

Naive Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard–Easy Effect

Peter Juslin, Anders Winman, and Henrik Olsson
Uppsala University

Two robust phenomena in research on confidence in one's general knowledge are the *overconfidence phenomenon* and the *hard–easy effect*. In this article, the authors propose that the hard–easy effect has been interpreted with insufficient attention to the scale-end effects, the linear dependency, and the regression effects in data and that the continued adherence to the idea of a “cognitive overconfidence bias” is mediated by selective attention to particular data sets. A quantitative review of studies with 2-alternative general knowledge items demonstrates that, contrary to widespread belief, there is (a) very little support for a cognitive-processing bias in these data; (b) a difference between representative and selected item samples that is not reducible to the difference in difficulty; and (c) near elimination of the hard–easy effect when there is control for scale-end effects and linear dependency.

Two well-known threats to scientific progress are naive empiricism and dogmatism. When one tries to explain to an untutored mind that the earth is round, one might be objected to by reference to the fact that the horizon looks flat to the naked eye. This is an illustration of *naive empiricism*, the uncritical acceptance of empirical observation. The classic example of *dogmatism*, theological conceptions that are upheld in the face of ever-increasing evidence to the contrary, is the scholastic reaction to the new cosmology advanced at the dawn of the modern age. Both of these examples benefit from the safety of a hindsight perspective, and, admittedly, there may exist no clear criterion delineating naive empiricism from mature science or sound skepticism from dogmatism. Nevertheless, there is little doubt that these two threats are genuine and serious problems in theory formation and methodology. When both problems co-occur, or even reinforce one another, things become particularly complicated.

The *overconfidence phenomenon* refers to the observation that the mean subjective probability (\bar{x}) assigned to the correctness of answers to general knowledge items like “Which country has the larger population: (a) Finland or (b) Zambia?” tends to exceed the proportion (\bar{c}) of correct answers ($\bar{x} - \bar{c} > 0$). The common observation of overconfidence has inspired ideas of information-processing biases. For example, it has been hypothesized that people are victims of selective retrieval of supporting evidence (Koriat, Lichtenstein, & Fischhoff, 1980), insufficient cognitive processing (Sniezek, Paese, & Switzer, 1990), overreliance on the

strength rather than the weight of evidence (Griffin & Tversky, 1992), and self-serving motivational biases (Taylor & Brown, 1988). The *hard–easy effect* refers to a covariation between over/underconfidence and task difficulty; overconfidence is more common for hard item samples, whereas underconfidence is more common for easy item samples.

In the early 1990s, the interpretation of overconfidence in terms of information-processing biases (e.g., confirmation biases) was challenged on two separate grounds. First, proponents of the so-called ecological models (Björkman, 1994; Gigerenzer, Hoffrage, & Kleinböling, 1991; Juslin, 1993a, 1993b, 1994) suggested that overconfidence could be a side effect of biased, or nonrepresentative, selection of items. Second, it was shown that “overconfidence” can arise as mere regression effects (the error models; Erev, Wallsten, & Budescu, 1994; Pfeifer, 1994; Soll, 1996; see also Dawes & Mulford, 1996).

The studies reported in support of the hypothesis that representative item selection decreases or even eliminates overconfidence (Gigerenzer et al., 1991; Juslin, 1993a, 1993b, 1994, 1995; Juslin, Olsson, & Björkman, 1997; Juslin, Winman, & Persson, 1995; Kleitman & Stankov, 1996; Winman, 1997a, 1997b) were soon dismissed, however, on the grounds that representative item selection was confounded with the hard–easy effect, that is, the representative samples were too easy to disclose the overconfidence phenomenon (Griffin & Tversky, 1992). This proposal elicited a burst of studies with difficult item samples that produced overconfidence, allegedly refuting the ecological models and demonstrating the realness of overconfidence (e.g., Brenner, Koehler, Liberman, & Tversky, 1996; Budescu, Wallsten, & Au, 1997; Griffin & Tversky, 1992; Suantak, Bolger, & Ferrell, 1996). These results—essentially amounting to the hard–easy effect—provided the principal support for a number of often-cited theoretical models (e.g., Griffin & Tversky, 1992; Suantak et al., 1996).

In this article, we propose that the hard–easy effect has been interpreted with insufficient attention to important methodological problems (something that also undoubtedly applies to our own research). In this sense, we have been victims of something akin to naive empiricism. Second, we demonstrate that, contrary to wide-

Peter Juslin, Anders Winman, and Henrik Olsson, Department of Psychology, Uppsala University, Uppsala, Sweden.

Henrik Olsson is now at the Department of Psychology, Umeå University, Umeå, Sweden.

This research was supported by the Swedish Council for Research in the Humanities and Social Sciences. We are indebted to Mats Björkman, Nils Olsson, Magnus Persson, and Pia Wennerholm for helpful discussions.

Correspondence concerning this article should be addressed to Peter Juslin, who is now at the Department of Psychology, Umeå University, SE-901 87 Umeå, Sweden. Electronic mail may be sent to peter.juslin@psy.umu.se.

spread belief, there is little data to support the hypothesis of a cognitive overconfidence bias when people respond to general knowledge items. With general knowledge items, the idea of an information-processing bias is approaching the status of a dogma, supported by naive empiricism and selective attention to particular data sets. Third, we demonstrate that there is a difference between selected and representative item samples, as predicted by the ecological models. In this article, we develop these arguments and substantiate our claims through a quantitative review of empirical data, which also shows that when we control for two of the methodological problems discussed in this article, the hard–easy effect is almost eliminated.

Naive Empiricism and the Hard–Easy Effect

The paradigmatic task in studies of the overconfidence phenomenon presents participants with two–alternative general knowledge items. For each item, the participants decide on one of the two answers and assess their confidence in their decision on a probability scale between .5 (*random choice*) and 1.0 (*certainty*). The participants are *realistic*, or *calibrated*, if, across items assigned a subjective probability of .*xx*, a proportion .*xx* is correct (Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990). The hard–easy effect was highlighted by Lichtenstein and Fischhoff (1977), who reported overconfidence for general knowledge items with a proportion of correctly answered items (\bar{c}) below approximately .75 and underconfidence for items with a proportion correct (\bar{c}) above .75. (Other names for the hard–easy effect are the “difficulty effect” [Griffin & Tversky, 1992] and the “discriminability effect” [Ferrell & McGoey, 1980].)

Reviews of the literature have concluded that the hard–easy effect is a robust and pervasive phenomenon (Keren, 1991; Lichtenstein et al., 1982; McClelland & Bolger, 1994; Wallsten & Budescu, 1983; Yates, 1990). A number of recent articles have been devoted to the psychological explanation of the phenomenon (Baranski & Petrusic, 1994; Griffin & Tversky, 1992; Suantak et al., 1996). The importance of the hard–easy effect was restated in a recent commentary: “The two major substantive and pervasive findings are overconfidence and the interaction between the amount of overconfidence and difficulty of the task, the so-called hard–easy effect” (Keren, 1997, p. 269). Although there have been occasional concerns about the effect (Brenner et al., 1996; Juslin et

al., 1997; Keren, 1991), it is hard to avoid the conclusion that it is often regarded as *the* principal finding in calibration research (Baranski & Petrusic, 1994, 1997; Griffin & Tversky, 1992; Suantak et al., 1996).

In the following section, we discuss three methodological problems associated with the hard–easy effect: scale-end effects, linear dependency, and regression effects. The important concepts and definitions are summarized in Table 1. All three of these problems are sufficient—alone or in combination—to produce an apparent hard–easy effect in the data. These problems have not been clearly distinguished in the literature, and their full importance has not been appreciated, as exemplified above.

Scale-End Effects

For two–alternative items, the over/underconfidence score is defined as the difference between the mean subjective probability assigned to the chosen answer and the proportion of correct answers, $\bar{x} - \bar{c}$. As such, there are definite mathematical constraints on the values that the score can take. In Figure 1A, the area between the upper and lower lines represents the region of possible values. Because the confidence scale starts at .5, when the proportion correct is .5 or less, the over/underconfidence score can only be zero or positive (overconfidence), attaining its maximum when the mean subjective probability is 1.0. When the proportion correct is 1.0, the over/underconfidence score can only be zero or negative (underconfidence), with a minimum of $-.5$ for a mean subjective probability of .5.

Any fitted linear function with proportion correct as the independent variable and over/underconfidence as the dependent variable that covers the entire interval (.5, 1) will have a zero or negative slope, with a crossover between over- and underconfidence somewhere in the interval (i.e., the correlation is zero or negative). Now consider a response error, e_{rt} , in the overt assessment, x_t , of the “true” subjective probability, T_t , at assessment trial t , that is, $x_t = T_t + e_{rt}$ (see Table 1). One limiting case is that in which all subjective probability assessments are perfectly calibrated, with no response error whatsoever in the overt expression, where the slope is zero. As soon as we enter a response error at the elicitation stage—or individual differences across participants in how confidence is mapped onto the scale, for that matter (other sources of error are also possible)—the slope will turn negative

Table 1
Summary of the Methodological Problems Associated With the Hard–Easy Effect

| Concept | Methodological problem | | | |
|-----------------------------|---|--|---|--|
| | Scale-end effect | Linear dependency | Measurement regression | Population regression |
| Error or origin of variance | Response error in overt probability assessments | Measurement error: proportion correct, \bar{c} | Measurement error: mean confidence, \bar{x} , and proportion correct, \bar{c} | Deviations between the population values \bar{X} and \bar{C} |
| Definition | $T_t - x_t = e_{rt}$ | $\bar{C} - \bar{c} = e_{\bar{c}}$ | $\bar{C} - \bar{c} = e_{\bar{c}}$ and $\bar{X} - \bar{x} = e_{\bar{x}}$ | $\bar{X} - \bar{C} = E$ |
| Locus | Response elicitation–measurement | Response elicitation–measurement | Response elicitation–measurement | Population values |

Note. T_t = true subjective probability assessment at assessment trial t ; x_t = overt probability assessment at assessment trial t ; e_{rt} = response error at assessment j ; \bar{C} = population proportion correct; \bar{c} = observed proportion correct; $e_{\bar{c}}$ = measurement error for proportion correct; \bar{X} = population mean confidence; \bar{x} = observed mean confidence; $e_{\bar{x}}$ = measurement error for mean confidence; E = deviation between population values for mean confidence and proportion correct (over/underconfidence).

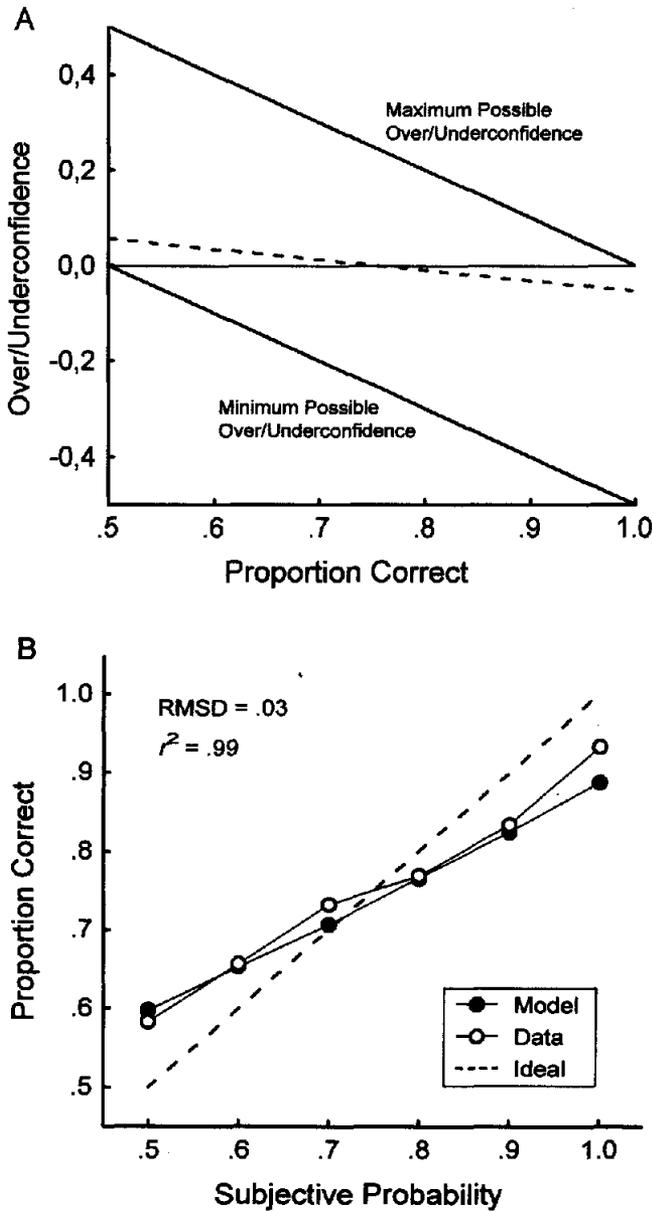


Figure 1. The region between the two lines in Panel A represents the possible values for the over/underconfidence score. The dashed line is the hard-easy effect predicted from scale-end effects alone, where the response error variance was estimated from the data collected in our laboratory (see the *Correction for scale-end effects* section of the text for details). Panel B presents a calibration curve based on all the data collected in our laboratory (almost 40,000 responses), along with the calibration curve predicted by the combined error model when fitted to these data. RMSD = root-mean-square deviation.

(i.e., because at the ends of the probability scale the errors have only one way to go). In the other extreme case—overt probability assessments that are uniformly distributed across the probability scale regardless of the proportion correct of the item sample—the linear function will have a negative slope of -1 with crossover at a proportion correct of .75. The most reasonable hypothesis, per-

haps, is that subjective probability is related to accuracy but that there is a response error in the use of the scale, suggesting a slope somewhere between 0 and -1 .

In a calibration diagram (see Figure 1B), the proportions correct are plotted against subjective probability to produce a calibration curve. A direct consequence of the response error is the rotation of the calibration curve illustrated in Figure 1B, in which the center of rotation is located close to the midpoint of the probability scale (i.e., .75 for two-alternative items). This rotation, or “regression,” of the curve implies a proportion correct greater than .5 in the subjective probability category of .5 and less than 1 in the subjective probability category of 1, even if the underlying judgments are unbiased. This effect is routinely observed in empirical calibration curves, suggesting a nontrivial response error in the overt assessment of subjective probabilities. When combined with the salient endpoints of the probability scale, this error alone will produce a hard-easy effect. (The hard-easy effect predicted from scale-end effects alone is represented by the dashed line in Figure 1A. Figure 1 is further commented on in connection with the quantitative review presented below.) Simulations in Juslin et al.’s (1997) article illustrated that this effect, hardly interpretable in terms of information-processing biases, is associated with a crossover between over- and underconfidence bias close to a proportion correct of .75. This is what has been observed in the empirical data (Juslin et al., 1997; Suantak et al., 1996).

Linear Dependency

In previous publications (Juslin et al., 1997, p. 193, footnote 2; Juslin, Olsson, & Winman, 1998, p. 20), researchers have pointed out that the linear dependency between proportion correct, \bar{c} , and over/underconfidence, $\bar{x} - \bar{c}$, is a second factor that may contribute to a hard-easy effect. Linear dependency is concerned with the measurement error for proportion correct (see Table 1), that is, the deviations (error), e_{ci} , between the population proportion correct, \bar{C}_i , and the observed proportion correct, \bar{c}_i , for observation unit i . Table 2 provides a schematic illustration in which the four units of measurement may be different participants, different target vari-

Table 2
Schematic Example of How a Hard-Easy Effect Arises From the Linear Dependency Between Proportion Correct and Over/Underconfidence

| Observation unit i | Variable | | | | | |
|----------------------|-------------|-------------|-------------|----------|-------------|-------------------------|
| | \bar{X}_i | \bar{C}_i | \bar{x}_i | e_{ci} | \bar{c}_i | $\bar{x}_i - \bar{c}_i$ |
| 1 | .75 | .75 | .75 | .05 | .80 | -.05 |
| 2 | .75 | .75 | .75 | -.05 | .70 | .05 |
| 3 | .75 | .75 | .75 | .05 | .80 | -.05 |
| 4 | .75 | .75 | .75 | -.05 | .70 | .05 |

Note. The (fictional) data here show a correlation of -1.00 between the proportion correct and the over/underconfidence score as a result of the correlated measurement errors alone (see the *Linear Dependency* section of the text for an explanation). \bar{X}_i = population mean subjective probability for unit i ; \bar{C}_i = population proportion correct for unit i ; \bar{x}_i = observed mean subjective probability for unit i (given no measurement error for subjective probability); e_{ci} = measurement error for the proportion correct of unit i ; \bar{c}_i = observed proportion correct for unit i ; $\bar{x}_i - \bar{c}_i$ = observed over/underconfidence score for unit i .

ables in the judgment task (e.g., population of countries or area of countries), or some other way to partition the data in a calibration study. In Table 2, \bar{X}_i is the population mean confidence for unit i , \bar{x}_i is the observed mean confidence for unit i , and $\bar{x}_i - \bar{c}_i$ is the observed over/underconfidence score for unit i . For illustrative purposes, we made two simplifying assumptions in Table 2: There is no measurement error with regard to mean subjective probability ($\bar{x}_i = \bar{X}_i$), and all units have the same population figures for proportion correct and mean subjective probability ($\bar{X}_i = \bar{C}_i = .75$).

In Table 2, the correlation between the proportion correct and the over/underconfidence score is -1.00 , with a negative slope equal to -1 and a crossover between over- and underconfidence at a proportion correct of $.75$. Three things are noteworthy about this example. First, there is no bias or hard–easy effect in the population. Second, there is no measurement error for the mean subjective probability (i.e., compare with the aforementioned discussion on scale-end effects). Third, there is no error in the relation between the population values for mean subjective probability and proportion correct, so this is not a regression effect with regard to the population values (see further discussion in the *Regression Effects* section). Nevertheless, measurement units (participants, judgment domains, etc.) with a low proportion correct will appear overconfident, whereas measurement units with a high proportion correct will appear underconfident.

With measurement error also for mean subjective probability, the correlation will rise above -1.00 , but as long as the errors are independent, the correlation is negative. Moreover, the measurement error for proportion correct will be particularly large because of its sensitivity to sampling error in the item selection (Juslin et al., 1998; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999),¹ and larger error variance is indeed observed for proportion correct than for mean subjective probability (e.g., Dawes & Mulford, 1996; Juslin, 1993b). Because the error variance for proportion correct is larger than the error variance for mean subjective probability, most of the error for the over/underconfidence score will be accountable in terms of the error for proportion correct. This, in turn, means that a correlation between proportion correct and the over/underconfidence score that is not clearly negative is indeed a surprising event (the correlated errors also make the interpretation of correlation and regression analysis with computed p values in a number of studies problematic; e.g., Arkes, Christensen, Lai, & Blumer, 1987; Ayton & Wright, 1990; Björkman, 1992). In sum, we expected a “hard–easy effect” because of linear dependency alone, again with a crossover close to a proportion correct of $.75$, as observed in the data.

Regression Effects

A trend in recent calibration research, primarily stimulated by an article by Erev et al. (1994), has been to point out how over- and underconfidence may arise from regression effects. The argument is that because of the merely correlative relationship between subjective and objective probabilities, there will be regression when one of the variables is plotted against the other. In calibration studies, in which objective probabilities are plotted against subjective probabilities, the regression will most often contribute to “overconfidence.” Moreover, this “overconfidence bias” will be particularly pronounced for tasks with a low proportion correct,

thus producing a further source of hard–easy effects in the data (see also Dawes & Mulford, 1996; Pfeifer, 1994; Soll, 1996).

This basic idea can be interpreted in two slightly different ways: (a) *population regression*, a regression that arises because the population values of the units (i.e., \bar{X}_i and \bar{C}_i in Tables 1 and 2) are merely correlated, and (b) *measurement regression*, a regression that arises because of measurement error (i.e., e_c and e_x in Table 1) and that includes the effects of response error and measurement error in proportion correct. We interpret the discussion in, for example, Dawes and Mulford (1996) to be concerned with both kinds of regression, whereas the discussion in Erev et al. (1994) seems to concentrate on the latter kind of regression.

The distinction between population and measurement regression has some importance. Whereas scale-end effects, linear dependency, and measurement regression primarily can be interpreted as artifacts that arise in the context of the observation, population regression seems like a more genuine effect. Measurement units (participants, judgment domains, etc.) that have a low proportion correct will have a higher mean subjective probability in a replicable and robust manner, and vice versa for units with a high proportion correct. Even if this is a more real hard–easy effect, it is debatable whether it is properly addressed by the notion of an information-processing bias. For instance, will the confirmatory search of memory (Koriat et al., 1980; McKenzie, 1997) turn into a disconfirmatory search for tasks that have a proportion correct that exceeds $.75$? Regression effects, of course, do not preclude genuinely cognitive interpretations (e.g., as noise in the memory process) but rather confound psychological, statistical, and environmental influences in a highly intricate manner.

In sum, the factors compiled in Table 1, scale-end effects, linear dependency, and regression effects, are sufficient—alone or in combination—for the observation of a hard–easy effect, in general with a crossover in the region of a proportion correct of $.75$. The assumptions are not radical: (a) variability in the subjective probability responses, either within the participants (e.g., a response error) or across the participants (e.g., in the mappings of confidence onto the probability scale), (b) a measurement error for proportion correct, and (c) a merely correlative relation between the mean subjective probability and the proportion correct of the measurement units. With a few exceptions (Budescu et al., 1997; Juslin et al., 1997, 1998; Klayman et al., 1999),² we were unable to find studies that controlled for even one of these three problems.

The naïveté with which the hard–easy effect has been interpreted is an obstacle to theoretical progress. For example, there is still no clear picture of the magnitude of the real hard–easy effect, the confirmation of which provides the principal support for a

¹ Participants often rely on probabilistic inferences to answer general knowledge items (see, e.g., Gigerenzer et al., 1991), and these inferences can be applied to a large number of distinct items, some of which are successful applications and some of which lead to the wrong answer. In the item selection, one might accidentally come up with unusually many of the first or the second kind of items, thus contributing to a sizable sampling error in proportion correct (see Juslin et al., 1998, for a discussion).

² Klayman et al. (1999) is the only study we know of that controlled for linear dependency. Juslin et al. (1997, 1998; Juslin, Wennerholm, & Olsson, 1999) explicitly modeled the end effects associated with response error, and Budescu et al. (1997) modeled, and corrected the data for, a stochastic component similar to the response error discussed in this article.

number of theoretical models (Griffin & Tversky, 1992; Suantak et al., 1996). However, the problem is only further aggravated by the fact that these results are also—and forcefully so—used to support one of the claims made by an influential research program.

Dogmatism and Cognitive Overconfidence Bias

Overconfidence in human judgment has developed into an established fact of psychology ubiquitously found in introductory textbooks (e.g., Myers, 1997; Plous, 1993; Sternberg, 1996), in which it is explained by a variety of psychological mechanisms. Recently, two journals specializing in judgment research devoted entire issues to research on calibration and primarily discussions of overconfidence (*Journal of Behavioral Decision Making*, 1997, Vol. 10, No. 3, and *Organizational Behavior and Human Decision Processes*, 1996, Vol. 65, No. 3). The commentaries in those issues concluded that, despite the recent criticisms of the overconfidence phenomenon, there is compelling evidence for the realness of an overconfidence bias in human judgment. We provide one example from Keren (1997), but the other commentaries came to similar conclusions: "There are now sufficient empirical studies demonstrating overconfidence even when items were carefully sampled in a random manner. . . . There is also sufficient evidence to dismiss the claim that overconfidence is entirely a statistical artifact" (p. 274). It seems fair to conclude that, although the recent criticisms in terms of biased item selection and regression effects have received some attention, there remains a pervasive majority opinion that overconfidence is a real and fundamental property of people's confidence in their general knowledge.

Given these conclusions, it may be worthwhile to scrutinize the data from studies with general knowledge items. Considering the aforementioned problems with the hard–easy effect, we propose the following criterion: Unequivocal evidence in favor of an information-processing 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 proportions correct.³ For proportions correct less than .75, we expect overconfidence for a multitude of reasons that have nothing to do with a cognitive-processing bias, and for item samples with a proportion correct greater than .75, we expect an underconfidence bias for similar reasons. Given the aforementioned strong conclusions, we would expect there to be plenty of data around with overconfidence in the medium and high regions of proportion correct, where the interpretation is most unequivocal.

In our review of studies with two-alternative general knowledge items (presented more extensively in the *What Is in the Empirical Data?* section), we were unable to find a *single* study with representative item selection, a proportion correct greater than .75, and a significant overconfidence bias. For example, Griffin and Tversky (1992) relied on three judgment topics, with proportions correct of .68 (population of U.S. states), .51 (voting rate in U.S. states), and .49 (education level in U.S. states). The observation of a mean subjective probability (on a scale between .5 and 1.0) greater than the proportion correct was taken as a "refutation" (cf. Griffin & Tversky, 1992, p. 411) of the ecological model in the form of probabilistic mental theory (Gigerenzer et al., 1991) and as a demonstration of the realness of overconfidence. By the same logic, of course, we could demonstrate a cognitive underconfidence bias by repeated observation of tasks with a proportion

correct close to 1, where—according to the hard–easy effect (whatever its true nature)—there will be "underconfidence."

We submit that with two-alternative general knowledge items there is little or no evidence for an information-processing bias in human judgment. In the next section, we substantiate this conclusion through a more careful examination of the empirical data.

What Is in the Empirical Data?

A crucial distinction in recent research and debate in the overconfidence literature is that between selected and representative item samples. The central argument presented by the ecological models (e.g., Gigerenzer et al., 1991; Juslin, 1993b)⁴ is that the item samples in traditional overconfidence studies have been generated in a way that inadvertently overrepresents those "trick items" for which the probabilistic inferences used by the participants lead to the wrong answer, at the expense of items for which the same inferences lead to the correct answer. The item-selection procedures involved in putting someone's knowledge to the test and the salience of surprising and interesting facts lead to item samples for which knowledge that is valid and useful in the natural (unselected) environment becomes less valid. Because the confidence judgments are (roughly) attuned to the validity of the inferences in a natural environment, the participants appear "overconfident" for these selected samples (see, e.g., Gigerenzer et al., 1991; Juslin, 1994; Juslin et al., 1997, for further details).

To test this conjecture, selected item samples have been contrasted with representative item samples. Representative item samples are generated in two steps: (a) A natural environment is defined in terms of a population of environmental objects (e.g., all German cities with more than 100,000 inhabitants, all world countries, all U.S. states), and (b) the objects of judgment (e.g., cities, countries, states) are randomly selected from this natural environment. The prediction by the ecological models is that confidence should be approximately the same in selected and representative item samples but that the proportion correct should be lower in the selected samples, yielding the overconfidence phenomenon.

Initial studies with representative item samples reported over/underconfidence biases close to zero at proportions correct in the interval of .7 to .8 (Gigerenzer et al., 1991; Juslin, 1994). Following the lead of Griffin and Tversky (1992), these results were dismissed on the grounds of confounding with the hard–easy effect:

The difficulty effect is one of the most robust findings in the calibration literature. . . . The difficulty effect can also explain the main findings of a study by Gigerenzer, Hoffrage & Kleinbölting (1991) . . . [who found that] average accuracy was 72% for the city judgments and only .53 for the general knowledge items. Hence, the

³ One example of the application of this criterion can be found in Juslin et al.'s (1998) article, in which it was applied to confidence in sensory discrimination (see Juslin & Olsson, 1997).

⁴ The most well-known, elaborate, and elegant formulation of these Brunswik-inspired ideas (e.g., Brunswik, 1956) is the theory of probabilistic mental models presented by Gigerenzer et al. (1991). At the time, similar ideas were developed in our lab and later were published (Björkman, 1994; Juslin, 1993a, 1993b, 1994). In a review (McClelland & Bolger, 1994), these approaches were referred to as the ecological models.

presence of overconfidence in the latter but not the former could be entirely due to the difficulty effect. (Griffin & Tversky, 1992, pp. 427–428)

This argument is routinely repeated in discussions of the topic (e.g., Kahneman & Tversky, 1996; McClelland & Bolger, 1994; Suantak et al., 1996). Griffin and Tversky may have been correct when they pointed out that with a single pair of data points, the predictions for representative and selected item samples were confounded with the hard–easy effect. By now, sufficient data have accumulated to analyze distributions of mean confidence and proportions correct across studies, thereby allowing us to resolve this issue.

A Cognitive Overconfidence Bias?

To test the idea of a cognitive overconfidence bias against a larger database, we performed a quantitative review of the data available to us (in June 1998). The criteria for inclusion in the analysis were as follows: (a) The data had to be based on two-alternative forced-choice items with subjective probability assessments on a probability scale from .5 to 1 (i.e., the half-range format); (b) the tasks had to be tests of general knowledge, that is, had to concern declarative facts with which the participants had no extensive professional experience; (c) there could not be a statistically significant effect of any independent variable on the data (e.g., debiasing and mood induction); and (d) both over/underconfidence scores and proportions correct had to be reported or deducible from the reported data.

These criteria may seem strict, but they coincide with the standard paradigm in studies of overconfidence in general knowledge and thus include the vast majority of the relevant data. The studies included in the analysis contained 95 independent data points with selected item samples and 35 independent data points with representative item samples, where *independence* refers to different participant samples (i.e., all data points were between subjects). For an item sample to be classified as “representative,” the study had to contain explicit statements that (a) a natural environment had been defined and (b) the items had been generated by random sampling from this environment. Any study that lacked a detailed description of the item-selection procedures or that explicitly described some other procedure was thus classified as a “selected item sample.” It is important to acknowledge the asymmetry in stimulus control for selected and representative item samples. We have much less knowledge of how the selected samples have been generated, and they may therefore be more heterogeneous (i.e., some item samples may be extremely selected for misleadingness, whereas others may effectively coincide with representative item samples). Note that this lack of stimulus control can only be conservative with regard to a test of a difference between selected and representative item samples.

Table 3 presents the mean subjective probability, the mean proportion correct, the mean over/underconfidence score, and the mean absolute bias, where “absolute bias” is over/underconfidence with sign ignored, for the selected and representative item samples, respectively. The means are presented with standard deviations and 95% confidence intervals (CIs) across the independent data sets ($n = 95$ for the selected samples and $n = 35$ for the representative samples). As we expected from the argument presented

by the ecological models, the mean subjective probability for the selected and representative item samples coincided, indeed with high precision (.73; 95% CIs of $\pm .01$ and $\pm .02$, respectively).

In Table 3, one can see that the proportion correct for representative item samples, .72, agreed quite well with the mean subjective probability, thus leading to an over/underconfidence score of almost zero (.01; 95% CI of $\pm .02$). In contrast, there was substantial overconfidence for the selected item samples (.09; 95% CI of $\pm .02$). The last row of Table 3 shows that the over/underconfidence score of almost zero for representative item samples was not a mere consequence of averaging. For the selected item samples, the mean absolute bias was .10 (95% CI of $\pm .01$); for the representative item samples, the mean absolute bias was .03 (95% CI of $\pm .01$). Hence, although there was some bias in the representative samples, the effects were modest and much lower than for the selected samples (i.e., for most of the representative item samples, the bias was statistically nonsignificant).⁵ Later in this article, we investigate whether scale-end effects and linear dependency account for the bias that remains with representative item samples.

A problem with evaluating the argument that the predictions by the ecological models are confounded with the hard–easy effect is that with representative samples we seldom expect very low proportions correct, whereas a low proportion correct is a characteristic attribute of selected item sets. For example, in the data that we compiled for the quantitative review, there were 12 (out of 35; 34%) representative item samples with proportions correct less than .7 but 64 (out of 95; 67%) selected samples with proportions correct less than .7, with generally lower proportions correct in the latter category.⁶ We controlled for difficulty by means of a simple matching procedure for item samples with proportions correct less than .7 (hard samples). For each representative item sample with a proportion correct less than .7, we entered all selected item sam-

⁵ Because there were relatively more studies with Swedish participants among the representative samples and more studies with U.S. participants among the selected samples, one might object that the observed difference reflects cultural differences, as reported by Svenson (1981). There was no significant difference between the data for representative item samples collected by Swedish participants and non-Swedish participants, nor for selected item samples. There was a significant difference between selected and representative item samples both in the studies with Swedish participants and non-Swedish (mainly U.S.) participants, so the difference was not explained by cultural differences. The difference was somewhat larger in the data for Swedish participants, in which the comparison primarily involved representative samples collected in our lab in Uppsala and selected item samples collected by Allwood and colleagues in Gothenburg (Allwood, 1994; Allwood & Björhag, 1990, 1991; Allwood & Granhag, 1996a, 1996b, 1996c; Allwood & Montgomery, 1987; Granhag, Strömwall, & Allwood, 1998). This difference was paralleled by a large difference in proportion correct, and for the reasons discussed in this article (scale-end effects, linear dependency, and regression), this difference led to the expectation of a larger difference also in the over/underconfidence score.

It should be noted, however, that there seem to be real differences in probability judgments between different cultures, such as between Asians and Americans (see Yates, Lee, & Bush, 1997). Whether these differences reflect discrepancies in how the response scale is used or deeper underlying psychological differences remains unknown.

⁶ It might seem that one solution to this problem would be to enter representative versus selected item samples as an independent variable in

Table 3
Means, 95% Confidence Intervals (CIs), and Standard Deviations for Selected ($n = 95$) and Representative ($n = 35$) Item Samples Collected in the Quantitative Review

| Dependent measure | Method for item selection | | | | | |
|-----------------------------|---------------------------|------|-----------|-----------------------------|------|-----------|
| | Selected item samples | | | Representative item samples | | |
| | <i>M</i> | CI | <i>SD</i> | <i>M</i> | CI | <i>SD</i> |
| Mean subjective probability | .73 | ±.01 | .04 | .73 | ±.02 | .04 |
| Proportion correct | .64 | ±.02 | .08 | .72 | ±.02 | .07 |
| Over/underconfidence | .09 | ±.02 | .08 | .01 | ±.02 | .05 |
| Absolute bias | .10 | ±.01 | .05 | .03 | ±.01 | .03 |

Note. Absolute bias is the over/underconfidence score with the sign ignored.

ples with the same proportion correct (i.e., as judged by two decimals). The mean over/underconfidence for selected hard items was .10 (95% CI = ±.02, proportion correct = .65, $n = 29$), and the mean over/underconfidence for representative hard items was .05 (95% CI = ±.02, proportion correct = .65, $n = 12$).

The matching of items from two item sets of unequal difficulty may in itself be susceptible to regression effects. However, the data points represent entire studies and have extremely small standard errors of measurement, so this effect is presumably very modest. Moreover, by virtue of the linear dependency, any measurement error in proportion correct, \bar{c} , is paralleled by a corresponding effect on the over/underconfidence score, $\bar{x} - \bar{c}$, which is conservative with regard to the interpretations made here. Any measurement error for a selected sample that leads to a \bar{c} that is too high also contributes to an $\bar{x} - \bar{c}$ that is too low, and any measurement error for a representative sample that leads to a \bar{c} that is too low also contributes to an $\bar{x} - \bar{c}$ that is too high, thereby, if anything, leading to a decrease in the observed difference between selected and representative samples.

In sum, on average, there is close to zero overconfidence with representative item samples but clear overconfidence with selected item samples. Although the data analyzed here were collected from earlier studies that varied on many dimensions—suggesting caution when we make causal inferences—this analysis does corroborate the predicted difference between selected and representative item samples. This holds even if we control for the (also predicted) overall difference in proportion correct. This selection effect—predicted by the ecological models—is left unaccounted for by other models. In the next section, we turn to a more careful examination of the distribution of individual studies and the hard–easy effect.

The Remains of the Hard–Easy Effect

Figure 2A presents the data points and regression lines that relate the over/underconfidence score to proportion correct for the

selected and representative item samples. The slopes (*bs*) of the regression lines and the coefficients of determination (r^2 s) are presented in the first row of Table 4. The proportions correct for the representative samples clustered around .7 and .8, with almost zero over/underconfidence. With selected samples, most proportions correct were less than .7, and overconfidence was observed. The slope was more steep for the selected item samples ($b = -.72$) than for the representative item samples ($b = -.53$), demonstrating the effect of item-selection procedures.

However, there was a negative slope with both methods of item selection, a hard–easy effect, with proportion correct appearing as an important predictor of the over/underconfidence score ($r^2 = .70$ for selected samples and $r^2 = .58$ for representative samples). For example, as inferred from the regression line for representative item samples, we would predict an over/underconfidence score of approximately .12 for a representative item sample with a proportion correct of .5 and an over/underconfidence score of approximately $-.14$ for a representative item sample with a proportion correct of 1.0 (the corresponding predictions for the selected samples were overconfidence of .20 for a sample with a proportion correct of .5 and underconfidence of $-.17$ for a sample with a proportion correct of 1.0).

We corrected a subset of the data for two of the methodological problems associated with the hard–easy effect: the scale-end effects and the linear dependency. The subset we used was the set of representative item samples collected in our lab, which for obvious reasons was conveniently available to us. First, of course, we wanted to ascertain that there were no important differences between the data based on our representative item samples and the data collected by other labs. In the second row of Table 4, we present the slope and the coefficient of determination for our subset of representative item samples, which are to be compared with the corresponding numbers for the overall set of representative samples presented in the preceding row in Table 4. The differences between our subset and the overall data set were small and not very reliable; if anything, there was a larger hard–easy effect in our data.

Correction for scale-end effects. The combined error model (Juslin et al., 1997, 1999) with one free parameter was used to estimate the response error variance in the use of the overt probability scale. The model was fitted to the calibration curve and distribution of confidence assessments for our subset of represen-

an analysis of covariance, with proportion correct as a covariate. However, because there is a larger hard–easy effect for selected item samples (as predicted by the ecological models; see, e.g., Juslin, 1993b), there is a strong interaction between the over/underconfidence score and the covariate, proportion correct.

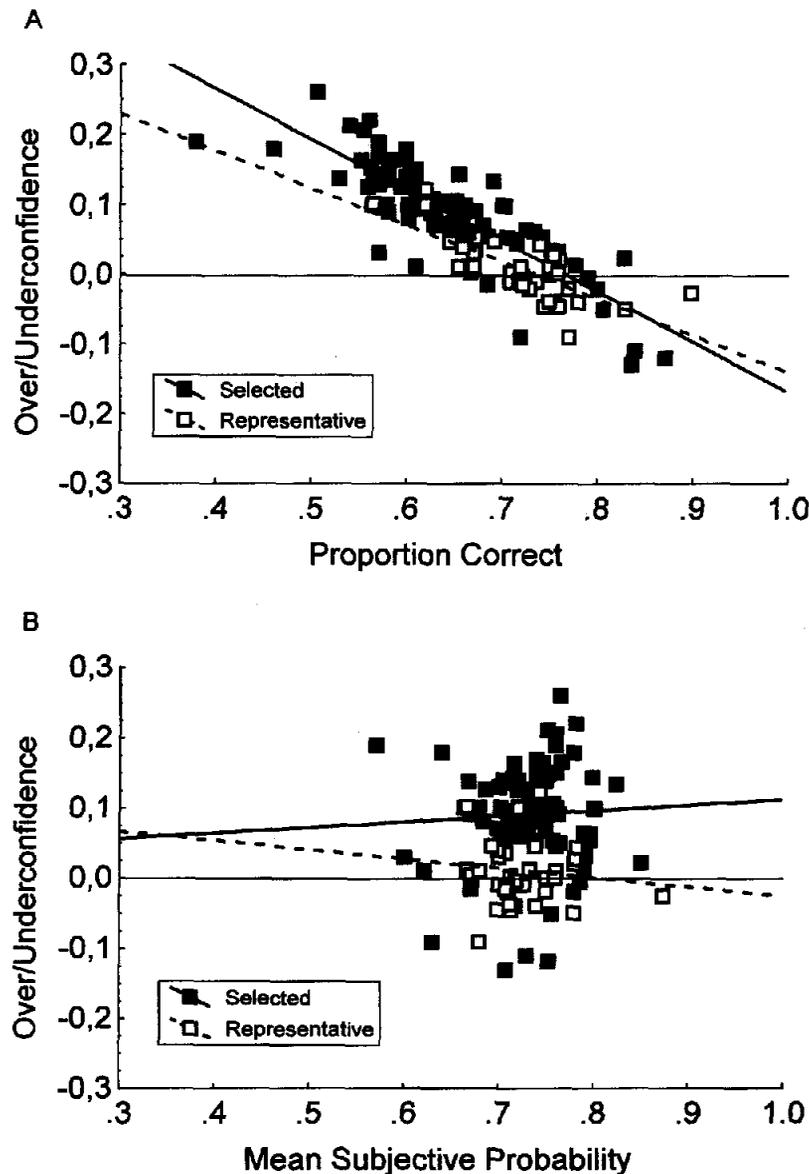


Figure 2. Panel A presents the regression lines that relate over/underconfidence scores to mean proportion correct for selected and representative item samples. Panel B presents the regression lines relating over/underconfidence scores to mean subjective probability for selected and representative samples. The latter slopes are not significantly different from zero ($p > .10$).

tative item samples (see the Appendix for details). The calibration curve is presented in Figure 1B.

Figure 1B suggests that the combined error model with only one free parameter for response error provides a good account for the data (root-mean-square deviation = .03; $r^2 = .99$; $n = 12$; $df = 10$, as computed over both the calibration curve shown in Figure 1B and the response distribution, not shown). This result replicates the previous findings that the combined error model with very small sampling error (actually zero in the present application) and modest response error accounts well for the data (Juslin et al., 1997, 1999). The best fitting value for the response error variance, .02, coincides quite well with the estimates from previous studies (e.g., .02 in Björkman, 1994, and .017 in Juslin et al., 1997).

The model was rerun with the response error variance fixed at the estimated value while we varied the predictability parameter so as to produce different levels of proportion correct (difficulty; see the Appendix and Juslin et al., 1997). This yielded the “hard–easy effect” predicted from the scale-end effects alone. As is evident in Figure 1A, the predicted over/underconfidence score was a negative linear function ($b = -.22$) of proportion correct. Again, to exemplify, the predicted over/underconfidence score was .06 for an item sample with a proportion correct of .5 and $-.06$ for an item sample with a proportion correct of 1.0.

The slope of the regression line for our subset of representative item samples after correction for scale-end effects is presented in Table 4 and Figure 3A. The slope was $-.41$, and the coefficient of

Table 4

Slopes (b) With 95% Confidence Intervals and Coefficients of Determination (r²) for the Regression Lines Relating Over/Underconfidence to Proportion Correct for Selected and Representative Item Samples Collected in the Quantitative Review

| Statistics | Method for item selection | | | |
|---|---------------------------|----------------|----------------------------------|---------------------------|
| | Selected item samples | | Representative item samples | |
| | b | r ² | b | r ² |
| All original data | -.72 ± .10 (n = 95) | .70 (n = 95) | -.53 ± .15 (n = 35) | .58 (n = 35) |
| Our original data | — | — | -.63 ± .31 (n = 17) | .51 (n = 17) |
| Corrected for scale-end effects | — | — | -.41 ± .31 (n = 17) | .31 (n = 17) |
| Corrected for linear dependency | — | — | -.42 ± .27 (n = 34) ^a | .23 (n = 34) ^a |
| Corrected for scale-end effects and linear dependency | — | — | -.20 ± .27 (n = 34) ^a | .06 (n = 34) ^a |

Note. The dashes indicate that the corrections for scale-end effects and linear dependency were applied to the representative item samples collected in our laboratory.

^a n increased from 17 to 34 as we also reversed the roles of definition and measurement sets.

determination was .31. With this corrected regression line, the predicted over/underconfidence score for an item sample with a proportion correct of .5 was .08 (95% CI = ±.08), and the predicted over/underconfidence score for an item sample with a proportion correct of 1.0 was -.12 (95% CI = ±.09). There was still a hard–easy effect in the data, but it was not astonishingly large.

Correction for linear dependency. Before we present the correction for linear dependency, we approach this issue in a somewhat roundabout manner, as illustrated in Figure 2B. Remember that the measurement error is generally much smaller for mean subjective probability than for proportion correct. Of course, there should also be a linear dependency between mean subjective probability and over/underconfidence (mean subjective probability minus proportion correct), but because the error is much smaller for mean subjective probability, this dependency should be weaker. In Figure 2B, we present the over/underconfidence scores regressed on mean subjective probability for the selected and representative item samples (i.e., based on all 95 selected and 35 representative item samples). Overall, the results confirmed our expectations. Neither of the two regression lines had a slope that differed reliably from zero. The data points for the representative item samples were scattered around an over/underconfidence score of zero, regardless of the mean subjective probability. In this sense, the participants seemed to conform to the normative analysis in calibration studies (i.e., across items assigned a subjective probability of .xx, one would expect .xx proportion to be correct). For the selected item samples, the over/underconfidence score was likewise fairly constant, but at the higher level, that corresponds to the “overconfidence phenomenon.”

We will benefit from a simple and clever method recently presented by Klayman et al. (1999) when we correct the data for linear dependency. The raw data for each of the 17 independent data samples are partitioned into one *definition set* and one *measurement set*, with different items and responses in the two sets. The definition set is used to estimate the proportion correct, and the measurement set is used to estimate the over/underconfidence score. Therefore, the same estimate of proportion correct never enters twice in the analysis, both as the independent variable and as part of the dependent variable, over/underconfidence. We also reversed the roles of definition and measurement sets to get twice as many data points.

Figure 3B illustrates the effect of correcting for linear dependency in this way. The slope ($b = -.42$) and the coefficient of determination ($r^2 = .23$) are presented in row 4 of Table 4. We found that the slope was closer to zero and that the proportion correct was no longer a very efficient predictor of over/underconfidence. This result fits nicely with the results reported by Klayman et al. (1999), who applied this procedure across eight target variables (topics) of general knowledge items. Once the linear dependency was corrected for in their study, there no longer was a significant hard–easy effect in the data.

Correction for both response error and linear dependency. Finally, the data in our subset of representative item samples were corrected for both scale-end effects and linear dependency (row 5 in Table 4 and Figure 3C). The negative slope was $-.20$ (ns), and the proportion correct accounted for 6% of the variance in the over/underconfidence score. The predicted overconfidence for item samples with a proportion correct of .5 was .03, and the predicted underconfidence for item samples with a proportion correct of 1.0 was $-.06$. Conditionally on the correctness of our estimates, these are the remains of the hard–easy effect.

We are well aware that, by now, the estimates should be interpreted with caution, because we have stacked the corrections on top of each other. Each correction necessarily involves assumptions, the appropriateness of which may be difficult to ascertain, and the mishaps and errors might accumulate when the corrections are added to each other. But there is a more general message here. We know with close to moral certainty that each of these effects is at work at least to some extent (e.g., response error, measurement error in the proportion correct) and that other problems could surely be added to the list. Notably, both of the factors that we attempted to correct for in this article have by themselves been sufficient to reduce the hard–easy effect to a modest level (in our judgment).

Discussion

In this article, we have presented a theoretical argument with two components. First, the hard–easy effect has been interpreted with insufficient attention to the scale-end effects, the linear dependency, and the regression effects that contribute to the effect. Very few studies control for even one of these problems; the vast majority fail to acknowledge them; and by the time the hard–easy

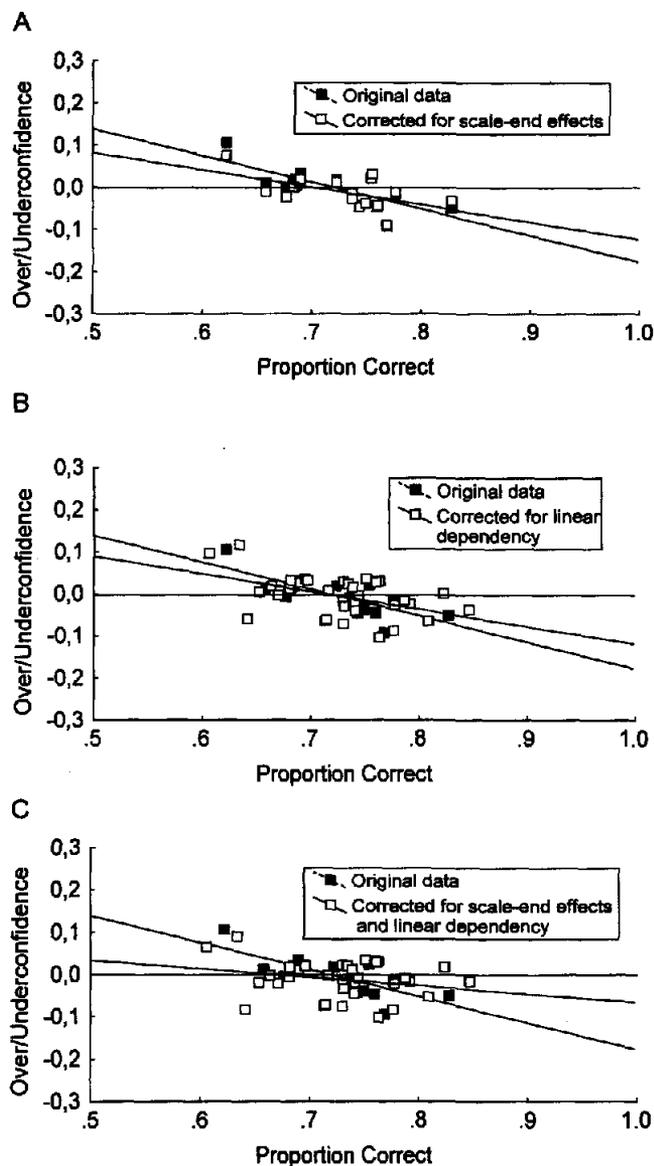


Figure 3. The data points and regression line between proportion correct and the over/underconfidence score for the original data from our laboratory ($n = 17$) are presented, along with the corresponding data points and regression lines after the data points were corrected for scale-end effects (A), linear dependency (B), and both scale-end effects and linear dependency (C; see *The Remains of the Hard-Easy Effect* section of the text for further details).

effect has entered into commentaries and reviews, it has become “a substantive and pervasive” finding (e.g., Keren, 1997, p. 269). The second part of the argument is that this naive empiricism co-occurs with a strong belief in a cognitive overconfidence bias. Therefore, with regard to the general knowledge items reviewed in this article, the overconfidence hypothesis is threatening to become a dogma entrenched by the hard-easy effect and selective attention to particular data sets (i.e., item samples with a low proportion correct).

This quantitative review makes three empirical contributions. First, contrary to the conclusions in comments, reviews, and introductory textbooks (e.g., Myers, 1997; Plous, 1993; Sternberg, 1996), the literature on confidence in general knowledge, primarily based on two-alternative items of the sort reviewed in this article, does not support the idea of a cognitive overconfidence bias that is due to, for example, confirmatory search of memory (Koriat et al., 1980; McKenzie, 1997) or self-serving biases (Taylor & Brown, 1988). The hypothesis is supported when there is a bias that covers most of the range of the proportion correct, in a way that is not better accounted for by scale-end effects, linear dependency, or regression effects. When carefully scrutinizing the data collected with two-alternative general knowledge items—the paradigmatic task in traditional overconfidence studies—we found that, with representative item samples, there was no such support. This, of course, does not preclude that confirmatory search or ego-enhancing biases operate in other task domains.

The second empirical contribution is the demonstration that there is a difference between selected and representative item samples in the direction predicted by the ecological models, with an elimination of the overconfidence phenomenon for representative item samples. This prediction has been challenged by the argument that representative item samples are too easy and by the finding that overconfidence reappears for difficult representative samples, as implied by the hard-easy effect (e.g., Griffin & Tversky, 1992). In this article, we have shown that the hard-easy effect can be explained by several factors other than information-processing biases, and when we corrected for these factors, the residual hard-easy effect was very modest. Finally, even if we were to control for the difficulty of selected and representative item samples by looking only at hard samples, there remains a clear difference between the two selection procedures.

Although only the ecological models account for these results, they do not demonstrate the validity of the specific mechanisms proposed by the ecological models, like the assumption of automatic and effortless encoding of frequencies (see Gigerenzer et al., 1991). Nonetheless, far from being a matter of “ecological hypocrisy” (Keren, 1997, p. 271, footnote 3), we have proposed that the ecological models capture one important aspect of the story. That aspect, the organism-environment relations that support behavior (Brunswick, 1956), is all too often neglected in psychology. Moreover, the ecological models need not be wedded to the naive belief that people are perfectly adapted to their natural environments but can also integrate various imperfections in the cognitive processes (see Juslin et al., 1997, 1999; Soll, 1996). The research with the ecological models and the error models illustrates that cognitive theories need to incorporate explicit theories of the measurements that, for example, specify relations between laboratory tasks and natural environments and the overt manifestations of error in measurement.

Third, the analysis presented here, in which we corrected the data for scale-end effects and linear dependency, suggests that the remaining hard-easy effect is modest (see Klayman et al., 1999, for similar results). If these results are correct, they refute the strength-and-weight model presented by Griffin and Tversky (1992) and remove the main support for the signal-detection-based decision variable partition model (Ferrell & McGoey, 1980). The hard-easy effect might be a robust phenomenon but perhaps for the wrong reasons.

One objection to the conclusions presented here could be that the hypothesis of a cognitive overconfidence bias seems to be supported by results from other tasks and paradigms. First, it should be noted that this objection does not invalidate the results of the quantitative review presented here for general knowledge items. Second, although the analysis of empirical data presented in this article concentrates on calibration of two-alternative general knowledge items, we must emphasize that the arguments apply to a wide range of experimental paradigms in judgment research. Scale-end effects, linear dependency, and regression effects are general phenomena and, indeed, are sometimes more problematic in other paradigms (e.g., for the full-range probability format with probability assessments on a scale between 0 and 1, the scale-end effects are perfectly confounded with overconfidence bias).

With the interval estimation format, participants assess probability intervals around their best guess about the value of some continuous quantity. It was soon realized that the observation of too low proportions of values falling within the stated probability intervals might be traced to a bias in the response process known as anchoring and adjustment (Tversky & Kahneman, 1974). Recent research (Juslin et al., 1999; Klayman et al., 1999) has suggested that there is a profound difference between the interval estimation format and the other assessment formats, with more overconfidence for interval estimation. Juslin et al. (1999) demonstrated that most of this difference can be accounted for in terms of the interaction between stochastic components of judgment and the assessment format, but there was also a modest anchoring-and-adjustment bias that specifically plagued interval estimation.

A third body of data is from the literature on personality and social psychology, where people seem to disclose unrealistic optimism and overly positive self-images (reviewed by Taylor & Brown, 1988). For example, people tend to believe that their chances of experiencing happy life events are larger and that their risks of being victims of sad life events are smaller than those of their average peer (Weinstein, 1980). The typical general knowledge item is likely to create a cognitive set in which the participants concentrate on the factual contents of the item (e.g., "Is New York really located farther north than Rome?"). The paradigms in personality and social psychology commonly direct attention to one's inherent capabilities, one's opportunities in life, or how one is viewed by others (e.g., "What is your intelligence as compared to the intelligence of your average peer?"). One sensible but as yet untested hypothesis is that there are important differences between these more person-oriented tasks and the general knowledge tasks used in calibration research (Griffin & Varey, 1996). Whatever the ultimate interpretation of unrealistic optimism turns out to be, we propose that there is little reason to further bolster the hypothesis of positive biases by pointing to a cognitive overconfidence bias in the processing of general knowledge.

References

References marked with an asterisk indicate studies included in the quantitative review.

- *Allwood, C. M. (1994). Confidence in own and others' knowledge. *Scandinavian Journal of Psychology*, *35*, 198–211.
- *Allwood, C. M., & Björhag, C.-G. (1990). Are two judges better than one?: On the realism of confidence judgements by pairs and individuals.

- In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 443–463). Amsterdam: Elsevier Science.
- *Allwood, C. M., & Björhag, C.-G. (1991). Mood and realism of confidence judgements of one's own answers to general knowledge questions. *Scandinavian Journal of Psychology*, *32*, 358–371.
- *Allwood, C. M., & Granhag, P. A. (1996a). Considering the knowledge you have: Effects on realism in confidence judgments. *European Journal of Cognitive Psychology*, *8*, 235–256.
- *Allwood, C. M., & Granhag, P. A. (1996b). The effects of arguments on realism in confidence judgements. *Acta Psychologica*, *91*, 99–119.
- *Allwood, C. M., & Granhag, P. A. (1996c). Realism of confidence as a function of working in dyads or alone. *Organizational Behavior and Human Decision Processes*, *66*, 277–289.
- *Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes*, *39*, 365–383.
- *Ariely, D., Zauberman, G., & Wallsten, T. S. (1995, August). *Statistical and psychological considerations in calibration research*. Paper presented at the 15th Subjective Probability, Utility, and Decision Making Conference, Jerusalem, Israel.
- *Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*, 133–144.
- Ayton, P., & Wright, G. (1990). Uncertain memories: Evaluating the competence of probabilistic cognition. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 465–476). Amsterdam: Elsevier Science.
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, *55*, 412–428.
- *Baranski, J. V., & Petrusic, W. M. (1995). On the calibration of knowledge and perception. *Canadian Journal of Experimental Psychology*, *49*, 397–407.
- Baranski, J. V., & Petrusic, W. M. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *American Journal of Psychology*, *110*, 543–572.
- *Björkman, M. (1992). Knowledge, calibration, and resolution: A linear model. *Organizational Behavior and Human Decision Processes*, *51*, 1–21.
- Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58*, 386–405.
- *Brake, G. L., Doherty, M. E., & Kleiter, G. D. (1998). *Overconfidence: Rethinking a fundamental bias in judgment yet again*. Manuscript submitted for publication.
- Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*, 212–219.
- Brunswik, E. (1956). *Perception and representative design of psychological experiments*. Berkeley: University of California Press.
- 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*, 172–188.
- *Dawes, R. M. (1980). Confidence in intellectual judgements vs. confidence in perceptual judgments. In E. D. Lanterman & H. Feger (Eds.), *Similarity and choice: Papers in honour of Clyde Coombs* (pp. 327–345). Bern, Switzerland: Huber.
- Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes*, *65*, 201–211.
- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-

- 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 Performance*, 26, 32–53.
- *Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–528.
- *Granhag, P. A., Strömwall, L. A., & Allwood, C. M. (1998). *Confidence judgment processes: Selective but possible to prime*. Unpublished manuscript, Department of Psychology, Göteborg University, Göteborg, Sweden.
- *Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Griffin, D. W., & Vary, C. A. (1996). Towards a consensus on overconfidence. *Organizational Behavior and Human Decision Processes*, 65, 227–231.
- *Hoffrage, U. (1994). *Zur angemessenheit subjektiver sicherheits-urteile: Eine exploration der theorie der probabilistischen mentalen modelle* [On the realism of subjective confidence judgments: An exploration of the theory of probabilistic mental models]. Unpublished doctoral dissertation, Department of Psychology, University of Salzburg, Salzburg, Germany.
- *Jansson, M. (1998). *An empirical test of the PROBEX-model of subjective probability assessment*. Unpublished undergraduate thesis, Department of Psychology, Uppsala University, Uppsala, Sweden.
- *Juslin, P. (1993a). *An ecological model of realism in one's general knowledge* (Acta Universitatis Upsaliensis: Studia Psychologica Upsaliensia, 14). Stockholm: Almqvist & Wiksell.
- *Juslin, P. (1993b). An explanation of the "hard–easy effect" in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, 5, 55–71.
- *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. (1995). Well calibrated confidence judgments for general knowledge items, inferential recognition decisions and social predictions. In J.-P. Caverni, M. Bar-Hillel, F. H. Barron, & H. Jungerman (Eds.), *Contributions to decision making* (Vol. 1, pp. 233–253). Amsterdam: Elsevier Science.
- Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104, 344–366.
- *Juslin, P., Olsson, H., & Björkman, 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., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes*, 73, 3–26.
- *Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1038–1052.
- *Juslin, P., Winman, A., & Persson, T. (1995). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Cognition*, 54, 99–130.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582–591.
- *Keren, G. (1985). *On the calibration of experts and lay people*. Paper presented at the 10th Subjective Probability, Utility, and Decision Making Conference, Helsinki, Finland.
- *Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, 67, 95–119.
- Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.
- Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, 10, 269–278.
- *Klayman, J., Soll, J., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence? It depends on how, what and whom you ask. *Organizational Behavior and Human Decision Processes*, 79, 216–247.
- *Kleitman, S., & Stankov, L. (1996). *Ecological and person-oriented aspects of metacognitive processes in test-taking*. Unpublished manuscript, Department of Psychology, University of Sydney, Sydney, New South Wales, Australia.
- *Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118.
- *Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.
- *Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26, 149–171.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities: The state of the art up to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.
- *May, R. S. (1986). Overconfidence as a result of incomplete and wrong knowledge. In R. W. Scholz (Ed.), *Current issues in West German decision research* (pp. 13–30). Frankfurt, Germany: Lang.
- McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–1993. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.
- McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes*, 71, 141–160.
- Myers, D. G. (1997). *Psychology* (5th ed.). New York: Worth.
- Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes*, 58, 203–213.
- *Pliske, R. M., & Mutter, S. A. (1996). Age differences in the accuracy of confidence judgments. *Experimental Aging Research*, 22, 199–216.
- Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.
- *Price, P. C. (1998). Effects of relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes*, 76, 277–297.
- *Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, 40, 193–218.
- *Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge–advisor decision making. *Organizational Behavior and Human Decision Processes*, 62, 159–174.
- *Sniezek, J. A., Paese, P. W., & Switzer, F. S. C. (1990). The effects of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, 46, 264–282.
- Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65, 117–137.
- Sternberg, R. J. (1996). *Cognitive psychology*. New York: Harcourt Brace.
- *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.
- *Subbotin, V. (1994). *Outcome feedback and confidence in general knowl-*

- edge. Manuscript submitted for publication, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
- *Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes*, 66, 268–276.
- Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47, 143–148.
- Taylor, S. E., & Brown, J. D. (1988). Illusion of well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103, 193–210.
- *Traffimow, D., & Sniezek, J. A. (1994). Perceived expertise and its effect on confidence. *Organizational Behavior and Human Decision Processes*, 57, 290–302.
- Tversky, A., & Kahneman, D. (1974, September). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29, 152–173.
- Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39, 806–820.
- *West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin and Review*, 4, 387–392.
- *Winman, A. (1997a). The importance of item selection in “knew-it-all-along” studies of general knowledge. *Scandinavian Journal of Psychology*, 38, 63–72.
- Winman, A. (1997b). *Knowing if you would have known: A model of the hindsight bias* (Comprehensive summaries of Uppsala dissertations from the faculty of social sciences, 69). Uppsala, Sweden: Acta Universitatis Upsaliensis.
- *Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, 34, 135–148.
- *Winman, A., & Juslin, P. (1994). [Effects of item selection strategies on over/underconfidence]. Unpublished raw data.
- *Winman, A., Juslin, P., & Björkman, M. (1998). The confidence-hindsight mirror effect in judgment: An accuracy-assessment model for the knew-it-all-along phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 415–431.
- *Wolfe, R. W., & Grosch, J. W. (1990). Personality correlates of confidence in one’s decisions. *Journal of Personality*, 58, 515–534.
- *Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, 52, 165–174.
- *Wright, G., & Wishuda, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology*, 23, 219–224.
- Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice Hall.
- Yates, J. F., Lee, J. W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and “reality.” *Organizational Behavior and Human Decision Processes*, 70, 87–94.
- *Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes*, 43, 145–171.
- *Zakay, D. (1992). The influence of computerized feedback on overconfidence in knowledge. *Behaviour and Information Technology*, 11, 329–333.
- *Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, 98, 103–115.

Appendix

Estimating Response Error With the Combined Error Model

The combined error model (Juslin et al., 1997) originally contained four parameters that are fitted to the data: (a) an *inference-proportion parameter* that concerns the relative proportion of responses that are based on probabilistic cue-based inference, as compared with retrieval or deduction; (b) a *predictability parameter* that models the distribution of ecological cue validities that are used in the probabilistic inferences, and thus the difficulty of the item sample; (c) a *sample-size parameter* that represents the amount of experience of the participant (the sample sizes from which the cue validities are estimated by the participants); and (d) a *response error variance parameter* that defines the extent of response error in the use of the overt probability scale. The response error of the combined error model implements the scale-end effects of concern in the present article. At the response elicitation, the true belief about the probability is perturbed by a normally and independently distributed random error, which is truncated at the extremes of the probability scales. All perturbed values less than .5 are assigned an overt probability of .5, and all perturbed values greater than 1 are assigned a probability of 1.0. This response error introduces exactly the kind of regression of the calibration curve exemplified in Figure 1B. The reader is referred to Juslin et al. (1997) for further details of the model.

In the present article, we used a simplified version of the combined error model as an instrument to estimate the response error variance in data. For this reason, we ignored the inference-proportion parameter (arguably, inference or no inference is of minor interest when one is interested in

estimating the response error). We also assumed that there is no sampling error in the participants’ experience and that they have correct estimates of the ecological cue validities. Admittedly, this may seem to be a strong simplification. However, when the combined error model has been fitted to other data sets (e.g., Juslin et al., 1997), the fitted value for sample size has been large, suggesting very little sampling error (e.g., sample size of 60). Moreover, if we were to refit the combined error model to the present data set and use the sample sizes estimated from the previous applications on comparable data sets, this would have an extremely small effect on the estimate of the response error variance. The parameter for predictability (or difficulty) was set so as to produce exactly the observed proportion of correct decisions in our data set (.73). Finally, the response error variance was adjusted to minimize the root-mean-square deviation between the observed calibration curve and distribution across the confidence scale (2×6 data points) and the corresponding predictions. Both the aggregated calibration curve for the entire subset of representative item samples from our lab and the corresponding best fitting predictions by the combined error model are shown in Figure 1B.

Received July 6, 1998

Revision received August 3, 1999

Accepted August 4, 1999 ■