Journal of Behavioral Decision Making
J. Behav. Dec. Making, **21**: 428–448 (2008)
Published online 6 March 2008 in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.597

Response Error and Processing Biases in Confidence Judgment

EDGAR C. MERKLE[1]*, WINSTON R. SIECK[2] and TRISHA VAN ZANDT[3]
[1] *Department of Psychology, Wichita State University, Wichita, Kansas, USA*
[2] *Applied Research Associates, Fairborn, Ohio, USA*
[3] *Department of Psychology, Ohio State University, Columbus, Ohio, USA*

ABSTRACT

Previous studies showed that random error can explain overconfidence effects typically observed in the literature. One of these studies concluded that, after accounting for random error effects in the data, there is little support for cognitive-processing biases in confidence elicitation. In this paper, we investigate more closely the random error explanation for overconfidence. We generated data from four models of confidence and then estimated the magnitude of random error in the data. Our results show that, in addition to the true magnitude of random error specified in the simulations, the error estimates are influenced by important cognitive-processing biases in the confidence elicitation process. We found that random error in the response process can account for the degree of overconfidence found in calibration studies, even when that overconfidence is actually caused by other factors. Thus, the error models say little about whether cognitive biases are present in the confidence elicitation process. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS overconfidence; response error; confidence calibration; probability judgment; cognitive bias

INTRODUCTION

Confidence, accuracy, and their relationship are immensely popular topics in judgment research. Part of this popularity comes from the fact that confidence judgments are intuitive: people usually have an idea of what confidence is, and how to express confidence, before stepping foot in a laboratory. A related part of this popularity comes from the fact that people frequently use confidence judgments, either implicitly or explicitly, in many applications such as medical diagnosis (Arkes et al., 1995), eyewitness identification (Wells, 1981), and meteorology (O'Hagan et al., 2006, who also discuss a number of other applications). From the researchers' perspective, the correspondence between confidence and accuracy is straightforward to assess, which makes confidence research more accessible. For example, if a decision maker gives an average

* Correspondence to: Edgar C. Merkle, Department of Psychology, Wichita State University, 1845 Fairmount Box 34, Wichita, KS 67260, USA. E-mail: edgar.merkle@wichita.edu

confidence of *x*% to a set of items, *x*% of those items should be correct. The further that accuracy strays from confidence, the more poorly calibrated is the decision maker.

Perhaps the most pervasive and well-known phenomenon on the literature on confidence calibration is that of overconfidence: confidence tends to exceed accuracy. In the last several years, a number of studies have called into question this overconfidence phenomenon. The arguments against overconfidence generally fall in one of two camps: ecological validity or statistical artifact. The ecological validity arguments (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994) typically revolve around experimenters (consciously or unconsciously) selecting test questions that are not representative of all possible test questions in a given domain. If there are a large number of trick questions in a test, the argument goes, then judges will be overconfident because their prior experience in the test domain conflicts with the administered questions.

Others have proposed that overconfidence should be at least partly dismissed as a statistical artifact (Erev, Wallsten, & Budescu, 1994; Pfeifer, 1994). These researchers argue that empirical findings are dependent on the method of analysis or that random error can account for the findings. For example, Erev et al. showed that the relationship between subjective and objective probability (e.g., between confidence and proportion correct) can change depending on whether subjective probability is computed as a function of objective probability or vice versa. They then presented a class of models in which a true, unbiased confidence judgment is perturbed by random error and showed that these models can produce classic findings in the overconfidence literature. Taken together, the ecological and statistical validity arguments imply that the overconfidence phenomenon is not necessarily an interesting topic of study. They suggest that there are no systematic cognitive biases at work in the confidence elicitation process and that only random error or biased test items are to blame for observed overconfidence.

Further complicating the situation, overconfidence is not ubiquitous: it is typically reduced or even reversed for very easy tests, a phenomenon generally known as the hard–easy effect. In its general form, the hard–easy effect occurs when people exhibit more overconfidence for more difficult sets of questions (and less overconfidence, or even underconfidence, for easy sets of questions). Lichtenstein and Fischhoff (1977) uncovered the hard–easy effect, and, to our knowledge, were also the first to address the possibility that the effect might be a statistical artifact. They tested the possibility that their results could be explained solely in terms of random error and found that they could not. These researchers also noted the critical implication that overconfidence did not amount to a simple systematic effect that could, for example, be subtracted out of a respondent's judgments (Lichtenstein & Fischhoff, 1980). In other words, if a systematic cognitive bias is present in the confidence elicitation process, it is not necessarily the case that this bias will translate into a simple, systematic overconfidence effect.

Other researchers have explored the artifactual and/or ecological contributions to overconfidence in more detail. For example, Budescu, Wallsten, and Au (1997) used clever experimentation to estimate both between-item and within-item variability (i.e., error) in confidence judgments. They then used a type of signal detection model (Wallsten & González-Vallejo, 1994) to examine the magnitude of overconfidence that would be exhibited by perfectly calibrated judges possessing the same amount of error as was observed in the real participants. These researchers found that error variability could not account for the full magnitude of overconfidence that was exhibited by their participants; thus, overconfidence cannot be explained solely by random error.

In contrast to Budescu et al. (1997), Juslin, Winman, and Olsson (2000) argued that overconfidence and the hard–easy effect can be attributed almost entirely to ecological and statistical artifacts. These researchers first compared representatively selected and informally selected general knowledge questions across several experiments that were matched for difficulty. Representatively selected questions are those that were randomly chosen from some larger set of potential questions, while informally selected questions are those that an experimenter explicitly chose. In their examination of these two types of tests, Juslin et al. found half as much overconfidence for the representatively selected tests as for the informally selected tests. They then

DOI: 10.1002/bdm

corrected the representatively selected data for statistical artifacts, using procedures that bear some similarity to the Budescu et al. procedure.

Specifically, Juslin et al. (2000) made statistical adjustments using two different procedures: a ''split-sample'' procedure devised by Klayman, Soll, González-Vallejo, and Barlas (1999), and a model-based procedure that is intended to statistically control for random error presumed to arise in the response system. They found that the hard–easy effect, and hence overconfidence for the difficult tests, was indeed reduced by their statistical adjustments to the data. Based on these results, they concluded that there is little evidence of cognitive biases contributing to the overconfidence effect. As summarized by O'Hagan et al. (2006), ''Juslin et al. (2000) argue vehemently that, when all potential artifacts associated with calibration tasks and their analysis are taken into account, there is little evidence of any meaningful cognitive overconfidence bias'' (p. 82).

In this paper, we focus on the role of random error in the confidence elicitation process, as compared to the role of systematic biases. Random error effects are unsystematic (and therefore uninteresting), while systematic biases such as alternative underweighting (e.g., McKenzie, 1997) are interesting research topics that can potentially be manipulated experimentally. Of specific interest to us is the ability of random error models to capture, or explain, data produced by other types of models. Most importantly, if estimated random error parameters change with cognitive biases present in the confidence elicitation process, then we can draw no conclusions about whether random error or cognitive biases drive miscalibration effects. This issue is important because it appears to run counter to the model selection literature (e.g., Myung, 2000; Pitt, Myung, & Zhang, 2002), which tells us to prefer the simplest explanation of a phenomenon. From this perspective, it is often assumed that random error provides the simplest explanations of psychological phenomena. We show here that the random error models are flexible enough to explain many calibration data sets, even when the data sources contain cognitive biases.

Stated differently, model selection research generally focuses on the number of distinct data patterns that a model can generate. We are as much concerned with the data production process, however, as we are with the exact data patterns that the model generates. To determine whether error models can capture data from other models, we follow a general approach employed by Ratcliff, Van Zandt, and McKoon (1995, among others). We first simulate realistic data from known models and then apply an error estimation procedure (specifically, the procedure developed by Juslin et al., 2000) to the simulated data. Because we know the data-generating mechanism for the simulated data, we know exactly how the procedure should perform if it is only estimating random error.

In the sections that follow, we first give a detailed description of Juslin et al.'s (2000) statistical adjustment procedure and clarify the parts of the procedure that we use to investigate random error in confidence calibration. We also examine general implications of the magnitude of response error found in previous data sets. Next, we systematically simulate data from four models of confidence: the combined error model (Juslin, Olsson, & Bjorkman, 1997), the decision variable partition model (DVPM) (Ferrell & McGoey, 1980), a novel DVPM with a cognitive bias component, and a novel combined error model with a cognitive bias component. We describe each model in detail and apply the correction procedure to each of them. Finally, we discuss the relationship between random error and cognitive biases, and we argue for the importance of pursuing experimental manipulations that are designed to illuminate cognitive overconfidence mechanisms.

CORRECTING OVERCONFIDENCE DATA

Juslin et al. (2000) describe and correct for two types of statistical artifacts associated with overconfidence data: ''linear dependency,'' and ''scale-end effects,'' which deal with random error. Based on a quantitative review of confidence experiments and on corrections for the statistical artifacts, Juslin et al. conclude that the hard–easy effect is due mostly to statistical artifacts and that the data do ‘‘not support the idea of a cognitive overconfidence bias that is due to, for example, confirmatory search of memory’’ (p. 393). Their conclusions are partly based on the assumption that the corrections accurately estimate linear dependency and scale-end effects. Below, we summarize these two types of problems and examine the correction procedures’ accuracy.

Linear dependency

For any decision maker, let $\overline{d}$ be proportion correct for a stated confidence level $\overline{f}$. Linear dependency refers to the fact that, in computing the correlation between proportion correct ($\overline{d}$) and overconfidence ($\mathrm{OC} = \overline{f} - \overline{d}$), proportion correct ($\overline{d}$) enters twice into the equation. Because we can expect some amount of error in sample proportions correct (e.g., variation of sample proportion correct around the population proportion correct), this causes the observed correlation (and thus, the observed hard–easy effect) to be artificially large.

For example, consider a decision maker with population proportion correct μ_d and population mean confidence μ_f. For any given sample of test items (denoted sample i here), random error means that the observed proportion correct ($\overline{d}_i$) will be different from μ_d:

$$\overline{d}_i = \mu_d + e_i$$

where e_i is the sample error associated with sample i. In words, any sample of test items will exhibit a proportion correct that is not exactly equal to the population proportion correct. The error term in the above equation, e_i, also has an impact on overconfidence for sample i. This is because $\overline{d}_i$ also enters into the overconfidence calculation. For a mean confidence of $\overline{f}_i$, we can calculate overconfidence (OC_i) for sample i as

$$\mathrm{OC}_i = \overline{f}_i - \overline{d}_i = \overline{f}_i - (\mu_d + e_i)$$

It is evident that the sign of e_i has opposite effects on the magnitude of $\overline{d}_i$ and OC_i: e_i is added in the proportion correct equation and subtracted in the overconfidence equation. That is, for positive e_i, proportion correct increases and overconfidence decreases. The opposite occurs for negative e_i. Thus, we could obtain a hard–easy effect by the simple addition of error to the population proportion correct.

To correct for the linear dependency problem, Juslin et al. (2000) make use of a partitioning scheme developed by Klayman et al. (1999). In this scheme, experimental data are partitioned into two sets. The first set is used to estimate proportion correct, and the second set is used to estimate overconfidence. Thus, one e_i enters into the calculation of $\overline{d}$ and a second, independent e_i enters into the calculation of OC. This removes the artificial relationship between $\overline{d}$ and OC that results from the same e_i entering into the calculation of both terms. The Klayman et al. method is an effective way of resolving the linear dependency problem in small samples. As the number of test items gets large, however, the impact of the linear dependency problem is reduced and the partitioning method is unnecessary. By the Weak Law of Large Numbers, as the number of test items increase, the sample mean $\overline{d}$ approaches the population mean μ_d. This implies that the e_i become vanishingly small and will not influence linear dependency. To summarize, for large sample sizes, the Klayman et al. procedure is unnecessary. In the simulations that we describe in this manuscript, we always generated a large sample of data for each test—large enough so that the partitioning method has no impact on the results. Hence, we do not discuss the Klayman et al. procedure further.

Scale-end effects

Scale-end effects refer to the fact that random error in confidence judgments can influence the magnitude of the overconfidence effect. To be more specific, as the magnitude of random error increases, miscalibration also tends to increase. Imagine a set of experiments where the decision makers give unbiased (perfectly

DOI: 10.1002/bdm

calibrated) confidence judgments on a 50–100% scale across a series of two-alternative, general-knowledge experiments. For each experiment, we could plot the experiment-wise overconfidence scores by the corresponding proportions correct. Such a plot is shown in Figure 1(a). Assuming the decision makers are perfectly unbiased, the points from each experiment would all fall along the solid line in Figure 1(a). This is because unbiased judges will have an overconfidence score of 0, regardless of their proportion correct.

If we were instead to first add random response error (ε_i) to the decision maker's confidence judgments over a series of experiments, then we might observe the points in Figure 1(a) (each point is one experiment,

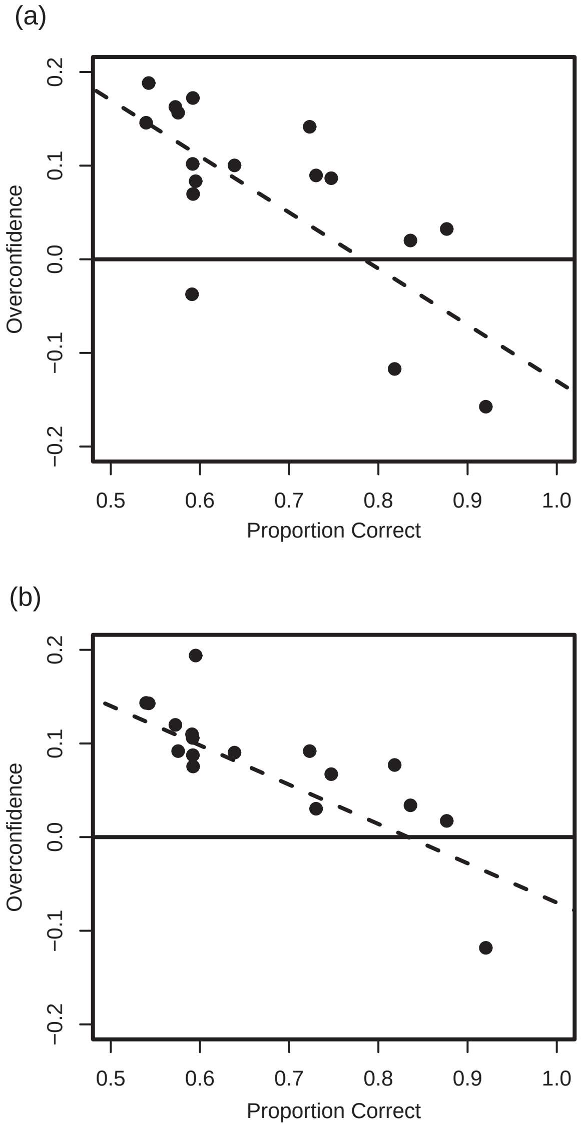

Figure 1. A depiction of the Juslin et al. (2000) response error correction procedure. In plot (a), a regression line is fit to observed proportion correct and overconfidence statistics collected across 17 experiments. In plot (b), a regression line is fit to data simulated from the estimated response error model. To correct for response error, the slope of the regression line in plot (b) is subtracted from the slope of the regression line in plot (a)

DOI: 10.1002/bdm

and the dashed line is a regression line fit to these points). This is because, at confidence levels near 100%, error can only make confidence go down. That is, decreases in confidence resulting from ε_i can be much larger in magnitude than increases in confidence. Correspondingly, at confidence levels near 50%, error can only make confidence go up. This results in overconfidence at the low end of the confidence scale and underconfidence at the high end, despite the fact that the decision maker's internal judgments were accurate. The effect implicitly depends on the bounds of the probability scale; if one instead used a log-odds scale, then there are no scale-end effects because the log-odds scale is unbounded (going from - ∞ to + ∞). Note that the scale-end effects can impact both overconfidence results and hard–easy effects, whereas linear dependency is specific to the relationship between proportion correct and overconfidence (e.g., hard–easy effects).

To correct for scale-end effects, Juslin et al. (2000) estimated the amount of random error in their data using the response error model (Juslin et al., 1997). Briefly, they fit the response error model to a calibration curve and confidence distribution aggregated over many experiments, and the fitted model yielded an estimate of the variance of ε. They then used the fitted model to generate new data that reflect miscalibration arising solely from random error (e.g., scale-end effects). From the generated data, they quantified the scale-end effects and removed them from their observed data. We describe both the procedure and the response error model in detail below.

The response error model

Model overview

Juslin et al. (1997) proposed that random error in the response system is an important contributor to observed overconfidence, and they conducted several simulations to demonstrate that overconfidence does indeed follow from their proposal. The idea of their response error model is that, when confronted with a general knowledge question, the respondent generates an ecologically valid internal probability that determines the choice response and confidence that the chosen alternative is correct. An overt probability that the correct choice has been made is then rendered according to the format provided by the experimenter. For example, two common formats are to have the person circle one of the six values 0.5, 0.6, . . ., 1.0, or to have the person write or type a number between 50 and 100%. There is some inconsistency associated with response execution, so that the same internal probability does not always lead to the same overt probability.

Formally, the ecological/internal probabilities are assumed to follow a symmetric beta distribution, and so range from 0 to 1. The single parameter of this distribution, α, represents difficulty and is set to reflect the observed proportion of correct responses. When α is less than 1, probabilities are close to 0 or 1, reflecting an easy task. When α is greater than 1, probabilities are closer to 0.5, reflecting a harder task. Response error follows a normal distribution with mean zero and variance σ^2. Juslin et al. showed by simulation that accurate ecological/internal probabilities combined with a sufficient amount of response error could produce overconfidence. In particular, they modeled response error with variances 0.04, 0.08, 0.12, and 0.16, though the authors indicated that these represented overly large variances chosen for illustrative purposes.

Model-based corrections

Juslin et al. (2000) used this response error model to correct overconfidence data for the scale-end effects described above. As a first step, they plotted overconfidence scores against proportion correct for 17 distinct data sets collected from a literature review. Figure 1(a) shows a plot with hypothetical data points, designed to be similar to Juslin et al.'s Figure 3. They fit a regression line (depicted by the dashed line in Figure 1(a)) to these observations, yielding a negative slope that characterizes the size of the observed hard–easy effect (i.e., as proportion correct increases, overconfidence decreases). Next, they fit the response error model in two steps. First, they set α so that the mean of the beta distribution was equal to the observed proportion correct.

DOI: 10.1002/bdm

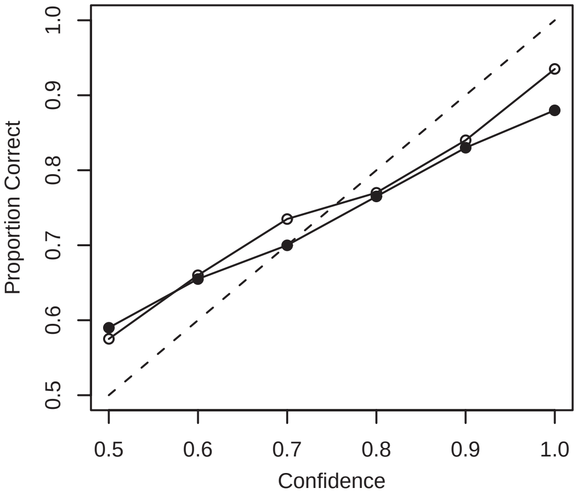

Figure 2. Fitted and observed calibration curves. In the Juslin et al. (2000) correction, the observed curve (solid points) is calculated over a number of experiments. The response error model is then fit to the curve, yielding the curve with empty points. Specific points on these curves were adapted from Figure 1(b) of Juslin et al. (2000)

Second, they fit the σ^2 parameter via least squares to a single calibration curve based on all 17 of the data sets, along with the corresponding distribution of confidence responses.

The calibration curve is a graph of proportion correct conditioned on different levels of confidence; Figure 2 presents a hypothetical calibration curve. Proportion correct for each confidence level (0.5, 0.6, . . ., 1) is calculated across all 17 experiments and plotted in the graph. The proportion of responses within each confidence level is also calculated; this is called the *confidence distribution*. The response error variance, σ^2, is then adjusted via a grid search so that the predicted calibration curve and confidence distribution most closely resemble the observed calibration curve and confidence distribution (as judged by the sum of squared error). The resulting σ^2 estimate represents the magnitude of random response error, or scale-end effects, that is present in the data.

After fitting the response error model to the observed calibration curve, Juslin et al. (2000) used the fitted model to estimate overconfidence effects that arise solely from response error. More specifically, they generated overconfidence predictions for each individual dataset using the fitted response error model. Figure 1(b) shows these predictions as hypothetical data points, which represent the level of overconfidence that can be attributed to the scale-end effects alone. Further, the slope of the regression line (the dashed line in Figure 1(b)) relating these predicted overconfidence values to the proportion correct characterizes the size of the hard–easy effect due to response error, and so provides the adjustment factor.

Although the fit of the model to the calibration curve may be extremely close, the model generally produces a smaller regression slope than does the empirical data representing the hard–easy effect. The difference is termed the ‘‘corrected slope,’’ and it represents the hard–easy effect remaining after effects due to response error have been subtracted out. In other words, the difference between the slope of the observed-data regression line (from Figure 1(a)) and the slope of the simulated-data regression line (from Figure 1(b)) measures the magnitude of the corrected hard–easy effect. Corrected slopes close to zero imply that the hard–easy effect is largely based on response error (scale-end effects), while corrected slopes further from zero imply a ‘‘real’’ hard–easy effect.

Preliminary examination of the correction

A key assumption underlying this scale-end correction is that the response error model describes the true confidence elicitation process closely enough that the estimates of response error are accurate. The model fits

DOI: 10.1002/bdm

aggregate confidence data well, but that fact alone does not mean that the model's description of the confidence process is a good one. The model's flexibility, or ability to accommodate different types of data patterns, also plays an important role in a model's goodness of fit (Roberts & Pashler, 2000). If the true confidence elicitation process differs substantially from the model, then model-based response error estimates may be inaccurate. Further, if the true confidence elicitation process changes substantially across experiments or judges, then the aggregate response error estimates may not reflect any single experiment or judge. Any of these factors could lead to inaccurate scale-end corrections and inaccurate conclusions about the impact of response error on overconfidence and the hard–easy effect.

As a first step toward examining this assumption, we can establish guidelines for a priori plausible values of response-error variances. In other words, how noisy is a response system that has an estimated variance of 0.02 (which is the estimated response error variance from Juslin et al., 2000)? On the surface, 0.02 appears to be a modest number; yet, it is prudent to keep in mind that almost any number associated with a probability scale is going to be small. It is thus useful to translate the variance into the percentage of misclassifications due to errors in the motor response system. Also, it is extremely difficult to separate error in the response system from other sources of error, especially in cognitively complicated tasks like general knowledge tests. However, as one point of comparison, we can consider other studies that have attempted to characterize the scope of response error in cognitively simpler tasks. An intuitively reasonable range that recurs in such research is 2–7% (Nosofsky, Palmeri, & McKinley, 1994). For example, Levine (1975) estimated that study participants who had acquired the correct hypothesis in a concept-learning task nevertheless misclassified stimuli about 5% of the time because of response error. Although these tasks are somewhat different than the general knowledge tasks of interest here, the judgment aspects are similar enough to provide a general reference to establish a plausible range of values. Furthermore, no study that we know of isolates misclassification rates due to response error in general knowledge tasks. To do so, it would be necessary to administer questions where judges directly retrieve the answers from memory (i.e., questions that judges ''know'') so that other cognitive mechanisms did not enter into the response process.

We can use the 2% to 7% figure in the current context to check the published variance estimate of 0.02 that Juslin et al. (2000) claim characterize a person's response error. Consider a two-alternative forced choice task with six confidence response options of 0.5, 0.6, . . ., 1.0. We first examine what percentage of the time such a person whose cognitive processing (e.g., in the ecological models, retrieval of learned frequency information) has led to a particular internal probability will inadvertently circle one of the other five values. A quick consultation of the normal probability tables reveals that the individual with 0.02 response error variance and an internal probability of 0.6, 0.7, 0.8, or 0.9 will involuntarily circle some other number on at least 72% of the trials.

Figure 3 depicts the above calculation for an internal probability of 0.7. After adding normally distributed response error (with variance of 0.02) to the internal probability of 0.7, confidence judgments follow the normal distribution in Figure 3. We assume that the judge will state ''0.7,'' if her judgment is anywhere between 0.65 and 0.75; this is the shaded region of Figure 3. Using normal probability tables, we find that the area of the shaded region is 0.276, and the area outside the shaded region is 0.724. This means that, 72% of the time, the judge will report a probability different from 0.7. The number is smaller when the internal probabilities are at the extreme values of 0.5 or 1.0; errors occur 36% of the time in these cases. Clearly, the 0.02 variance estimate implies a massive amount of response error. What magnitude of response error variance yields a misclassification rate of 5%? For an internal probability of 0.7, a response error variance of 0.00065 yields the 5% rate. This value is about 30 times smaller than the estimated value of 0.02.

These observations demonstrate that the estimated error variance reported by Juslin et al. (2000) in their review of empirical data is too large to reflect only post-cognitive processes in the response system. Because the variance estimate cannot be explained solely by response error, it is likely that the variance estimate results from cognitive processes, including cognitive biases and limitations. In the following sections, we test this idea by simulating data from four confidence models and applying the Juslin et al. response error

DOI: 10.1002/bdm

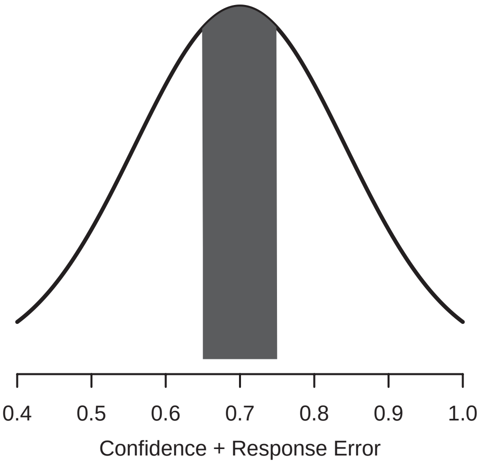

Figure 3. Calculation of confidence responses implied by a response error variance of 0.02. The judge's true, internal judgment is assumed to be 0.7, and the shaded area depicts the probability that a judge's external judgment (with response error added) is still 0.7

correction to each dataset. We are most interested in observing the stability of the response error estimate as we change model parameters that do not reflect response error.

Combined error model

As suggested by its name, the combined error model (Juslin et al., 1997) introduces error into probability judgments via two routes: the response error route described above, and an ecological route. These errors lead to overconfidence and general miscalibration. Within the ecological component (e.g., Gigerenzer et al., 1991) of the model, judgments arise from a judge's prior experience with similar stimuli in the environment. Error arises because a judge's prior experience in a given environment will not correspond exactly to the parameters of the environment. For example, if a judge is deciding which of two cities is larger, she might pick a city based on the fact that one city has a professional baseball team and the other does not. Furthermore, the judge's probability of being correct might be based on the fact that she has successfully used this sports-team cue 7 out of 10 times in the past (hence, a probability judgment of 70%). Error results from the fact that the judge has received feedback on this cue only 10 times; if she used the cue thousands of times, she might find that the cue is actually accurate 65% of the time. Correspondingly, the ecological error described here is modeled via random draws from a binomial(n, p_0) distribution, where n is the amount of prior experience that a judge has with the cue and p_0 is the true environmental probability that the cue leads to a correct choice.

Response error occurs in the model when internal probability judgments are translated into external probability judgments. The model assumes that, after encoding an appropriate stimulus and choosing an answer, the judge makes a probability judgment. This probability judgment may be based on ecological cues, or it may be exactly accurate for a given stimulus. In any case, such an internal probability judgment may be described as a ‘‘feeling of confidence’’ that must be translated into an overt probability. The translation from feeling of confidence to overt probability is not completely accurate, and error in the overt probability judgment results. Such error is modeled as the addition of a random draw from a normal(0, σ^2) distribution to the internal probability judgment.

In estimating response error in experimental data, Juslin et al. (2000) considered using the combined error model instead of the response error model. They decided against this, however, because the combined error

DOI: 10.1002/bdm

model's estimated n parameter tended to be large when fitted to previous data sets.[1] Note, however, that the combined error model has two parameters that affect proportion correct. The beta distribution's parameter α represents test difficulty, and the binomial n represents degree of experience. Given the model flexibility issues described earlier, it is not clear that these two parameters are uniquely identifiable. That is, they may have achieved equivalently good fits to the data with the combined error model by assuming less experience and easier tests (lower n and lower α).[2]

Simulation

Our implementation of the combined error model was similar to that of Juslin et al. (1997). We first set the model parameters (n, α, and σ^2) and generated many ‘‘experiments’’ of data from the combined error model across a range of proportions correct. Next, we used Juslin et al.'s (2000) procedure to estimate the amount of error variance in the generated data. This method essentially consists of fitting the response error model to the observed data (which consist of a calibration curve and the corresponding confidence distribution), where the free parameters are σ^2 (the response error variance) and α (which is determined by the observed proportion correct). Finally, we examined how the error variance estimates change with n. If the response error correction is robust to the data-generating process, then it should accurately recover the error variance (σ^2) regardless of the value of n.

For each simulation reported in this paper, we generated 17 experiments of 20 000 trials each. The 17 experiments were chosen to match the number of experiments that Juslin et al. (2000) used. The 20 000 trials per experiment were chosen to make the summary statistics very precise, removing the need for formal statistical tests on the differences between the true error variances and estimated error variances.

Results and discussion

The results of several simulations are presented in Table 1, with the values of n and σ^2 that we used to generate data noted in the first and second columns. Table 1 shows that the error correction procedure described above can behave quite liberally in some situations. When $n = 64$, the procedure is reasonably accurate: σ^2 estimates (column 4) are nearly equal to the actual σ^2 value used to generate the data (column 2), and the corrected slopes representing the hard–easy effects are all (correctly) close to 0. Said differently, the response error corrections (column 5) are accurate for $n = 64$. As n decreases, however, σ^2 is overestimated because it is capturing the added binomial sampling error. This can again be seen by comparing columns 2 and 4. The overestimate of σ^2 leads to response error corrections that are too large (in an absolute value sense); in other words, too much of the original slope relating accuracy to overconfidence is attributed to response error/scale-end effects.

In summary, Juslin et al.'s (2000) scale-end correction procedure overestimates response error in data that are generated from the combined error model with small values of n. This happens because, in the combined error model, error enters into judgments via both response error and ecological error. When estimating response error, the response error model confuses ecological error for response error, resulting in the overestimates. Thus, if judges have a small amount of experience with a particular cue, their response error will likely be overestimated. This may not be the case with the data in the Juslin et al. (2000) paper (the $n = 64$ condition is most relevant here), but studies with other paradigms have suggested that n may be small because it reflects limitations in retrieval in addition to low experience.

[1] When n is large, the combined error model becomes the response error model.

[2] Note, however, that joint estimation of the calibration curve and confidence distribution may reduce identifiability issues. Furthermore, Juslin et al. (1997) present evidence that the α parameter has a larger effect on proportion correct than the n parameter.

DOI: 10.1002/bdm

Table 1. Simulation results for the combined error model

Parameters		Results			
n	σ^2	Original slope	Est. σ^2	Scale-end slope	Corrected slope
4	0.01	−0.499	0.042	−0.277	−0.222
	0.03	−0.569	0.060	−0.362	−0.207
	0.05	−0.613	0.087	−0.423	−0.190
8	0.01	−0.385	0.019	−0.151	−0.233
	0.03	−0.473	0.036	−0.253	−0.220
	0.05	−0.532	0.065	−0.371	−0.160
64	0.01	−0.152	0.013	−0.123	−0.029
	0.03	−0.274	0.030	−0.222	−0.052
	0.05	−0.362	0.049	−0.314	−0.048

For example, small values of n have been used with the PROBEX model (Juslin & Persson, 2002) in the domain of exemplar retrieval.[3] *Exemplars* are specific cases in memory that a judge believes to be representative of a test category. For example, a doctor trying to diagnose appendicitis may recall past patients with appendicitis as exemplars for making the current diagnosis. The PROBEX model generally describes how judges form probability judgments based on these exemplars. The combined error model can be viewed as a special case of the PROBEX model, so the use of small n with PROBEX implies that small n may also be used with the combined error model. In general, random error could mask ecological effects present in the confidence elicitation process. This is problematic if we are trying to conclude that response error is responsible for confidence calibration effects, rather than systematic cognitive mechanisms such as retrieval limitations.

We have demonstrated above that the scale-end correction procedure has the ability to confuse ecological error for response error, even when the confidence-generating mechanism is very similar to the response error model. How does the correction procedure perform when the confidence-generating mechanism is different from the response error model, as we can expect in true experiments? We investigate this question below.

Decision variable partition model

The DVPM is one of the earliest and most general formal models of probability judgment specifically designed to address questions of calibration (Ferrell & McGoey, 1980). It is based on signal-detection theory and was developed for use in a wide variety of task formats, including the two-alternative forced-choice (2AFC) tasks that are the focus of the current paper. We describe the model only with respect to this kind of task.

The DVPM derives its name from the fact that confidence arises from an arbitrarily scaled random variable, the "decision variable." The range of the decision variable is segmented into a succession of partitions that correspond to reported confidence on whatever scale is provided (e.g. "8" or "very likely"). In the 2AFC task, the respondent examines each of the possible alternatives and then generates a numeric plausibility for each alternative. The plausibilities, which are internal to the judge and do not follow any particular scale, arise from two normal distributions in a standard signal detection framework. This framework is presented in Figure 4(a), where one distribution corresponds to the chosen alternative's

[3] In this domain, n can be considered the number of exemplars retrieved before making a decision. In one instance of fitting PROBEX to data, Juslin and Persson found that, on average, participants sample only two exemplars before making a decision.

DOI: 10.1002/bdm

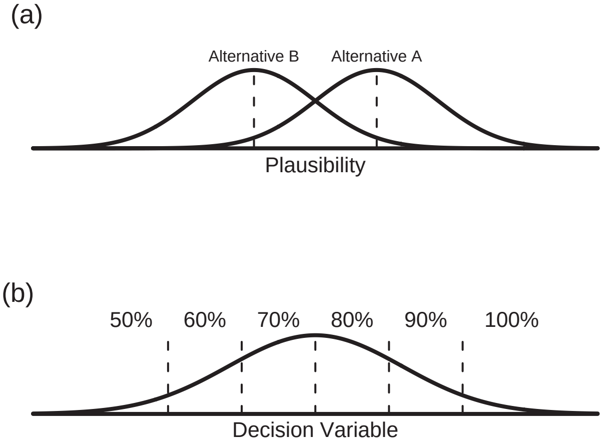

Figure 4. Illustration of Ferrell and McGoey's (1980) decision variable partition model. In panel (a), the judge samples a plausibility value from the distribution of each alternative. In panel (b), a confidence judgment is attained by partitioning the decision variable. This variable is simply the difference between plausibility values of the two alternatives

potential plausibility values and the other distribution corresponds to the unchosen alternative's potential plausibility values. The respondent chooses the alternative that produced the larger plausibility value (in Figure 4(a), this would most often be Alternative A). Confidence is then based on the difference between the larger plausibility value and the smaller plausibility value (i.e., plausibility of the chosen alternative minus plausibility of the unchosen alternative), such that larger differences yield higher confidence. The distribution of differences between plausibility values (i.e., the distribution of the ''decision variable''), presented in Figure 4(b), can be derived from the distributions in Figure 4(a). The judge places partitions on this distribution, and the partitions determine which confidence judgment is reported (in Figure 4(b), the partitions are the dotted vertical lines). For a given test item, the respondent considers the magnitude of the decision variable and gives a confidence response corresponding to the partition in which the decision variable lies.

The DVPM provides good fits to overconfidence and hard–easy effects observed in a wide variety of data (Ferrell & McGoey, 1980; Suantak, Bolger, & Ferrell, 1996). The model explains the hard–easy effect by assuming that the placement of partitions is relatively insensitive to task difficulty; they tend to be too narrow for hard tasks and too wide for easy tasks. As a result, more overconfidence is typically observed for harder sets of questions. The partitions do not reflect response error, so, as partition parameters change, any scale-end effects in the model should remain constant.

Simulation

We first used the DVPM to generate data, supposing that people actually operate as specified by the DVPM. We then applied the response error correction to the data generated by the DVPM. If the response error correction is reasonable, we should not find any effect of the correction in this instance. That is, the hard–easy effect exhibited by the DVPM should not be reduced. This is because there is no explicit response error present in the DVPM. If the scale-end correction does reduce the DVPM's hard–easy effect, then the reduction should at least be approximately equal across different parameter values. Partition parameters in the DVPM do not reflect response error, so partition manipulations should not influence the response error estimates and scale-end corrections.

DOI: 10.1002/bdm

For each of 5 sets of partitions, we simulated 17 experiments of data from the DVPM with overall levels of proportion correct ranging from 0.62 to 0.83. This range mimics the data points in Juslin et al.'s (2000) analyses. We manipulated proportion correct in the model by changing the mean separation between distributions for the correct and incorrect alternatives. We then applied the scale-end correction to data generated from each set of partitions.

Results and discussion

Table 2 shows the results of several simulations. The values of the partition parameters are shown in the first column, and the slopes for the DVPM-generated datasets are shown in the second column. These slope values are all approximately the same as the value Juslin et al. (2000) reported for their data. We fit the response-error model to each set of data produced by the DVPM, and the resulting σ^2 estimates are displayed in the third column. Modifications of the partitions influenced the estimated σ^2 substantially, with very modest changes in slope. This is problematic because the partition parameters do not reflect response error; thus, estimated σ^2 should remain the same across the rows of Table 2. Finally, the response-error correction (i.e., the scale-end slope) associated with the two lowest levels of estimated σ^2 are approximately the same as the response-error corrections that Juslin et al. reported.

Juslin et al. (2000) claimed that their results ''. . . remove the main support for the signal-detection-based decision variable partition model (Ferrell & McGoey, 1980)'' (p. 393). If the participants in Juslin et al.'s (2000) studies operated essentially as described by the DVPM, then response error corrections could mask the DVPM processes. We can make no claims about the validity of the DVPM based on data transformations that result from the response-error model.

In the next section, we extend this argument to include models that explicitly incorporate cognitive biases, an ''underweighted alternative'' bias in particular (e.g., McKenzie, 1997). Alternative underweighting occurs when, in evaluating confidence in a particular choice, participants ignore or downplay the plausibility of the unchosen alternative, resulting in overconfidence. To arrive at a formal model, we added an alternative-underweighting bias to the DVPM.

Underweighted alternative model

The underweighted alternative decision variable partition model (UADVPM) incorporates a cognitive processing bias into the regular DVPM. The main difference between the DVPM and the UADVPM is the way that probability judgments are constructed. In the DVPM, probability judgments arise based on partitioning of the ''decision variable''. For example, if the value of an observer's ''decision variable'' falls in partition A (see Figure 4(b)), then she may give a probability judgment of, say, 0.8. If the value of the ''decision variable'' instead falls in partition B, then the observer's probability judgment may be assigned a value different from 0.8 (see DVPM description).

Table 2. Simulation results for the decision variable partition model

Partition parameters	Original slope	Est. σ^2	Scale-end slope	Corrected slope
(0.45, 0.65, 1, 1.3, 1.8)	−0.68	0.023	−0.19	−0.49
(0.5, 0.7, 1, 1.4, 1.8)	−0.68	0.027	−0.21	−0.47
(0.5, 0.65, 0.85, 1, 1.8)	−0.68	0.041	−0.28	−0.40
(0.35, 0.57, 0.74, 0.98, 1.2)	−0.68	0.077	−0.40	−0.28
(0.45, 0.65, 1, 1.2, 1.8)	−0.65	0.149	−0.53	−0.12

DOI: 10.1002/bdm

Like the DVPM, the UADVPM initially assumes that the judge assigns a plausibility value to each alternative. In this case, however, there is a decision variable for each alternative. That is, the judge first evaluates the truth of Alternative A and assigns a probability to Alternative A. Next, the judge evaluates the truth of Alternative B and assigns a probability to Alternative B.[4] The judge chooses whichever alternative has the higher probability. The overt (final) confidence judgment for the chosen alternative is then calculated as

$$f = (1 - w/2)p_{\mathrm{c}} + (w/2)(1 - p_{\mathrm{u}})$$

where f is the overt confidence judgment, p_{c} is the probability that the chosen alternative is true, p_{u} is the probability that the unchosen alternative is true, and w is a weighting factor. (McKenzie, Wixted, Noelle, & Gyurjyan, 2001, present a number of equations that are similar to this one.)

In other words, the reported confidence that the chosen alternative is correct equals a weighted average of: (1) the probability that the chosen alternative is true, and (2) the probability that the unchosen alternative is false. The extent of underweighting of the alternative depends on the parameter w: no underweighting occurs when w equals 1, and maximal underweighting occurs when w equals 0. McKenzie (1997) showed that this sort of underweighting can yield an overconfidence effect similar to that observed in the literature.

Simulation

We simulated the UADVPM the same way we simulated the DVPM: assuming that people behave in a manner consistent with the UADVPM, we first generated data from the UADVPM. We then applied Juslin et al.'s (2000) scale-end correction procedure to estimate response error in the simulated data. If Juslin et al.'s procedure only corrects for response error, then changes in the weighting factor w should have no effect on the error correction. This is because changes to the cognitive underweighting bias do not influence the amount of response error in the data.

Results and discussion

Table 3 displays simulation results for different magnitudes of w across constant values of partition parameters. We observe that the estimated σ^2 values from Juslin et al.'s (2000) correction procedure (column 4) change considerably across values of w. Thus, besides correcting confidence data for response error, Juslin et al.'s procedure can also reduce or eliminate calibration effects that stem from a cognitive underweighting bias.

Table 3. Simulation results for the underweighted alternatives decision variable partition model

Parameters		Results			
Partitions	w	Original slope	Est. σ^2	Scale-end slope	Corrected slope
(0.1, 0.3, 0.4, 0.5, 0.7)	0	−0.815	0.194	−0.582	−0.232
	0.5	−0.755	0.044	−0.291	−0.464
	1	−0.731	0.028	−0.216	−0.515
(0.1, 0.3, 0.5, 0.7, 0.9)	0	−0.808	0.185	−0.576	−0.233
	0.5	−0.759	0.025	−0.194	−0.565
	1	−0.757	0.013	−0.104	−0.653

[4] Ferrell and McGoey (1980) describe application of the DVPM to evaluating the truth of a single alternative.

DOI: 10.1002/bdm

When a cognitive underweighting bias is involved in the generation of confidence data, response error estimates change as a function of the underweighting bias. This demonstrates that, while random error may be conceptualized as response error within a specific model, random error can also mask cognitive biases in the confidence elicitation process. As was the case for our other simulation models, this finding implies that the scale-end correction procedure may correct for more than simple response error in confidence data.

In the best-case scenario for the scale-end correction procedure, confidence judgments arise from the addition of random error to a true internal probability. This would imply a cognitive process that is similar to the response error model, in which case the procedure could plausibly yield accurate corrections. In the worst-case scenario for the procedure, confidence judgments arise from a complex confidence elicitation process, a process that includes some degree of underweighting. In this case, we have demonstrated that the discrepancy between the actual elicitation process and the response error model will likely render the resulting response error corrections meaningless.

Combined error model with confirmation bias

Up to now, we have described three systematic mechanisms that can be at least partially masked by random error: ecological error, partition placement, and a cognitive underweighting bias. These findings are problematic for the notion that random error effects are separable from systematic effects. If error estimates are influenced by systematic mechanisms, then we can draw no conclusions about the contribution of random error versus the contribution of systematic biases to confidence calibration.

While we have shown that error models can mimic data generated by many confidence elicitation mechanisms, the models cannot fit data from any mechanism. That is, the error models do put (weak) constraints on potential ‘‘true’’ confidence elicitation processes. To demonstrate this, we now simulate confidence data from a combined error model with a confirmation bias (CEMCB).

With regard to the confidence elicitation process, the term ‘‘confirmation bias’’ generally refers to a predisposition to rely on evidence supporting one’s choice in ignorance of evidence contradicting it. For example, consider a judge who is completing a series of two-alternative trivia questions. For a given question, the judge may choose Alternative A and then assess her confidence in that choice. In assessing confidence, the judge with a confirmation bias would be predisposed to consider only evidence supporting Alternative A, in ignorance of evidence against Alternative A. Koriat, Lichtenstein, and Fischhoff (1980) and Lee et al. (1995) describe confirmation biases and show how they can contribute to the overconfidence effect.

The CEMCB that we use in this section is very similar to the combined error model that we previously described in this paper. As before, ecological error is modeled via a binomial distribution, with the *n* parameter reflecting a judge’s experience with a specific cue or, in the confirmation bias case, the number of recruited arguments for a particular alternative. Also, response error is modeled via the addition of normal, random error to an internal confidence judgment. The new part of the model implements a confirmation bias, where judges create confidence judgments by focusing on evidence in support of the chosen alternative and disregarding evidence against the chosen alternative.

In the CEMCB, we assume that the judge first assesses evidence for each alternative using his or her knowledge of the environment (i.e., evidence for each alternative is obtained via the binomial distribution). In translating this evidence into a probability judgment, we implement the confirmation bias as

$$f = S/[S + (1 - \beta)A]$$

where S is evidence supporting the chosen alternative, A is evidence contradicting the chosen alternative, and β is a weighting factor representing the extent of confirmation bias. We then obtain the final, overt confidence judgment by adding random error to f.

In words, the extent to which a judge considers evidence contradicting the chosen alternative depends on the magnitude of β. When $\beta = 0$, the judge fully considers evidence against the chosen alternative, resulting

DOI: 10.1002/bdm

in no confirmation bias. When $\beta = 1$, the judge completely ignores evidence against the chosen alternative, resulting in a massive confirmation bias.

Simulation

We simulated the CEMCB in the same manner as for the other models. First, we generated data from the CEMCB using specific values of the model parameters n, σ^2, and β. Next, we applied the Juslin et al. (2000) procedure to estimate random error in the simulated data. If error models cannot mimic confirmation biases, then the estimated error variance and corresponding response error correction should be the same across different values of β.

Results and discussion

Table 4 displays simulation results for different values of β and n across a constant error variance σ^2. In contrast to the other models, where the estimated σ^2 and scale-end slopes varied greatly with model parameters, the estimated σ^2 and scale-end slopes here (columns 4 and 5) are similar from $\beta = 0$ (no confirmation bias) to $\beta = 0.5$ (sizeable confirmation bias). This implies that random error cannot mask all forms of confirmation biases present in overconfidence data.

To examine how our simulated data compared to empirical data, we explored the magnitudes of overconfidence present in our simulated data. Focusing on data for $n = 64$, the range of overconfidence varies greatly for different values of β. For example, when $\beta = 0$, experiment-wise overconfidence in our simulated data ranged from -0.01 to 0.03. When $\beta = 0.5$, however, experiment-wise overconfidence ranged from 0.05 to 0.15. These overconfidence ranges are generally higher than those observed by Juslin et al. (2000), which was approximately -0.1 to 0.1 (see their Figure 3).

These findings support the Juslin et al. (2000) statement that a confirmation bias ''is obtained when the bias is observed regardless of the proportion correct, or, at least, if we find a clear dominance of the bias for most levels of proportion correct'' (p. 388). Furthermore, the CEMCB simulations corroborate Juslin et al.'s claim that the balance of evidence does not favor this specific kind of cognitive bias for the judgment tasks considered here. Specifically, when Juslin et al. corrected overconfidence data with the response error model, the hard–easy effect disappeared. Because the response error model cannot mimic this form of confirmation bias, the correction must not have removed confirmation bias effects of this type. As a result, there is no remaining effect that can be attributed to this specific kind of confirmation bias.

Table 4. Simulation results for the combined error model with confirmation bias

Parameters		Results			
n	β	Original slope	Est. σ^2	Scale-end slope	Corrected slope
4	0	−0.502	0.037	−0.266	−0.236
	0.25	−0.552	0.041	−0.286	−0.266
	0.5	−0.656	0.047	−0.314	−0.342
8	0	−0.375	0.022	−0.177	−0.198
	0.25	−0.478	0.020	−0.166	−0.312
	0.5	−0.613	0.025	−0.192	−0.421
64	0	−0.140	0.013	−0.121	−0.019
	0.25	−0.285	0.010	−0.106	−0.179
	0.5	−0.455	0.017	−0.149	−0.306

DOI: 10.1002/bdm

In summary, we have found evidence that random error cannot mimic all forms of confirmation biases in the confidence elicitation process. While random error can mimic many systematic processes in confidence elicitation, it places some constraints on potential contributors to overconfidence.

GENERAL DISCUSSION

Error models of confidence calibration generally specify overconfidence as arising from the addition of random error to calibrated confidence judgments. The fact that these models provide good fits to confidence data implies that overconfidence may not be a ''real'' psychological phenomenon; overconfidence may only be an artifact of random error or other methodological problems. For example, Juslin et al. (2000) claim that there is little evidence of a cognitive-processing bias in the confidence elicitation process and that the hard–easy effect is an artifact related to linear dependency and scale-end (i.e., response error) effects. The intent of this paper has been to examine specific psychological processes that the error models can mimic. The approach involved simulating data from known mechanisms and examining the corrections across differing mechanisms. Our results lead us to conclude that error models can mimic confidence data that are generated by a number of systematic psychological processes. Thus, use of error models to disconfirm systematic biases in confidence judgments can be misleading because these procedures also eliminate systematic biases that are actually present in the judgments.

We began our examination of random error by investigating the magnitude of response error estimates that are typically produced by error models. By looking at response error estimates in terms of the percentage of misclassifications, we concluded that the response error estimates are very large, and, thus, may also be estimating systematic processes. Next, we simulated data from four models of confidence with specific parameter settings: the combined error model (Juslin et al., 1997), the DVPM (Ferrell & McGoey, 1980), a novel version of the DVPM that includes one kind of cognitive bias, and a novel version of the combined error model that includes a distinct cognitive bias.

For the first three models, we found that the magnitude of the response error estimate is influenced by model parameters that do not reflect response error. For the combined error model with a confirmation bias, we found that the response error estimate is largely unaffected by the magnitude of cognitive bias. Taken together, these findings are problematic for researchers who treat random error and cognitive biases as generally separable, distinct explanations for overconfidence. While random error cannot mimic all types of cognitive biases, it can mimic the effects of many systematic information processing mechanisms.

Overconfidence causes versus effects

In estimating the effects of different contributors to overconfidence, it is important to distinguish between systematic cognitive processing biases and systematic overconfidence effects. In describing the linear dependency and scale-end effects present in overconfidence data, Juslin et al. characterize cognitive biases as simple effects in the data across a range of proportions correct. For example, they state that the idea of a cognitive overconfidence bias is ''supported when there is a bias that covers most of the range of proportion correct, in a way that is not better accounted for by scale-end effects, linear dependency, or regression effects'' (p. 393).

An alternative definition of ''cognitive bias'' is a systematic bias in the way information is processed within the cognitive system. Such a bias may or may not lead to a simple, constant effect in the data. Indeed, it is quite plausible that such a systematic contributor within the confidence construction process, when combined with other aspects of the process, leads to complex effects in the data. This definition and the results of our simulation studies highlight the need to discriminate between properties of a generating process and properties of the resulting effects.

DOI: 10.1002/bdm

An important implication of this distinction is that attempts to separate effects that are due to systematic versus random components of the cognitive process are difficult and rely heavily on the particular model that is chosen. What the various investigations into random processes have contributed, on the other hand, is that mere observation of overconfidence in a data set does not, by itself, entail a systematic bias in the generating system. Together, these implications force us to consider that quite different approaches are required to advance our understanding of the nature of the cognitive processes underlying confidence judgment. Clever experimental design (e.g., Budescu et al., 1997) has proven to be one useful approach. The joint examination of multiple confidence measures could also prove useful. While some researchers have focused on modeling multiple confidence measures (e.g., Dougherty, 2001; Merkle & Van Zandt, 2006), Wallsten (1996) points out that researchers often place undue emphasis on overconfidence and the hard–easy effect. Finally, other procedures intended to separate out individual contributors to overconfidence should be subjected to the sort of detailed testing that we present in this paper.

Problems in confidence research

There are many mathematical models of confidence, and these models' explanations for overconfidence include ideas like random error, cognitive biases, and ecological validity. In reviewing these models, we can appreciate that the mere finding that some factor could yield overconfidence does not mean that that factor is necessarily at work in the confidence elicitation process. Furthermore, all models have weaknesses and are incomplete. It is probable that response error, systematic cognitive biases, and ecology all contribute to overconfidence effects to some degree. After taking into account all overconfidence results found in the research (e.g., alternative-underweighting effects, training effects, cultural effects, etc.), each class of models probably can explain some effects that no other class of models can explain.

That said, research involving error models often places random error in an elevated position over other explanations for overconfidence: the error model is used to account for random error in the elicitation process, and other explanations for overconfidence are validated only to the extent that an overconfidence effect remains. The current research highlights the fact that teasing apart the different contributors to overconfidence is more difficult than it may seem. While we can use models to estimate the effects of these different contributors, the extent to which these estimates are accurate is often unknown.

On a related note, the extent to which a model differs from the true elicitation process may also have an impact on parameter estimates (MacCallum, 2003; Ratcliff & Tuerlinckx, 2002; Van Zandt, Colonius, & Proctor, 2000; Van Zandt and Ratcliff, 1995; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). The true confidence elicitation process may be thought of as a large equation with many parameters, some of which are not identifiable. Only clever empirical designs, and not quantitative methods that control for random error, will allow us to isolate and investigate individual contributors to overconfidence. To this end, mathematical models of confidence that describe a plausible confidence elicitation process can help us to design relevant experiments. The effects of different parameters representing different aspects of the elicitation process can be assessed, and experiments can be designed based on these assessments. Such confidence models include not only the Combined Error Model, but also Minerva-Decision Making (Dougherty, 2001), Underweighted Alternatives Models (McKenzie, 1997), the Random Support Model (Brenner, 2003), the Exemplar Retrieval Model (Sieck, 2003; Sieck & Yates, 2001), and the Poisson Race Model (Merkle & Van Zandt, 2006). Furthermore, Sieck, Merkle, and Van Zandt (2007) present two experiments whose designs were based on the Assess-Search-Construct model. This model posits that familiarity drives choice of an alternative, which subsequently biases the confidence elicitation process. The model implies that confidence calibration will improve if participants explicitly judge the truth of each alternative during the confidence elicitation process, and the two experiments supported this implication.

DOI: 10.1002/bdm

CONCLUSIONS

The random error estimated by many confidence models may not be truly ‘‘random:’’ our analyses have shown that the magnitudes of the error estimates are highly influenced by factors that do not represent response error. That is, the error in confidence models is analogous to the error term in ANOVA and regression contexts, which represents other sources of variation that are not included in the model (e.g., Dean & Voss, 1999). While we agree that random error can play a role in the overconfidence phenomenon, our results imply that researchers who claim evidence against cognitive overconfidence biases have overstated their case. Response error, ecological validity, and cognitive biases likely all contribute to the overconfidence effect in some situations, and future studies can move toward examining the interplay between different contributors to overconfidence in specific situations. A careful consideration of a range of mathematical models of the confidence process will go a long way toward this goal.

ACKNOWLEDGEMENTS

This work was supported by Grants SES-9911301, SES-0214574, and SES-0437251 from the National Science Foundation. Portions of this work were presented at the 25th Meeting of the Society for Judgment and Decision Making, Minneapolis, MN.

REFERENCES

Arkes, H. R., Dawson, N. V., Speroff, T., Harrell, F. E., Jr., Alzola, C., Phillips, R., & et al. (1995). The covariance decomposition of the probability score and its use in evaluating prognostic estimates. *Medical Decision Making*, *15*, 120–131.

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes*, *90*, 87–110.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*, 173–188.

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer-Verlag.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, *130*, 579–599.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Decision Processes*, *26*, 32–53.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*, 226–246.

Juslin, P., Olsson, H., & Bjorkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189–209.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A ‘‘lazy’’ algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*, 384–396.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216–247.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.

DOI: 10.1002/bdm

Lee, J. W., Yates, J. F., Shinotsuka, H., Singh, R., Onglatco, M. L. U., Yen, N.-S., & et al. (1995). Cross-national differences in overconfidence. *Asian Journal of Psychology*, *1*, 63–69.

Levine, M. (1975). *A cognitive theory of learning: Research on hypothesis testing*. Hillsdale, NJ: LEA.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance*, *20*, 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149–171.

MacCallum, R. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113–139.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, *71*, 141–160.

McKenzie, C. R. M., Wixted, J. T., Noelle, D. C., & Gyurjyan, G. (2001). Relation between confidence in yes-no and forced-choice tasks. *Journal of Experimental Psychology: General*, *130*, 140–155.

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*, 391–408.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., & et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester, UK: Wiley.

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes*, *58*, 203–213.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *107*, 358–367.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review*, *9*, 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General*, *124*, 352–374.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? *Psychological Review*, *107*, 358–367.

Sieck, W. R. (2003). Effects of choice and relative frequency elicitation on overconfidence: Further tests of an exemplar-retrieval model. *Journal of Behavioral Decision Making*, *16*, 127–145.

Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: A cognitive contributor to overconfidence. *Organizational Behavior and Human Decision Processes*, *103*, 68–83.

Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: A comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1003–1021.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201–221.

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin and Review*, *2*, 20–54.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin and Review*, *7*, 208–256.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes*, *65*, 220–226.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, *101*, 490–504.

Wells, G. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, *66*, 688–696.

Authors' biographies:

Edgar C. Merkle is Assistant Professor of Psychology at Wichita State University. He earned a Ph.D. in Quantitative Psychology from The Ohio State University. His interests include subjective probability and statistical computing.

DOI: 10.1002/bdm

Winston R. Sieck is currently a Principal Scientist working at the Klein Associates Division of ARA. He combines cognitive field research with experimentation and quantitative modeling to address basic and applied issues in culture, judgment, and decision making.

Trisha Van Zandt is Associate Professor of Psychology at The Ohio State University. She earned a Ph.D. in Quantitative Psychology from the Purdue University. She studies mathematical models of choice and statistical methods.

Authors' addresses:

Edgar C. Merkle, Department of Psychology, Wichita State University, 1845 Fairmount Box 34, Wichita, Kansas, 67260, USA.

Winston R. Sieck, Applied Research Associates, 1750 Commerce Center Blvd. North, Fairborn, OH, 45324-3987, USA.

Trisha Van Zandt, Department of Psychology, The Ohio State University, Columbus, OH, 43210, USA.

DOI: 10.1002/bdm