach could be to consider elements from prospect theory (e.g., Tversky & Kahneman, 1992; for an example in the context of experience-based decisions, see Ahn, Busemeyer, Wagenmakers, & Stout, 2008). Another could be to consider alternative implementations of recency, for instance, by modeling valuations of experience based on principles of ACT-R (Anderson & Lebiere, 1998; for applications to decisions from experience, see, e.g., Gonzalez & Dutt, 2011).

Active sampling hampers the recovery of the underlying processes in valuations from experience by affecting the diagnosticity of the data. If a person samples only a few times, the chances are high that only one of the lotteries' outcomes is observed, which renders the models' predictions identical and their discrimination impossible. Problems also arise, however, if a person samples many times, because the SUM, VUMr, and SWIM converge in their predictions when sample size is large. To address these issues, we see three avenues for improving experimental designs testing models of valuations from experience. First, researchers

[6] Model weights are defined as $w_M = \frac{e^{-\frac{1}{2}\Delta \mathrm{crit}_M}}{\sum_i e^{-\frac{1}{2}\Delta crit_i}}$, where $\Delta crit$ is the difference between model M and the best-performing model (in the set of competing models) on the respective information criterion (i.e., AIC_c or BIC; see Lewandowsky & Farrell, 2010).

[7] The only exceptions occurred with a sample size of 2, due to the fact that the VUMr (with $\varphi = 1$ or $\varphi = 0$) can perfectly mimic the two possible predictions of the SWIM for a sample size of 2.

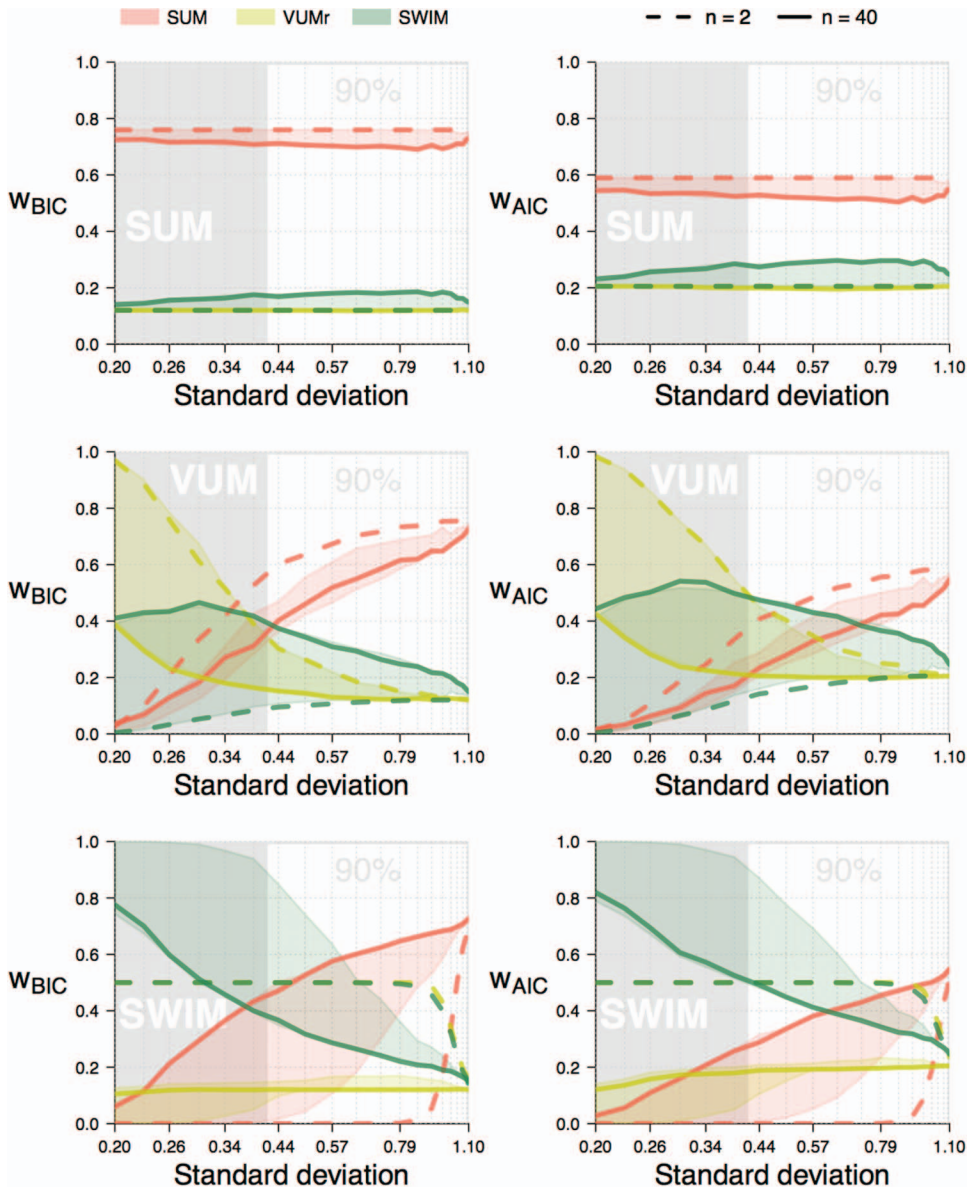

Figure 2. Model recovery as a function of noise (*SD*) and sample size of the observations on which a valuation is based. Panels in the upper, middle, and lower rows show the fits for data generated by the SUM, VUMr, and SWIM, respectively. Panels on the left show fits expressed as median-model weights based on the BIC; panels on the right show the fits based on the AIC_c. The solid lines indicate the performance for a sample size of 2, the dashed lines for a sample size of 40. The transparent shapes show the range of performance for mean sample sizes between 4 and 38. AIC = Akaike information criteria; BIC = Bayesian information criteria; SWIM = sliding window mode; SUM = summary model; VUM = value-updating model. See the online article for the color version of this figure.

could use a larger number of lotteries, as this will increase the chance of observing diagnostic sequences. Second, instead of using the common two-outcome lotteries, researchers could use multi-outcome lotteries, or even lotteries with continuous outcomes (Wulff et al., 2014). Third, by providing an incentive to sample more, researchers could decrease the chance that only one type of outcome is observed (see Hau, Pleskac, Kiefer, & Hertwig, 2008). Note, however, that excessive sampling will also decrease model discriminability; these efforts therefore need to be well balanced. To evaluate and improve the diagnosticity of different design variants several formal approaches have been developed (e.g., landscaping, Navarro, Pitt, & Myung, 2004; optimal design, Myung & Pitt, 2009).

Our results have shown that differences in sample size due to active sampling also impact the relative flexibility of the models. Constant punishment terms, as implemented in the AIC and BIC, are incapable of accounting for this. Less constrained model selection criteria, such

as marginal likelihood or normalized maximum likelihood (e.g., Myung, Navarro, & Pitt, 2006), or evaluations based on out-of-sample prediction (Erev et al., 2010) might therefore be considered. Note that any assessment of the appropriateness of a given model should also include consideration of absolute model fit (e.g., Heathcote, Brown, & Wagenmakers, 2015).

Conclusion

Ashby and Rakow's (2014) SWIM presented an interesting alternative to existing approaches to model recency in valuations from experience. We welcome their proposal, as it emphasizes the overdue need to develop and rigorously compare models of experience-based judgment and decision making. Our analyses suggest, however, that evidence for the SWIM (or any of the other models) must currently be considered limited and that the validity of the model's key parameter (i.e., its window-size estimate) may be compromised. We identified a number of general challenges in current approaches to testing models of valuations from experience and highlighted several methodological and conceptual issues that warrant consideration. We hope that future research will take up these challenges and suggestions, as valuations based on active information search are ubiquitous in real-world decision making.

References

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E. J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32,* 1376–1402. http://dx.doi.org/10.1080/03640210802352992

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1153–1162. http://dx.doi.org/10.1037/a0036352

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1–29. http://dx.doi.org/10.1146/annurev-psych-120710-100422

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9,* 226–232. http://dx.doi.org/10.1002/bs.3830090304

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Carbone, E., & Hey, J. (2000). Which error story is best? *Journal of Risk and Uncertainty, 20,* 161–176. http://dx.doi.org/10.1023/A:1007829024107

Correction to Ashby and Rakow. (2014). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1509. http://dx.doi.org/10.1037/xlm0000087

Cowan, N. (2001). The magical Number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. http://dx.doi.org/10.1017/S0140525X01003922

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College Press. Original work published 1885.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23,* 15–47. http://dx.doi.org/10.1002/bdm.683

Fox, P. (1997). *The Port Mathematical Subroutine Library, Version 3*. Murray Hill, NJ: AT&T Bell Laboratories.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118,* 523–551. http://dx.doi.org/10.1037/a0024558

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518. http://dx.doi.org/10.1002/bdm.598

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4939-2236-9_2

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 75–91). New York, NY: Cambridge University Press.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience. Do our patterns of sampling foreshadow our decisions? *Psychological Science, 21,* 1787–1792. http://dx.doi.org/10.1177/0956797610387443

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J

James, D. (2007). Stability of risk preference parameter estimates within the Becker-DeGroot-Marschak procedure. *Experimental Economics, 10,* 123–141. http://dx.doi.org/10.1007/s10683-006-9136-y

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition, 124,* 334–342. http://dx.doi.org/10.1016/j.cognition.2012.06.002

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. London, UK: Sage.

March, J. G. (1996). Learning to be risk averse. *Psychological Review, 103,* 309–319. http://dx.doi.org/10.1037/0033-295X.103.2.309

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50,* 167–179. http://dx.doi.org/10.1016/j.jmp.2005.06.008

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116,* 499–518. http://dx.doi.org/10.1037/a0016104

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47–84. http://dx.doi.org/10.1016/j.cogpsych.2003.11.001

Pachur, T., & Scheibehenne, B. (2012). Constructing preference from experience: The endowment effect reflected in external information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1108–1116. http://dx.doi.org/10.1037/a0027637

Plott, C. R., & Zeiler, K. (2005). The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *The American Economic Review, 95,* 530–545. http://dx.doi.org/10.1257/0002828054201387

R Core Team. (2013). R [Computer software]: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Safra, Z., Segal, U., & Spivak, A. (1990). The Becker-DeGroot-Marschak mechanism and nonexpected utility: A testable approach.

Journal of Risk and Uncertainty, 3, 177–190. http://dx.doi.org/10.1007/BF00056371

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review, 22,* 391–407. http://dx.doi.org/10.3758/s13423-014-0684-4

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

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323. http://dx.doi.org/10.1007/BF00122574

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228–243. http://dx.doi.org/10.1037/a0027127

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. http://dx.doi.org/10.3758/BF03206482

Wulff, D. U., Hills, T. T., & Hertwig, R. (2012). Adaptive information search and decision making over single and repeated plays. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Building bridges across cognitive sciences around the world: Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1167–1172). Austin, TX: Cognitive Science Society.

Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*. Advance online publication.

Appendix A

Parameter Estimation in the Monte Carlo Simulation and Model Recovery

The models were fitted to the valuation of each simulated participant by maximizing log-likelihood over all J lotteries, using a normally distributed noise:

$$LL = \log\left(\prod_{j}^{J} \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{v_j - m_j}{\sigma}\right)^2}}{\Phi(4, m_j, \sigma) - \Phi(0, m_j, \sigma)}\right), \quad \text{(A1)}$$

with m_j being the predicted valuation of the model, v_j being the data, and Φ being the cumulative normal probability distribution. Note that this formalization does not specify the source of the noise; it could, for instance, result from error in response selection (e.g., trembling hand), noise in the actual valuation process (e.g., error in retrieval from memory), or an invalid model of the process. VUM and SWIM were implemented as defined in Equations 1 and 2. The predicted valuation of SUM at the nth sample, based across all n sampled outcomes, was defined as

$$v_n = \frac{1}{n}\sum_{i}^{n} x_i. \quad \text{(A2)}$$

The VUM parameters σ and φ (see Equations A1 and 1) were jointly estimated using R (R Core Team, 2013) via a quasi-Newton minimization procedure from the PORT library (Fox, 1997). The same procedure was applied for the SUM. To estimate the SWIM parameters, we used a different approach, as stable estimates were obtained only when the starting points for the window size were close to the true window size. Instead, we performed an exhaustive search for the window size ζ and optimized the LL for each discrete value of ζ to find the optimal value of σ, again using PORT routines. In the rare cases (less than 1% of runs) in which multiple window sizes led to the same optimal fit, the smaller value was chosen, consistent with Ashby and Rakow (2014). To avoid local minima, we repeated the fitting for several start values for σ and ζ.

Appendix B

Simulation: Estimation of SWIM's Window Size from Noisy Valuations

Our simulation demonstrating the impact of noise on the estimation of the SWIM window size was conducted as follows. For a broad range of sample size and noise levels, we simulated 1,000 agents completing the valuation task used in Ashby and Rakow (2014). The sample sizes and noise levels were set such that they covered 95% of the average (across trials) sample sizes and standard deviations, σ, reported by Ashby and Rakow, namely $1 \leq$ sample size ≥ 40 and $.01 \leq \sigma \geq 1.15$. Each of the agents first took a fixed number of samples for each of 40 lotteries (which are reported in Ashby & Rakow's 2014 supplemental material) and then provided valuations based on a truncated normal distribution centered on the mean outcome of the respective sample (i.e., it was assumed that the window size equaled the sample size, $\zeta = n$). We then fitted the SWIM to each simulated agent's data (corresponding to Appendix A, but using a nontruncated normal to match the approach taken by Ashby and Rakow) and estimated the window size and noise level for each simulated agent. Finally, to evaluate the empirical window-size estimates, we matched each participant in Studies 1 and 2 to the conditions in our simulation that were closest in terms of sample size and noise level. This resulted in an average proportion of sample size to window size of .87, which is very close to the value of .86 reported by Ashby and Rakow for their empirical data.

Received January 5, 2015
Revision received June 2, 2015
Accepted June 3, 2015 ■