Psychological Review
2000, Vol. 107, No. 4, 943–946

0033-295X/00/$5.00 DOI: 10.1037//0033-295X.107.4.943

THEORETICAL NOTES

Should Observed Overconfidence Be Dismissed as a Statistical Artifact? Critique of Erev, Wallsten, and Budescu (1994)

Lyle Brenner
University of Florida

I. Erev, T. S. Wallsten, and D. V. Budescu (1994) showed that the same probability judgment data can reveal both apparent overconfidence and underconfidence, depending on how the data are analyzed. To explain this seeming paradox, I. Erev et al. proposed a general model of judgment in which overt responses are related to underlying "true judgments" that are perturbed by error. A central conclusion of their work is that observed over- and underconfidence can be split into two components: (a) "true" over- and underconfidence and (b) "artifactual" over- and underconfidence due to error in judgment. It is argued in the present article that decomposing over- and underconfidence into true and artifactual components is inappropriate. The mistake stems from giving primacy to ambiguously defined model constructions (true judgments) over observed data.

A central and extensively researched question in the study of judgment under uncertainty concerns the correspondence between subjective and objective probabilities. In an influential article, Erev, Wallsten, and Budescu (1994) raised several cautions regarding common analyses of subjective probability data. Erev et al. illustrated cases in which the same data can appear to reveal both overconfidence and underconfidence, depending on the method of data analysis. Across a wide range of conditions, Erev et al. demonstrated that when objective probability (OP) is predicted from subjective probability (SP), one may observe overconfidence, but when SP is predicted from OP, one may observe underconfidence. To be precise, the former demonstration of overconfidence consists of the observation that for a given level of SP, average OP is less extreme (i.e., closer to 50%) than SP. The latter demonstration of underconfidence consists of the observation that, for a given level of OP, average SP is less extreme than OP.

Erev et al. (1994) attributed these seemingly paradoxical results to random error in judgment and response processes. They proposed a general model of judgment in which overt statements of subjective probability are related to underlying "true judgments" perturbed by error. In this model, when assessing the likelihood of an uncertain event, a judge experiences a covert degree of confidence (x), modeled as a function of two arguments: a true judgment (t) and a random error component (e):

$$x = f(t, e).$$

This covert degree of confidence is translated into an overt response (y) by some monotonic function g:

$$y = g(x).$$

Erev et al. (1994) described in detail a special case of this general model, in which covert confidence is composed of the log-odds corresponding to the true judgment plus a normally distributed error term:

$$x = \ln\left(\frac{t}{1-t}\right) + e.$$

The transformation g converts from log-odds back to probabilities:

$$y = \frac{1}{1 + \exp(-x)}.$$

For this particular model, Erev et al. (1994) showed that even if the true judgments t are perfectly calibrated, the overt responses y may not be. Consequently, they concluded that patterns of over- or underconfidence in the responses may, in part, reflect statistical artifacts driven by the error term e. In a more recent article, Budescu, Erev, and Wallsten (1997) have argued that this conclusion of (partially) artifactual miscalibration also follows from several other specific cases of the general model described earlier. In several additional articles (e.g., Budescu, Wallsten, & Au, 1997; Wallsten, Budescu, Erev, & Diederich, 1997), the authors have extended the model, while maintaining the basic framework of true judgments perturbed by error. Other authors have made similar arguments regarding the potential artifacts of "error" in judgments of probability (e.g., Juslin, Olsson, & Björkman, 1997; Pfeifer, 1994; Soll, 1996).

Consistent with the notion that observed over- or underconfidence may be partially artifactual, Erev et al. (1994) argued that calibration should ideally be evaluated based on the true judgments (t), after removing the influence of error: "The relation between SP and OP in a particular context needs to be established after con-

Lyle Brenner, Warrington College of Business Administration, University of Florida.

I thank Yuval Rottenstreich, Derek Koehler, Dale Griffin, Ido Erev, Tom Wallsten, David Budescu, Alan Sawyer, and Andrew Ward for helpful comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Lyle Brenner, who is now at Jones Graduate School of Management, Rice University, 6100 Main Street, MS531, Houston, Texas 77005-1892. Electronic mail may be sent to lyle@rice.edu.

trolling for random factors in judgment or response" (p. 523). Thus, Erev et al. argued for a stricter empirical standard for documenting deviations from good calibration. In another article, Budescu, Erev, and Wallsten (1997) stated that "to establish the existence of true over- and underconfidence, one must first be able to discount the possibility that the observed patterns are artifactual and due to the effects of random error" (p. 167). Erev et al. were careful *not* to claim that the common finding of overconfidence *is* artifactual, but rather that in some cases it may be partially artifactual. Indeed, both Erev et al. and others (e.g., Brenner, Koehler, Liberman, & Tversky, 1996; Keren, 1997) have argued that, even if the purported artifacts of random error are removed, overconfidence can still be reliably documented.

My goal in this article is to examine the logic and foundations of Erev et al.'s (1994) model of true judgments perturbed by error. I argue that, although their model may do an excellent job of accounting for subjective probability data, it is inappropriate to separate observed over- and underconfidence into true and artifactual components based on such a model. As a result, even the claim that observed over- and underconfidence *may* be a statistical artifact needs to be reconsidered. Although well-intentioned, the attempt to discriminate between true and artifactual miscalibration suffers from a basic flaw: the lack of a clear definition of either the true judgment t or the error term e. Despite the ambiguous definition of t, Erev et al. nonetheless treated the constructed true judgment as more fundamental than the data to be explained. In the absence of compelling definitions and interpretations for the concepts of *true judgment* and *error* there is no sound basis for separating so-called artifactual miscalibration from true miscalibration.

Erev et al.'s (1994) model makes a valuable contribution by reminding researchers that there are many different (and potentially useful) ways to analyze subjective probability data and that seemingly similar analyses can often yield divergent conclusions. Furthermore, Erev et al. encouraged modelers to explicitly recognize variability in judgment, a recommendation I wholeheartedly endorse. However, Erev et al. overreached by implicitly claiming to identify the fundamental criterion for evaluating "true" calibration.

Example of the Erev, Wallsten, and Budescu (1994) Model

Let us first consider a simple illustration of how Erev et al.'s (1994) log-odds model can lead to an interpretation of spurious over- or underconfidence. Assume that a judge's true judgment (as defined in Erev et al.'s model) for a particular class of events is $t = 0.90$, and that this true judgment is perfectly calibrated (i.e., the appropriate objective probability for this class of events is also 0.90). This true judgment corresponds to a log-odds of $\ln(0.90/0.10) = 2.2$. For simplicity, assume that the error term e consists of either a two-point log-odds increase or a 2-point log-odds decrease with equal likelihood. The equally likely covert confidence levels (in log-odds) would then be $x = 0.2$ and $x = 4.2$, and the corresponding overt responses would be $y = .55$ and $y = .99$. Thus, for events with an objective probability of .90, the average overt subjective probability is only $(.55 + .99)/2 = .77$, yielding substantial underconfidence in terms of average probabilities, but perfect calibration in terms of average log-odds.

If one concludes that the observed underconfidence in stated probabilities in this case is merely an artifact, misrepresenting the true psychological state of confidence, one is implicitly assuming that log-odds are the appropriate units for evaluating true calibration. But why use log-odds? No compelling rationale was provided by Erev et al. (1994) for the log-odds model, beyond the suggestion that covert confidence should be represented on an unbounded $(-\infty, \infty)$ rather than bounded [0, 1] scale. Indeed, they noted that the log-odds transformation is just one of many possible forms of the general model $x = f(t, e)$. But in order to make claims about the calibration of the true judgments, the researcher specifies the function f, and by doing so implicitly defines the fundamental units in which calibration is to be evaluated. Given that the specification of f (and consequently t) is left to the judgment of the researcher, it does not seem prudent to evaluate the calibration of primary interest in terms of the true judgments t.

There is a natural alternative way to interpret the data in the example discussed here without invoking the "true judgment" and "error" interpretations of Erev et al.'s (1994) model. We observe that for events with OP = .90, judges respond with overt subjective probabilities of .99 and .55, sometimes overestimating the target value slightly and other times underestimating it substantially. On average, then, in conditioning on OP = 0.90, we observe underconfidence in judged probabilities. If, for whatever reason, we wished to evaluate the calibration of the judge's responses converted to log-odds, they could easily do so and would find that for an event with log-odds of 2.2, judges respond with overt subjective probabilities corresponding to log-odds of 4.2 and 0.2. On average, in conditioning on objective log-odds of 2.2, we observe perfect calibration of judged log-odds.

Of course, many other analyses are available to the researcher: Probabilities could be transformed to odds, or to z-scores using the inverse cumulative normal distribution, or to any other units of interest. To determine "true" calibration, Erev et al. (1994) needed to identify the units that the judge "thinks in" and evaluate calibration in those units. However, this is not a goal that any psychological model can ever achieve because the "true units" of thought do not exist outside the assumptions embedded in the model. Suggesting that the judge may be well-calibrated internally, where the internal confidence measure is some transformation of the stated probability, presupposes the very issue to be determined.

To illustrate further the arbitrary nature of the true judgments, note that if one observed well-calibrated responses in terms of average probabilities, it would be straightforward to construct a model in which people are "truly" miscalibrated and the observed good calibration is merely due to the transformation from psychological units to response units (and an appropriately chosen error distribution). Depending on the specific transformation assumed, one could conclude either true overconfidence or true underconfidence. The central problem lies in declaring the constructed true judgments to be more fundamental than the judgment data from which they are derived.

Having said this, I should stress that I do not in any way wish to discourage model building in the study of probability judgment. Much of my own work has been directed toward developing and extending models of probability judgment (e.g., see Brenner & Koehler, 1999), using the framework of support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). The problem with Erev et al.'s (1994) conclusion does not reside in the math-

ematical structure of the model they proposed, but rather in the (overly) fundamental interpretations they gave to the terms in the model. Although their model may be extremely useful in accounting for probability judgment data, such utility does not justify claims of identifying true or underlying calibration.

Defining True Scores and Errors

In addition, I would suggest that the terminology (*true judgment* and *random error*) used in describing the model may contribute to the problem of treating the model as more fundamental than the data. Once the mathematical entity t is labeled as a "true judgment" and "error" (as represented by e) is introduced into the process leading to an overt response (y), it seems natural to evaluate the calibration of t rather than y. The language stacks the deck in favor of the model (which purports to identify the "underlying" process) over the error-contaminated data.

At an even more basic level, neither t nor e is unambiguously defined. This problem is illustrated by Erev et al.'s (1994) circular definition of the true judgment in terms of error: "The subject's true judgment of the likelihood of event i, t_i, is the estimate from 0 to 1 the subject would provide if he or she could operate in a fully repeatable, *error-free* manner" (p. 524, emphasis added). In addition to being circular, this definition is puzzling in that it rests on a condition (that the judge could operate in a "fully repeatable" manner) that is always assumed to be false.

Consider an alternative labeling for the mathematical terms in this model. In most classic psychometric contexts, a *true score* is defined as a mean across some population of items. For example, a test taker's true score for an aptitude test is the mean of the scores she or he would receive across a set of equivalent tests (e.g., tests with questions sampled from some population of items). *Error* is interpreted as variability around this mean. In these terms, then, t in Erev et al.'s (1994) model can be defined as the average log-odds across a (hypothetical) population of judgments of the same event; e can be defined as the difference between the log-odds of any particular judgment and the population mean t. The key insight from using these definitions is that t represents a *particular population mean* (which requires some specification of the population) defined over a *particular empirical operation* (in this case, a judge making a probability judgment that is then transformed to log-odds). In the absence of auxiliary arguments about the importance of the particular population and empirical operation invoked in defining t, strong conclusions about the fundamental nature of the true judgments are not warranted.

These changes in interpretation are not trivial, nor are they merely semantic. In many cases the connotations of the terms used in one's model may be critically important and may imply quite different analyses and conclusions. Once the terms *true judgment* and *error* are used for t and e, it is a natural next step to remove the "nuisance" influence of error and to evaluate the calibration of t. In contrast, using definitions in which t is interpreted as a population mean and e as within-subject variability, it becomes less clear why one should evaluate the calibration of the mean log-odds. Of course, such an analysis may be of interest, *not* because it identifies true or fundamental underlying processes, but rather because it may reveal structure in the data that was not apparent from analysis of the raw, untransformed data.

The implications of within-subject variability of judgment are of substantial theoretical and practical interest; we would be wise to follow Erev et al.'s (1994) lead and build models that explicitly incorporate variability of judgment. However, we must be especially careful that our models, and the terms we use to describe those models, do not inadvertently distract us from the central goal of describing and characterizing the systematic regularities in the data. Once the leading terms *true judgment* and *error* are removed, the conclusion flowing from Erev et al.'s model is that a judge can be well-calibrated in terms of one metric (e.g., average log-odds of probability judgments) but nevertheless show over- or underconfidence when evaluated in terms of a different metric (e.g., average probability judgments). This discrepancy will occur whenever there is variability in the judgments and the transformation from one metric to another is nonlinear. Although based on the same model, this conclusion is quite different from Erev et al.'s stated conclusion that some over- or underconfidence may be an artifact of error in judgment and response processes.

Summary and Conclusion

I have argued that it is inappropriate to treat constructed true judgments as more fundamental than the data from which they are derived, and therefore that over- or underconfidence in observed responses should not be dismissed based on models of well-calibrated true judgments perturbed by error. Furthermore, use of the terms *true judgment* and *error* may contribute to these problems.

The oft-asked question "Are people really overconfident?" (cf. Ayton & McClelland, 1997) is best answered, like many other questions, by "It depends." Empirically, calibration quality, however defined, depends on the tasks presented and the particular judges tested. Furthermore, the answer might be complex because different data analyses may yield somewhat different conclusions. This is not a limitation of the data analysis so much as a direct consequence of asking an overly general question. We can best understand human behavior by describing the behavioral data we gather in its full richness and complexity, and then extracting the simple laws or regularities in the data. A specific interpretation of an extracted regularity ought not take precedence over the regularity itself. If, in some task, people provide confidence ratings of 90% but are correct only 75% of the time, we can reasonably say that their statements of 90% confidence are, on the whole, overconfident. Whether this empirical result is seen as a consequence of variability in judgment, or a transformation from psychological units to verbal units, does not change the empirical pattern.

The constructed true judgments would be more appropriately named if it were shown that they predicted a measure that is seen as a more important manifestation of subjective probability. For example, perhaps the true judgments t (defined in log-odds) are more closely related to a measure of subjective probability derived from choices among actions, rather than a direct numerical probability statement. If we define the choice-derived measure as the "true subjective probability of interest," then it might be reasonable under appropriate conditions to say that calibration of stated probabilities is partially biased or artifactual, relative to the choice-derived measure. Such claims, of course, require that we operationally define the measure of fundamental interest. A definition in terms of log-odds (or any other transformation) without justifica-

tion to support that definition as fundamental is not a firm foundation on which to make strong claims about "true" calibration.

The famous statement that "all models are wrong, but some models are useful" (usually attributed to statistician George Box) encourages both the use of mathematical and statistical models, as well as caution in the interpretation of the entities in the models. The models proposed by Erev et al. (1994) indeed do an excellent job of clarifying important structure in judgment data, as do many other models that invoke stochastic latent variables. It is essential, though, to maintain appropriate interpretations of the entities created in the models. In the case of Erev et al.'s model, the constructed true judgments should not trump the observed responses as the appropriate measure to be evaluated for judgment quality or coherence.

References

Ayton, P., & McClelland, A. G. R. (1997). How real is overconfidence? *Journal of Behavioral Decision Making, 10,* 279–286.

Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: Local-weight models for decomposition of evidential support. *Cognitive Psychology, 38,* 16–47.

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.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making, 10,* 157–171.

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.

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.

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,* 189–210.

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making, 10,* 269–278.

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

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104,* 406–415.

Soll, J. (1996). Determinants of miscalibration and over/underconfidence: The roles of random noise and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117–137.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243–268.

Received May 6, 1998
Revision received December 22, 1999
Accepted December 30, 1999 ■