Journal of Behavioral Decision Making, Vol. 10, 157–171 (1997)

On the Importance of Random Error in the Study of Probability Judgment. Part I: New Theoretical Developments

DAVID V. BUDESCU,[1]* IDO EREV[2] AND THOMAS S. WALLSTEN[3]
[1]*The University of Illinois, USA*
[2]*The Technion, Israel Institute of Technology, Israel*
[3]*The University of North Carolina at Chapel Hill, USA*

ABSTRACT

Erev, Wallsten, and Budescu (1994) demonstrated that over- and underconfidence can be observed simultaneously in judgment studies, as a function of the method used to analyze the data. They proposed a general model to account for this apparent paradox, which assumes that overt responses represent true judgments perturbed by random error. To illustrate that the model reproduces the pattern of results, they assumed perfectly calibrated true opinions and a particular form (log-odds plus normally distributed error) of the model to simulate data from the full-range paradigm. In this paper we generalize these results by showing that they can be obtained with other instantiations of the same general model (using the binomial error distribution), and that they apply to the half-range paradigm as well. These results illustrate the robustness and generality of the model. They emphasize the need for new methodological approaches to determine whether observed patterns of over- or underconfidence represent real effects or are primarily statistical artifacts. © 1997 by John Wiley & Sons, Ltd.

Journal of Behavioral Decision Making, 10, 157–171 (1997)

No. of Figures: 3 No. of Tables: 3 No. of References: 39

KEY WORDS overconfidence; calibration; bias

In an earlier paper, we (Erev, Wallsten, and Budescu, 1994) analyzed an apparent paradox in the empirical judgment literature. Numerous studies of revision of opinion (reviewed, for example, by Edwards 1968; Fischhoff and Beyth-Marom, 1983; Rapoport and Wallsten, 1972; Slovic and Lichtenstein, 1971) concluded that subjects' estimates of probability following the observation of data are not sufficiently extreme relative to the predictions of Bayes' theorem. We refer to this pattern, which was called conservatism in the earlier literature, as *underconfidence*. In a different paradigm, when subjects are asked to estimate the probabilities that statements, answers to questions, or forecasts are correct, their assessments generally are more extreme than the corresponding relative frequencies

* Correspondence to: David V. Budescu, Department of Psychology, The University of Illinois, Champaign, IL 61820, USA. Tel: 217 333 6758; Fax: 217 244 5876; Email: dbudescu@s.psych.uiuc.edu

CCC 0894–3257/97/030157–15$17.50

Accepted 8 January 1997

correct. This pattern suggests that subjects are *overconfident* in their judgments. (For reviews of this literature on *calibration*, see Keren, 1991; Lichtenstein, Fischhoff, and Phillips, 1977, 1982; McClelland and Bolger, 1994; Wallsten and Budescu, 1983; Yates, 1990.)

To determine whether the conflicting pattern of results are due to fundamental differences in how subjects handle internal and external sources of uncertainty (Budescu and Wallsten, 1987; Howell, 1972; Howell and Burnett 1978; Kahneman and Tversky, 1982; Teigen, 1994) or to the methods of analysis (i.e. analyzing objective probabilities (OP) as a function of subjective probabilities (SP), or vice versa), Erev *et al.* (1994) applied both types of analyses to three sets of judgments involving internal and/or external uncertainty in paradigms in which the two definitions of OP can, in principle, coincide. In all cases they found simultaneous over- and underconfidence as a function of the analysis used (also see Wagenaar and Keren, 1985). These results suggest that the dual patterns do not necessarily reflect different information processing modes. They are, however, consistent with a general pattern of reversion to the mean (Samuels, 1991), which generalizes the better-known (linear) regression to the mean. Thus, it is possible that the phenomena of over- and underconfidence are, at least to some degree, statistical consequences of how the data have been analyzed. To further illuminate this possibility, Erev *et al.* (1994) proposed and illustrated a class of stochastic models that explicitly incorporate a random error component and reproduce the empirical regularities observed in these data sets.

These judgment models assume four constructs, each represented by a random variable: true judgment, T, error, E, covert confidence, X, and response, R. The values of T, t_i, denote the subject's true judgment of the likelihood of event i. Thus the distribution of T depends on the items in the set being judged. We assume only that $0 \leqslant t_i \leqslant 1$. The values of E, e_j, denote a random error component on trial j. When considering the likelihood of event i at time j, the subject experiences a degree of confidence, $x_{ij} \in X$, which depends on t_i and e_j. That is,

$$X = f(T, E) \tag{1}$$

A response rule translates the covert feeling x_{ij}, into an overt response, $r_{ij} \in R$ in the [0,1] interval. That is,

$$R = g(X) \tag{2}$$

First, we need to elaborate on the appropriate interpretation of *true judgment* and *error* in this context. Our discussion relies heavily on the views articulated by Lord and Novick (1968) in the context of the theory of mental tests, and applied to analysis of probability judgments by Wallsten and Budescu (1983). It would be unnatural, and totally unrealistic, to think of the *true subjective probability* of an event (e.g. 'the Dow Jones industrial average will increase by more than 50 points next week' or 'rain tomorrow') in the same Platonic sense as one thinks of *true* physical measures, such as the *true* distance between two points or the *true* mass of an object. Rather, the concept of a *true subjective probability* is based on the idealized assumption that a given respondent can repeatedly and independently form judgments regarding a class of equivalent events I. The *true* value is the judgment, t_i, the subject would have if he or she could operate in a fully repeatable and error-free manner on all these occasions. For example, a weather forecaster operating during a certain season in a given location forms numerous probability judgments regarding the likelihood of precipitation under a set of atmospheric and meteorological conditions which, for all practical purposes, can be considered equivalent. However, the forecasts provided on these various occasions will not necessarily be identical. The variation may be due to subtle differences in how the forecaster perceives, codes, interprets, or aggregates the information; to random fluctuations in his or her mood or physical state; or to all the above factors. We refer to the combined effect of all these momentary and unpredictable factors as

error. We assume, quite naturally, that these errors cancel out at the judgment level (i.e. have a mean of zero) and refer to the judgment variability induced by these factors as *error variance*. Under mild technical assumptions, it is possible to show that the average judgment will converge to a constant value. It is this hypothetical constant that we refer to as the *true* judgment. This description is fully consistent with Thurstone's classical model for the laws of comparative and categorical judgments (Thurstone, 1927a,b). Similar models of true score + random error were invoked in many other instances in the judgment and behavioral decision literature (e.g. Schoemaker and Hershey, 1992, for the measurement of utility; Ravinder, Kleinmuntz, and Dyer, 1988, in the analysis of decomposed subjective probabilities; and Ravinder and Kleinmuntz, 1991 for a variety of measurement problems in additive multiattribute utility models).

This formulation applies quite naturally to repeatable events in which the uncertainty stems primarily from external factors. Typical examples are meteorological or financial forecasts provided under highly similar circumstances. The same formulation can, however, be applied to almanac or general-knowledge items in which uncertainty is primarily associated with internal sources. It is well established that when respondents are not fully certain with regard to the answer *of a specific item* their estimates may vary across replications. This explains why test–retest reliabilities of objective aptitude and intelligence tests are less than perfect, and also the high attractiveness of probabilistic test forms as a means of expressing partial knowledge (e.g. Ben Simon, Budescu, and Nevo, in press). In a more general sense, for each respondent, one can consider *items of equal difficulty from a well-defined domain* to be repeatable events for our purposes (e.g. Was the 1995 population of [Nashville, Cincinnati, New Orleans, Columbus] greater than one million? Did [Frankfurt, Bremen, Bielefeld, Regensburg] have soccer teams in the Bundesliga in the 1995–6 season of play?).

Equations (1) and (2) constitute a very broad class of models. Various special cases can be derived by making assumptions about the nature of f, g, and the error distribution. We (Erev *et al.*, 1994) used one instantiation of this class of models to demonstrate the hypothesized pattern of results. For purposes of the illustration, we assumed a fifth random variable, P, with $p_i \in P$ representing the objective probability of event I, and, furthermore, that true judgment is accurate, i.e. $T = P$. To simulate a wide range of subject differences and experimental conditions, we derived predictions from the particular instantiation for various joint probability distributions over the true probabilities, t_i, and response categories, r_k, $(i, k = 0, \ldots, 10)$, which we then summarized in either of two ways depending on which analysis we were mimicking (see Murphy and Winkler, 1992). We examined all combinations of three distributions of true probabilities and four levels of error variability, σ, across eleven response intervals. The three distributions were uniform, U-shaped, and W-shaped. The four values of σ employed were 0.5, 1.0, 1.5, and 2.0. The paradoxical double effect was obtained in all cases. Its magnitude depended on the error variance and, to a lesser extent, on the shape of the distribution of true probabilities. We concluded in Erev *et al.* (1994) that there is good evidence that observed over- and underconfidence in judgment tasks is due, at least in part but not necessarily entirely, to the statistical methodology employed in the depiction and analysis of the results.

The voluminous literature on the inaccuracy and miscalibration of human judgments has, relatively, little to say about the actual psychological processes used by the judges, and no universally accepted model dominates this literature (see McClelland and Bolger's review from 1994). Given this state of affairs, general theoretical claims regarding the relevance, or importance, of certain factors must be shown to hold under a wide variety of models and circumstances. In this paper we seek to establish the generality and robustness of the Erev *et al.* (1994) results by showing that they are replicated under alternative assumptions about the nature and magnitude of the error components, and using other response scales. These illustrations re-emphasize the importance of considering and explicitly modelling the relation between judgment and response and the concomitant effects of random error in judgment studies. In the companion paper (Budescu, Wallsten, and Au, 1997) we propose and illustrate

an alternative method of analysis that allows one to determine whether there is evidence of systematic over- or underconfidence *beyond the level expected given the random error in the data*.

ALTERNATIVE MODELS OF ERROR

The most general form of the Erev *et al*. (1994) model relies on very weak and reasonable assumptions. However, more specific and, possibly, restrictive assumptions are required to implement the model and derive of quantitative predictions. Erev *et al*. (1994) assumed a log-odds relationship between a 'full range' (0–1.0) response scale and the underlying confidence. Moreover, we assumed for purposes of illustration that judges are perfectly calibrated in their true judgments (see details below). In this section we will present three variations on this theme. The first model uses a binomial distribution of error and a different response rule. Our second and third models apply to judgment tasks using the 'half range' (0.5–1.0) scale, and invoke the log-odds and the binomial models respectively.

The full-range log-odds model

Erev *et al*. (1994) assumed that the degree of confidence experienced by an individual when considering the likelihood of event i is not limited to a bounded closed interval, rather $-\infty < x_{ij} < \infty$. More specifically, the degree of confidence was represented as an additive combination of error plus a log-odds transformation of t_i. That is,

$$x_{ij} = \ln\left(\frac{t_i}{1 - t_i}\right) + e_j \tag{3}$$

$0 < t_i < 1$. Errors were assumed to be independently, identically, and normally distributed with $\mathrm{EV}(E) = 0$, and $\mathrm{Var}(E) = \sigma^2$. To specify the response rule, Erev *et al*. postulated that X is mapped via a continuous strictly increasing response function into the [0,1] interval:

$$Y_{ij} = \frac{\mathrm{e}^{x_{ij}}}{1 + \mathrm{e}^{x_{ij}}} \tag{4}$$

Finally, to obtain a discrete response scale consistent with the common observation that subjects tend to respond in units that are multiples of 0.05 or 0.10 (e.g. Wallsten, Budescu, and Zwick, 1993), Erev *et al*. assumed that y_{ij} is mapped into response categories $r_0, r_1, \ldots, r_n$ by means of response cut-offs, $\pi_k, k = 1, \ldots, n$, such that $0 < \pi_1 < \ldots < \pi_n < 1$. The response rule is to respond:

$$\begin{aligned} r_0 \quad &\text{iff} \quad Y_{ij} \leqslant \Pi_1 \\ r_{\mathrm{i}} \quad &\text{iff} \quad \Pi_i < Y_{ij} \leqslant \Pi_{i+1} \qquad i = 1, \ldots, n-1 \\ r_{\mathrm{n}} \quad &\text{iff} \quad \Pi_n < Y_{ij} \end{aligned} \tag{5}$$

A full range binomial model

Our first model applies to the same response format, i.e. full-scale using the [0,1] range, but assumes a different internal judgment process that is captured by another model. Whereas Erev *et al*. (1994) assumed that one's internal degree of confidence is mapped onto a continuous and unbounded scale (which is translated to a bounded probability scale only at the response stage), the current version constrains the internal representation of confidence to be in the [0,1] range as well. The model implicitly assumes that one constructs confidence in the truth of an event by successively sampling and

accumulating evidence. The degree of confidence is quantified as the proportion of confirming items. The level of random error is modelled by the number of items considered. Specifically, the smaller this number, the higher the error variance. The level of error can be affected by external factors (e.g. time pressure that prevents one from considering all relevant information), or internal ones (e.g. one's inability to come up with more pieces of relevant information regarding 'hard items'). The model to this point is essentially identical to the one developed and used by Soll (1996), as well as by Juslin, Olsson, and Björkman (1997) and can be thought of as a sampling version of Björkman's (1994) internal cue model. Juslin *et al.* (1997) have argued, convincingly, that in a process of judging probabilities, random error can play an important role at the judgment and/or response stage. Therefore, unlike our log-odds or Soll's (1996) model that invoke deterministic response rules, we propose a probabilistic response mechanism, so random error is present at the response stage as well.

To present these ideas formally, we assume that the subject's judgment,

$$x_{ij} = \frac{b_{ij}}{m} \tag{6}$$

arises according to a binomial distribution with parameters t_i and m (a positive integer inversely related to degree of error in judgment). b_{ij} is an integer from 0 to m reflecting momentary strength of confidence on trial j given underlying probability t_i. To interpret b_{ij}, one might assume that in a calibration context the respondent considers each item m times, or comes up with m pieces of information about that item, of which b_{ij} suggest the statement is true and the remaining suggest it is false. When the environmental probabilities correspond to relative frequencies, the subject might imagine m trials, on b_{ij} of which the event occurs.

As with the log-odds model, we assume that the subject partitions the [0,1] response interval into $n + 1$ categories by thresholds denoted $\pi_i, 0 < \pi_1 < \ldots < \pi_n < 1$, and selects the centers of those categories, $r_0, \ldots, r_n$, as the response set. We also assume that the values of the response set are equivalent to the set of environmental probabilities, $p_i, i = 1, \ldots, n + 1$, and that the subject's true judgment is accurate, i.e. $t_i = p_i$.

A response rule analogous to equation (5) will not work now, because some values of b/m may span two categories and values of $m < 10$ will exclude some categories altogether. Instead, we adopt a probabilistic 'matching' model. Realizing that b can be generated by several of the possible t_i categories, we assume the respondent selects one of them by a random process in which the probability of selecting each response category is proportional to the posterior probability (conditional on b and m) of the corresponding judgment category:

$$\Pr(r_k \mid b, m) \propto \Pr(t_k \mid b, m)$$

But because the constant of proportionality $\sum_{i=1}^{n+1} \Pr(t_i \mid b, m)$ sums to 1, this reduces to:

$$\Pr(r_k \mid b, m) = \Pr(t_k \mid b, m) \tag{7}$$

We applied the model to 12 cases using 11 response categories with $r_k, t_i = 0.025, 0.10, 0.20, \ldots, 0.90, 0.975$ and the same three prior distributions (uniform, W, and U-shaped) over p_i, and therefore over t_i. The four levels of error variance were obtained by using $m = 5,\ 10,\ 15,$ and 20. Exhibits 1 and 2 summarize the results. Calculation of $S(p_i) = \mathrm{EV}(R|p_i)$ for all i corresponds to analyses when OP is defined. In words, $S(p_i)$ is the mean judgment associated with events that have true probability p_i when OP is defined, as in revision of opinion tasks. Calculation of $O(r_k) = \mathrm{EV}(P|r_k)$ for all k corresponds to analyses when OP is not defined. $O(r_k)$ is the mean objective probability associated with all items within each response class (actually, expected relative frequency correct for that response class), as typically calculated in calibration experiments.

Exhibit 1. Mean relative confidence measures, $CONF_o$ and $CONF_s$, for twelve illustrative examples of the Binomial model with eleven response categories

Prior distribution	m	$-CONF_o = CONF_s$	O/U
U-shaped	20	0.008	0.008
	15	0.010	0.010
	10	0.015	0.014
	5	0.029	0.028
Uniform	20	0.022	0.020
	15	0.028	0.025
	10	0.040	0.036
	5	0.071	0.065
W-shaped	20	0.027	0.019
	15	0.034	0.024
	10	0.047	0.033
	5	0.080	0.056

To obtain the entries in Exhibit 1, we calculated for each case the average signed deviation, defined such that for both measures, positive values indicate overconfidence and negative values imply underconfidence. The indices, $CONF_s$ and $CONF_o$, are the average (over/under)confidence when the analysis is conditional upon SP and OP, respectively. Formally, the measures are given by:

$$CONF_o = \frac{1}{N}\left[\sum_{p_i<0.5} N_i(p_i - S(p_i)) + \sum_{p_i>0.5} N_i(S(p_i) - p_i)\right] \tag{8a}$$

and

$$CONF_s = \frac{1}{N}\left[\sum_{r_k<0.5} N_k(O(r_k) - r_k) + \sum_{r_k>0.5} N_k(r_k - O(r_k))\right] \tag{8b}$$

where N_i/N is the proportion of observations in the corresponding probability or response category, respectively.[1]

Unlike other measures of over-or underconfidence (e.g. Juslin, 1994; Ronis and Yates, 1987), ours excludes the central category. We do so because the concepts of over- and underconfidence are not well defined for SP or OP = 0.5 in a full-scale response paradigm, and our model implies perfect accuracy (i.e. $S(P_i) = p_i$ and $O(r_k) = r_k$) in this interval. Others have handled this problem by treating the contribution of SP = 0.5 responses to the over/underconfidence score (O/U in their terms) as 0.[2] In fact, because the binomial and log-odds models guarantee that the calibration curves (Exhibit 2) go through (0.5, 0.5), our index relates to O/U according to $O/U = (1 - q)CONF_s$, where q is the proportion of responses in 0.5 category. The O/U values are given in the last columns of Exhibit 1.

[1] Erev *et al.* failed to include the N_i or N_k in their formulae. That was an error in writing the equations. All calculations with those equations were done correctly.

[2] In fact, other authors first convert full-scale to half-scale responses by assuming that any event eliciting a response greater than 0.5 would be called *true* and assigned the stated probability, SP, of being correct; an event eliciting a response less than 0.5 would be called *false* and assigned the probability of 1-SP of being correct; events eliciting SP = 0.5 would be called *true* or *false* each half the time and be assigned 0.5 of being correct. The effect of this rule is to guarantee that full-scale responses of 0.5 contribute 0 to the O/U score.

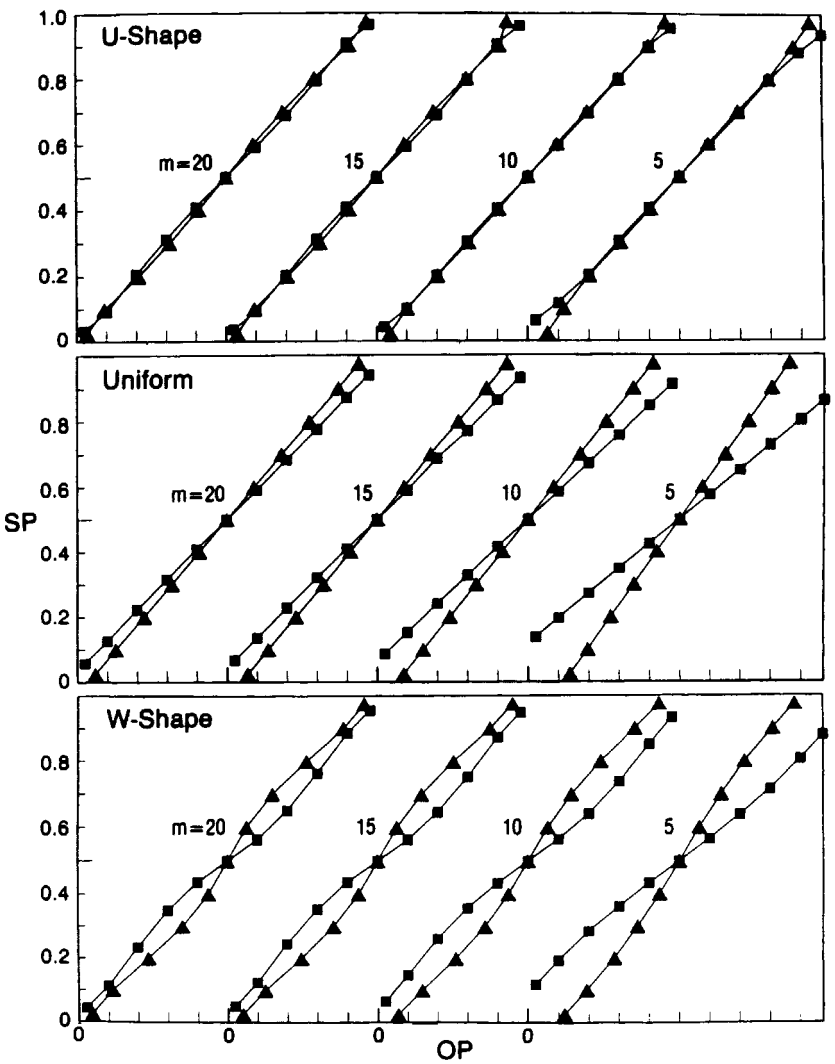

Exhibit 2. Mean objective probability (OP) as a function of subjective probability (SP) (The S-curve, denoted by ▲) and mean SP as a function of OP (the O-curve, denoted by ■) for twelve instances of the binomial model defined by three prior distributions (U-shaped, Uniform, W-shaped) and four sample sizes (m = 20, 15, 10, 5). The abscissa increments in units of 0.1 beginning at 0, as indicated for each pair of curves.

The results, shown in Exhibits 1 and 2, are very similar to those from the log-odds model. The plots in Exhibit 2 show that the double effect is obtained. This plot (and some of the others in the sequel) is oriented in a manner which is common in studies where OP is defined and manipulated (i.e. OP on the abscissa and SP on the ordinate). Indeed, the O-curve shows the typical pattern of underconfidence, i.e. SP > OP when OP < 0.5, SP < OP when OP > 0.5, and SP = OP at 0.5. The orientation of the S-curve is opposite to the one usually used in the calibration research but after making the appropriate adjustment (e.g. by rotating the exhibit so that SP is on the abscissa), the standard pattern of overconfidence for the S-curve emerges (i.e. OP > SP when SP < 0.5, OP < SP when SP > 0.5, and SP = OP at 0.5). The numerical values in Exhibit 1 illustrate that the magnitude of the effects depends

on the amount of response error. The effect is weakest in the U-shaped distribution.[3] (In fact, it does not appear in the case where $m = 15$ or 20 for $p_i = 0.1$ and 0.9.) Because of the matching nature of the response rule used in this model, $\mathrm{CONF_s} = -\mathrm{CONF_o}$ by definition.

Half-range models

So far we have assumed that the subject responds in the [0,1] range. In fact, many experiments employ a half-range procedure in which the subject first identifies the most likely of two alternatives and then states its probability in the [0.5, 1] range. This format is especially popular in calibration experiments (e.g. Lichtenstein *et al.*, 1982), where subjects first select the more likely of two possible answers (e.g. 'Is absinthe a precious stone or a liqueur?' or 'Will the Chicago Bulls win the next NBA championship?') and then judge the probability of being correct. In some revision of opinion experiments (e.g. Phillips and Edwards, 1996), subjects first identify which of two data generators (e.g. urns) is most likely and then give a half-scale probability estimate of their confidence in this determination.

This seemingly minor change in the procedure has several important implications. First, Sniezek, Pease, and Switzer (1990) have shown that the act of choosing prior to the expression of confidence affects the subsequent judgments. Their results suggest that the choice followed by half-scale judgments leads to deeper processing of the information and to superior judgments. In other words, the change in the procedure may affect the psychological process invoked by the judge. Second, the change in the procedure imposes certain restrictions on the interpretation of the results. Note that the judge always starts with a binary classification reflecting his subjective beliefs about the 'true' state (e.g. absinthe is a liqueur, the sample is from urn A), and then provides a probability judgment that the decision is correct *conditional* upon this classification. Thus the probabilistic judgments made by the subjects (and analyzed by the experimenter) do not refer to the truth or falsehood of *objective* events with observable counterparts in the external world. Rather these are estimates of the probability that a certain *subjective* decision is correct. Clearly, in the half-scale format the analysis must, by definition, be conditional on the subjects' responses. Therefore, in this section we no longer examine the two possible types of analysis (and the paradoxical pattern of simultaneous over- and underconfidence). Instead, we focus on the case where the analysis of probability judgments is conditional upon the responses of the judges.

Finally, the half-scale procedure induces a slightly different pattern of results. Most empirical calibration studies using general-knowledge items document a consistent pattern of overconfidence under the half-range procedure (e.g. Lichtenstein *et al.*, 1982). Within this result, however, the relative frequency of correct judgments at 0.5 exceeds 50% (e.g. Figure 2 in Lichtenstein *et al.*, 1982, summarizing four experiments; Ronis and Yates, 1987). It is of interest to determine whether 'True + Error' models of the sort developed here can capture this overall pattern.

A natural extension of our models to the two-alternative procedure assumes that the subject considers one of the alternatives and, as in the regular full-scale task, experiences a degree of confidence x_{ij}. If $x_{ij} > \mathrm{f}(0.5,\ 0)$ (cf.equation (1)), the subject chooses it as the answer, and assesses the probability of being correct as y_{ij} (the full-scale response given x_{ij}). If $x_{ij} < \mathrm{f}(0.5,\ 0)$, the subject chooses the other alternative as the answer and assesses the probability of being correct as $1-y_{ij}$.

We used this approach in applying the log-odds model to the two-alternative case with responses restricted to 0.5, 0.6, ..., 0.9, 0.975. Exhibit 4 shows the results under a subset of the conditions examined in the full-range case (omitting the U-shaped distribution). Two comments are in order.

[3] Juslin *et al.* (1997) showed that this result is not universal. They examined a different version of the Binomial model in which the matching response mechanism is replaced by a deterministic one and found, invariably, overconfident judgments.

Exhibit 3. Mean relative confidence measures, $CONF_s$ and, O/U_s for eight illustrative examples of the log-odds model with six response categories (half-range procedure)

Prior distribution	m	$CONF_s$	O/U_s	Bias at $p = 0.5$
Uniform	0.5	0.020	0.013	−0.003
	1.0	0.070	0.058	−0.010
	1.5	0.111	0.096	−0.015
	2.0	0.156	0.138	−0.016
W-shaped	0.5	0.037	0.025	−0.017
	1.0	0.077	0.063	−0.017
	1.5	0.116	0.100	−0.017
	2.0	0.154	0.136	−0.018

Unlike Exhibit 2 where we were looking at the dual pattern, these results are plotted in the usual fashion for calibration studies, i.e. OP = f(SP). Also note that although we continue to use the labels SP and OP for the two axes, their interpretation is slightly different: SP denotes the subjective probability that the event was classified/predicted correctly by the judge and OP is the objective proportion of correct decisions. The calibration curves show agreement between the full- and half-range results at all responses except that of 0.5. Whereas in the full-range procedure the judgments were perfectly calibrated with $OP(0.5) = 0.5$, now they are slightly underconfident with $OP(0.5) > 0.5$ as observed in some empirical studies. The effects are summarized quantitatively in Exhibit 3 in terms of the mean $CONF_s$, the corresponding O/U and bias at $p = 0.5$. As in the full-scale case, the level of overconfidence (measured by $CONF_s$ and O/U_s) increases monotonically with the level of error.

We applied the binomial model to the half-range case in a slightly different fashion using other assumptions regarding the subject's judgments. Assume that for a given item whose alternatives have complementary probabilities correct of (0.5, 0.5), (0.6, 0.4) (0.7, 0.3), (0.8, 0.2) (0.9, 0.1) and (0.975, 0.025), the judge can always identify the most probable alternative (or pick one at random in the equiprobable case). On the basis of m independent pieces of information, b of which support the truth of the first alternative, the respondent needs to select a response. The same probability matching rule used in the full-range case is applied here, but the matching is performed across the six pairs.

We used the same values of m to model levels of error as in the full-range case in conjunction with a uniform distribution and a U-shaped distribution over the six pairs of probabilities, obtained by folding the W-shaped distribution considered earlier. The results are summarized in Exhibits 5 and 6. On average, all cases display greater overconfidence at higher levels of error variance (lower m). Whereas in the full-scale case the judgments were perfectly calibrated at $p = 0.5$, here they are biased in a direction that matches empirical data. This pattern is more pronounced and evident than in the log-odds model for the half-scale task. The S-curves obtained for high-variability cases resemble empirical results reported by Lichtenstein *et al.* (1977) for 'easy items' and by Gigerenzer, Hoffrage, and Kleinbölting (1991) for a 'representative set' of items in calibration studies.

FINAL REMARKS

This paper continues the line of work originally presented in Erev *et al.* (1994). We showed that a variety of prevalent empirical patterns of over- and underconfidence can be obtained from a single data set due simply to the presence of random error. Our illustrative models assumed that overt responses depend on the subject's true (well-calibrated) judgment as well as an error component. The results were that, unless the error variance is very small, the dual over- and underconfidence patterns appear. The

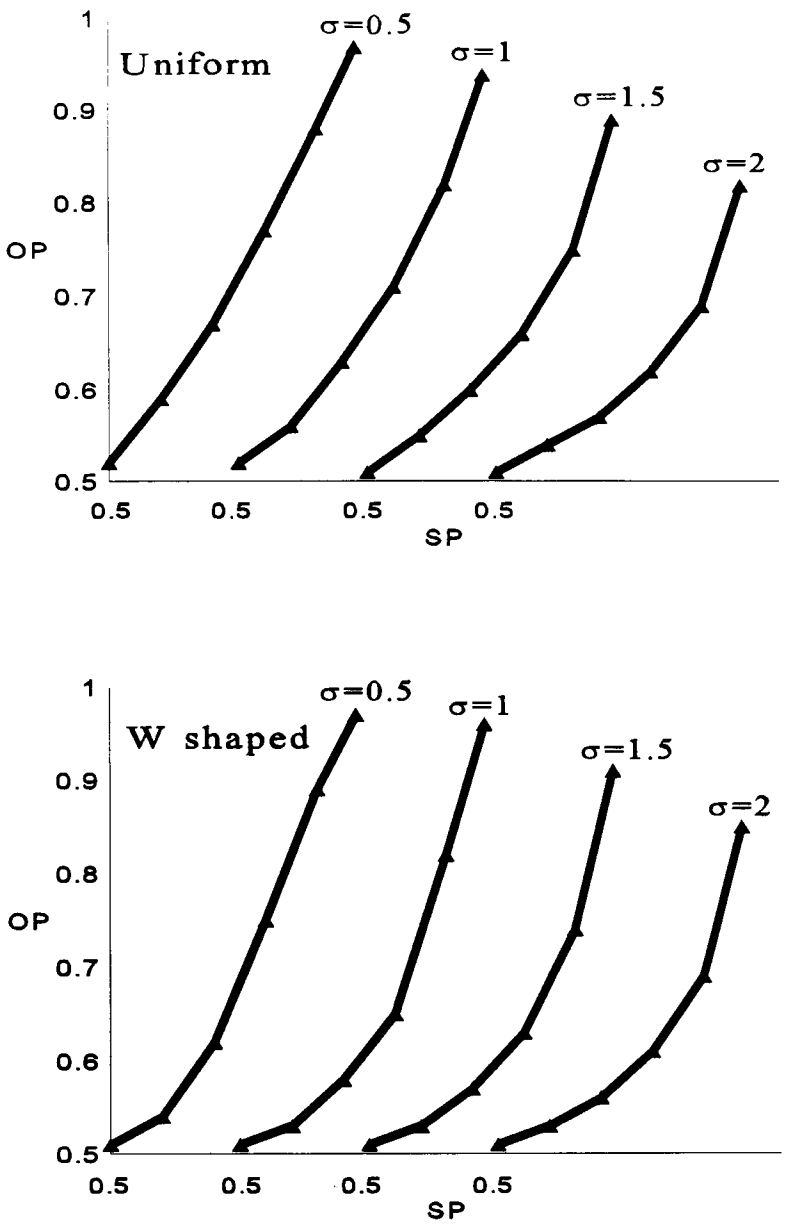

Exhibit 4. Mean objective probability of being correct (OP) as a function of subjective probability of being correct (SP) (the S-curve, denoted by ▲) for eight instances of the half-range log-odds model defined by two prior distributions (uniform and U-shaped) and four levels of error variance ($\sigma = 0.5, 1, 1.5, 2$). The abscissa increments in units of 0.1 beginning at 0.5 as indicated for each curve.

models we used had little to say about the psychological processes that may give rise to these judgments, nor are they wedded to a particular error distribution or a specific rule for combining true judgment and error. For demonstration purposes we assumed two very different formulations (log-odds and binomial) and applied them to both full- and half-range responses.

Exhibit 5. Mean relative confidence measure, $CONF_s$ and O/U_s, for eight illustrative examples of the Binomial model with six response categories (half-range procedure)

Prior distribution	m	$CONF_s$	O/U_s	Bias at $p = 0.5$
Uniform	20	0.014	0.000	−0.070
	15	0.016	0.000	−0.082
	10	0.020	0.000	−0.101
	5	0.027	0.000	−0.137
U-shaped	20	0.072	0.037	−0.044
	15	0.083	0.042	−0.053
	10	0.100	0.050	−0.068
	5	0.130	0.060	−0.103

It is possible, of course, to explore many alternative error models and experimental paradigms. Several recent papers have done precisely this and, without exception, have reached similar conclusions. Kleiter (1995) and Pfeifer (1994) have obtained spurious overconfidence with a model assuming beta distributions of error (which, of course, are closely related to the binomial); Gigerenzer, Hoffrage, and Kleinbölting's (1991) Probabilistic Mental Model assumes subjects have accurate knowledge of cue validities over an entire knowledge domain (corresponding to the assumption that $t_i = p_i$ in our system). Björkman (1994) extended this work by suggesting that cue validity is neither internally stable nor translated precisely into overt responses. As it happens, our binomial model is an ideal representation of Björkman's Internal Cue Theory. It handles all of his results, simply as a consequence of yielding overlapping response distributions conditional on t_i, and, as already shown, mimics the standard empirical findings. Soll (1996) simulated judgments generated according to another variation of the PMM (one with multiple and interrelated cues) combined with a binomial distribution of random errors and concluded that measures of (apparent) overconfidence vary directly and depend to a good degree on the amount of noise in the responses. Most recently, Juslin, Olsson, and Björkman (1997) considered effects of adding an error component at either the judgment or the response stage of a PMM-like model.

We are not making special claims regarding the log-odds and binomial models, although they are both reasonable. In fact, we believe that most models that (1) distinguish overt response from covert judgment and (2) explicitly disentangle true judgments and random error will lead to quite similar results. Recent work by Björkman (1994), Juslin *et al.* (1996), Kleiter (1995), Pfeifer (1994) and Soll (1996), as well as the four models described in this paper, strongly support this conjecture. We hope that these results will convince judgment researchers and applied decision analysts that any inference regarding the nature and magnitude of biases in probability judgment must take into account the type and amount of error involved in the process. In other words, 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.

We are not arguing that under- and overconfidence are necessarily entirely, or even primarily, statistical artifacts. The results presented by Erev *et al.* (1994), Pfeifer (1994), and in the previous sections indicate how error *can* induce the apparent over- and underconfidence. All these simulations show is that, if random error is sufficiently large, it can create the appearance of over- or underconfidence in cases where, in fact, the judge is well calibrated. In light of these results, our main thesis is that the relation between SP and OP in a particular context, and all inferences about the existence of biases in judgment, can be established only after controlling for random factors in judgment or response. It is possible that in the absence of error, real under- or overconfidence will emerge. But other

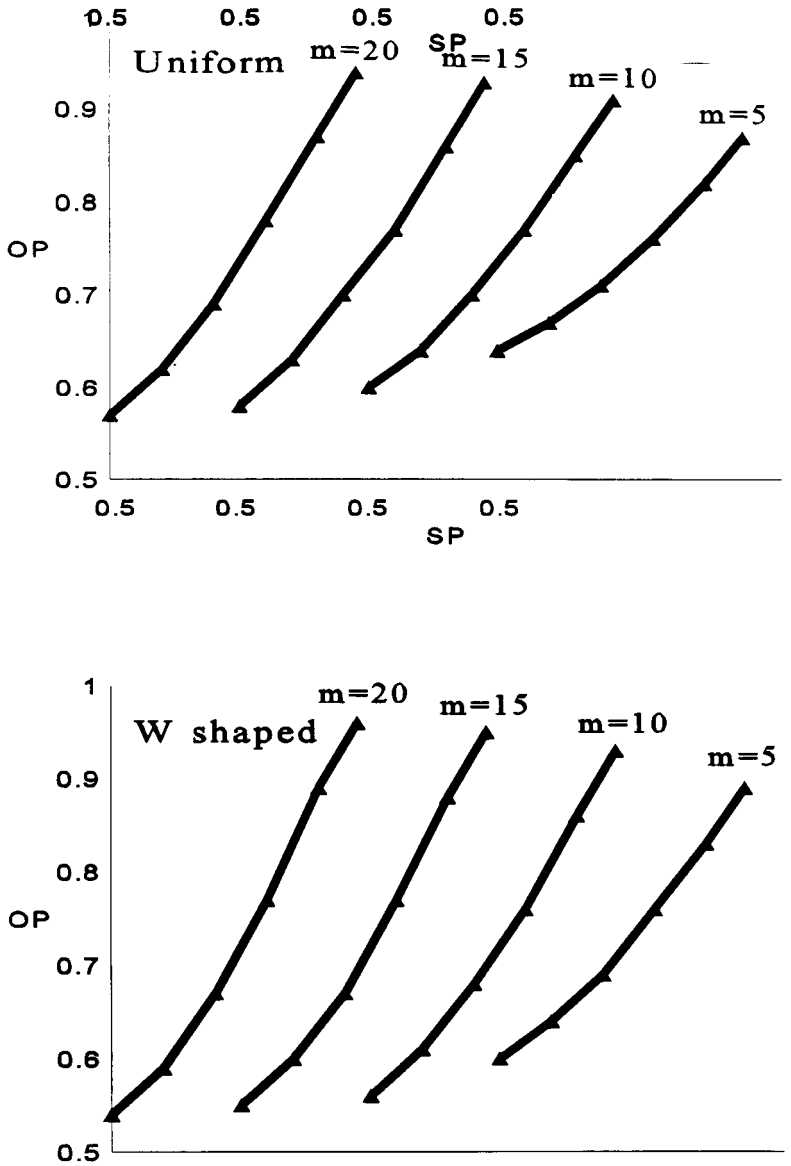

Exhibit 6. Mean objective probability of being correct (OP) as a function of subjective probability of being correct (SP) (the S-curve, denoted by ▲) for eight instances of the half-range binomial model defined by two prior distributions (uniform and W-shaped) and four sample sizes (n = 5, 10, 15, 20). The abscissa increments in units of 0.1 beginning at 0.5 as indicated for each curve.

patterns may obtain as well. A pleasant outcome under such circumstances would be that no bias is observed, but that true judgments are accurate.

One potential weakness of the previous results is that they are based on a set of assumptions about the magnitude of the error and of the prior distribution of opinions. Thus, their relevance to the

empirical study of probability judgment depends on the appropriateness of the assumptions regarding the magnitude of the error. We consider the assumptions to be reasonable, and feel reassured by the robust pattern observed over a variety of models. Others, however, may feel differently.

In the companion paper, (Budescu *et al*., 1997) we propose and illustrate a methodology designed to assess and characterize the degree of true over- or underconfidence with empirical data, i.e. in situations where the assumptions invoked in the various simulations are replaced with parameters estimated from the subjects' empirical judgments. More specifically, we show how this methodology can be applied to calibration studies to determine whether there is sufficient evidence to reject the hypothesis that judgments are perfectly calibrated.

ACKNOWLEDGEMENTS

Preparation of this paper was supported by National Science Foundation Grants No. SBR-9222159, SBR-9601281, and SBR-9632448, and by an Arnold O. Beckman Research Award from the Research Board of the University of Illinois.

We wish to thank Frank Yates, two anonymous reviewers, and the participants of the symposium on overconfidence at SPUDM-15, Jerusalem, Israel, 1995, for many useful and helpful comments.

REFERENCES

Ben Simon, A., Budescu, D. V. and Nevo, B. 'A comparative study of measures of partial knowledge in multiple choice tests', *Applied Psychological Measurement* (in press).

Björkman, M. 'Internal cue theory: Calibration and resolution of confidence in general knowledge', *Organizational Behavior and Human Decision Processes*, **58** (1994), 386–405.

Budescu, D. V. and Wallsten, T. S. 'Subjective estimation of vague and precise uncertainties', in Wright, G. and Ayton, P. (eds), *Judgment Forecasting* (pp. 63–82). Chichester: Wiley, 1987.

Budescu, D. V., Wallsten, T. S. and Au, W. 'On the importance of random error in the study of probability judgment. Part II: Using the Stochastic Judgment Model to detect systematic trends', *Journal of Behavioral Decision Making*, **10** (1997), 173–188.

Edwards, W. 'Conservatism in human information processing', In Kleinmuntz, B. (ed.), *Formal Representations of Human Judgment* (pp. 17–52), New York: Wiley.

Erev, I., Wallsten, T. S. and Budescu, D. V. 'Simultaneous over- and underconfidence: The role of error in judgment processes', *Psychological Review*, **101** (1994), 519–27.

Fischhoff, B. and Beyth-Marom, R. 'Hypothesis evaluation from a Bayesian perspective', *Psychological Review*, **90** (1983), 239–60.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. 'Probabilistic mental models: A Brunswikian theory of confidence', *Psychological Review*, **98** (1991), 506–28.

Howell, W. C. 'Compounding uncertainty from internal sources', *Journal of Experimental Psychology*, **95** (1972), 6–13.

Howell, W. C. and Burnett, S. A. 'Uncertainty measurement: A cognitive taxonomy', *Organizational Behavior and Human Performance*, **22** (1978), 45–68.

Juslin, P. 'The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items', *Organizational Behavior and Human Decision Processes*, **57** (1994), 226–46.

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

Kahneman, D. and Tversky, A. 'Variants of uncertainty', *Cognition*, **11** (1982), 143–157.

Keren, G. 'Calibration and probability judgments: Conceptual and methodological issues', *Acta Psychologica*, **77** (1991), 217–73.

Kleiter, G. D. 'A hidden probability model of overconfidence and hyperprecision. Comments on Erev, Wallsten & Budescu', unpublished working paper (1995).

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. 'Calibration of probabilities: The state of the art', in Jungermann, H. and de Zeeuw, G. (eds), *Decision Making and Change in Human Affairs* (pp. 275–324), Amsterdam: D. Reidel, 1977.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. 'Calibration of probabilites: The state of the art to 1980', in Kahneman, D., Slovic, P., and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–34), Cambridge: Cambridge University Press, 1982.

Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley, 1968.

McClelland, A. G. R. and Bolger, F. 'The calibration of subjective probabilites: Theories and models 1980–94', in Wright, G. and Ayton, P. (eds), *Subjective Probability* (pp. 453–84), Chichester: Wiley, 1994.

Murphy, A. H. and Winkler, R. L. 'Diagnostic verification of probability forecasts', *International Journal of Forecasting*, **7** (1992), 435–55.

Phillips, L. D. and Edwards, W. 'Conservatism in a simple probability inference task', *Journal of Experimental Psychology*, **72** (1966), 346–54.

Pfeifer, P. E. 'Are we overconfident in the belief that probability forecasters are overconfident?' *Organizational Behavior and Human Decision Processes*, **58** (1994), 203–13.

Rapoport, A. and Wallsten, T. S. 'Individual decision behavior', *Annual Review of Psychology*, **23** (1972), 131–76.

Ravinder, H. V. and Kleinmuntz, D. N. 'Random error in additive decompositions of multiattribute utility (with commentaries and reply)', *Journal of Behavioral Decision Making*, **4** (1991), 83–100.

Ravinder, H. G., Kleinmuntz, D. N., and Dyer, J. S. 'The reliability of subjective probabilities obtained through decomposition', *Management Science*, **34** (1988), 186–99.

Ronis, D. L. and Yates, F. 'Components of probability judgement accuracy: Individual consistency and effects of subject matter and assessment method', *Organizational Behavior and Human Decision Processes*, **40** (1987), 193–218.

Samuels, M. L. 'Statistical reversion toward the mean: More universal than regression toward the mean', *The American Statistician*, **45** (1991), 344–6.

Schoemaker, P. J. H. and Hershey, J. C. 'Utility measurement: Signal, noise and bias', *Organizational Behavior and Human Performance*, **52** (1992), 397–424.

Slovic, P. and Lichtenstein, S. 'Comparison of Bayesian and regression approaches to the study of information processing in judgment', *Organizational Behavior and Human Performance*, **6** (1971), 649–743.

Sniezek, J. A., Pease, P. W. and Switzer, F. S. *Organizational Behavior and Human Decision Performance*, **46** (1990), 246–82.

Soll, J. B. 'Determinants of miscalibration and over/underconfidence: The roles of random noise and ecological structure', *Organizational Behavior and Human Performance*, **65** (1996), 117–37.

Teigen, K. H. 'Variants of subjective probabilities: Concepts, norms and biases', in Wright, G. and Ayton, P. (Eds), *Subjective Probability* (pp. 211–38), Chichester: Wiley, 1994.

Thurstone, L. L. 'A law of comparative judgment', *Psychological Review*, **34** (1927a), 273–86.

Thurstone, L. L. 'Psychophysical analysis', *American Journal of Psychology*, **38** (1927b), 368–89.

Tversky, A. and Koehler, D. J. 'Support theory: A nonextensional representation of subjective probability', *Psychological Review*, **101** (1994), 547–67.

Wagenaar, W. A. and Keren, G. B. 'Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people', *Organizational Behavior and Human Decision Processes*, **36** (1985), 406–16.

Wallsten, T. S. and Budescu, D. V. 'Encoding subjective probabilities: A psychological and psychometric review', *Management Science*, **29** (1983), 151–73.

Wallsten, T. S., Budescu, D. V., and Zwick, R. 'Comparing the calibration and coherence of numerical and verbal probabilistic judgments', *Management Science*, **39** (1993), 176–90.

Yates, J. F. *Judgment and Decision Making*, Englewood Cliffs, NJ: Prentice Hall, 1990.

Authors' biographies:
David V. Budescu is a professor of Quantitative Psychology at the University of Illinois in Urbana-Champaign. His research interests are individual and group decision making, subjective probability, and behavioral statistics.

Ido Erev received a PhD in Psychology from UNC and postdoctoral training in Economics at the University of Pittsburgh. He is now a Senior Lecturer at the faculty of Industrial Engineering and management at the Technion. His research interests are human judgment and cognitive game-theoretic analysis of repeated decision tasks.

Thomas S. Wallsten is professor and director of the Cognitive Psychology Program at the University of North Carolina–Chapel Hill, as well as a former editor of the *Journal of Mathematical Psychology*. His research focuses on various aspects of subjective probability judgment and statement verification.

Authors' addresses:
David V. Budescu, Department of Psychology, The University of Illinois, Champaign, IL 61820, USA.

Ido Erev, Faculty of Industrial Engineering and Management, The Technion, Israel Institute of Technology, Haifa, Israel.

Thomas S. Wallsten, Cognitive Psychology Program, University of North Carolina — Chapel Hill, NC, USA.

