*Perception & Psychophysics*
*1995, 57 (2), 255–259*

# Reply to William R. Ferrell's paper "A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination"

MATS BJÖRKMAN, PETER JUSLIN,
and ANDERS WINMAN
*Uppsala University, Uppsala, Sweden*

*Ferrell's decision-variable partition model and our subjective distance model belong to the same family of Thurstonial models. The subjective distance model is limited to sensory discrimination with the method of constant stimuli and rooted in such notions as discriminal dispersion and sense distance. Ferrell's model is intended to be wider in scope and to apply to both cognitive and sensory tasks. Both models need supplementary assumptions to predict calibration phenomena. The point of departure for us is the fact that the model predicts underconfidence under "guessing" and the empirical finding that people are about 100% correct when they report "absolutely certain." Ferrell makes assumptions about cutoffs on the decision variable. The respondent is assumed to adjust or not adjust cutoffs according to "cues to difficulty." We disagree with Ferrell's claim that the hard–easy effect is* explained *by the respondent's failure to adjust cutoffs sufficiently when there is a change in level of difficulty, and argue that this amounts to little more than a translation of the hard–easy effect into the lingua of Ferrell's decision-variable partition model. Our argument is that the hard–easy effect is a consequence of the post hoc division of items according to solution probability. In addition, error variance may contribute to regression effects that enlarge the hard–easy effect. Finally, in contrast to Ferrell's position, we regard inference (cognitive uncertainty) and discrimination (sensory uncertainty) as different psychological processes. An understanding of calibration in these two areas requires separate models.*

Ferrell's (1995) critique of Björkman, Juslin, and Winman (1993) centers on four issues: (1) the model of sensory discrimination we proposed is the same as the two-alternative, forced-choice, half-range [2AFC(HR)] format in the "signal-detection model" (decision-variable partition model) presented by Ferrell and his colleagues (e.g., Ferrell & McGoey, 1980); (2) the model does not predict underconfidence; (3) the "hard–easy effect" has been observed with sensory judgments and is predicted by Ferrell's signal-detection model; and (4) there is no need for separate models of calibration of sensory and cognitive judgments. Below we reply to each of the four issues raised by Ferrell.

1. Ferrell's decision-variable partition model and our model belong to the same family of models, sometimes referred to as Thurstonian models (e.g., Luce, 1977, 1994). The model we presented was based on Thurstone's Case V (normal distributions, equal variances, zero correlation; see Thurstone, 1927a, 1927b), to which we added the most natural assumption that confidence increases monotonically with increasing distance between stimuli, a common assumption in psychophysical research (see, e.g., Johnson, 1939, and Festinger, 1943, for early examples). The main difference from these early studies lies in the fact that we interpreted confidence judgments as subjective probabilities with the purpose of investigating calibration, or realism of confidence.

Formally, Ferrell's 2AFC(HR) model is the same as our Thurstonian model, but we make *different assumptions* in order to derive predictions from the model, and we have *different applications* in mind. Turning to applications first, the subjective distance model was intended to deal with *sensory discrimination with the method of constant stimuli.* The assumptions of the model are rooted in a particular conception about the underlying psychological processes, namely, the discriminal activity in the nervous system of single subjects as modeled by Thurstonian notions such as discriminal dispersion, discriminal difference, and sense distance. This limits the applicability of the model to situations in which this account of the psychological processes seems justified. For instance, uncertainty in the context of general-knowledge questions will most often not reflect discriminal processes that are due to the inherent variability of the nervous system (see Point 4 below).

While we apply Thurstone's model to the sensory discrimination task for which it was originally developed, Ferrell has far-reaching ambitions with his more abstract "decision-variable model." For example, the 2AFC(HR) model was applied to sets of general-knowledge questions of the type "Absinthe is (a) a liquor, (b) a precious stone" (Ferrell & McGoey, 1980). We are not convinced that this generalization captures the psychological processes involved in acquiring and using general knowledge. In our view, answers to general-knowledge questions are typically *inferences* from self-generated, environmental cues and needs to be modeled accordingly (see Point 4 below and Juslin & Winman, in press). The signal-detection model for calibration proposed by Ferrell seems to focus on the process of associating an internal decision variable with a subjective probability judgment, quite regardless of the cognitive origin of this internal decision variable. That confidence assessment will involve generation and "scaling" of some internal variable seems fairly obvious, of course.

Our model, focused as it is on sensory processes, states that each stimulus (not a calibration experiment) generates a distribution of sensory magnitudes through the

The authors wish to express their gratitude to Henrik Olsson for valuable comments on an earlier version of this paper. The authors' mailing address is Department of Psychology, Uppsala University, Box 1854, S-751 48 Uppsala, Sweden.

constant fluctuations of the nervous system. Hence, although both models are formally identical Thurstonian models, we have different sample spaces in mind, namely, entire calibration experiments versus repeated presentations of single stimuli. Considerations like these are, of course, also relevant when we evaluate the appropriateness of the assumptions and claims about the quantitative fit of the models.

The difference between the models is evident from the fact that Ferrell discusses data from several published calibration experiments that concern a variety of tasks and contents, but *nowhere does he review and discuss data from sensory discrimination on single sensory continua with the method of constant stimuli, which was the topic of our paper* (i.e., aside from his discussion of data from our study). That the scope of our study is sensory discrimination as investigated with the method of constant stimuli is clearly stated in the introduction to our paper.

In order to derive strong predictions about calibration phenomena, Thurstonian models need to be supplemented with additional assumptions. Ferrell often makes the assumption that the cutoffs remain fixed across different data sets. We do not make this assumption. We assume only that confidence increases monotonically with the subjective distance between stimuli; cutoffs may or may not differ between sets of stimuli. Instead, we hypothesized that the imbalance that creates underconfidence in the lowest confidence category persists into the higher confidence categories and leads to a general underconfidence bias. This seemed most natural in view of previous studies showing strong underconfidence in the guessing category and close to 100% correct when subjects reported "absolutely certain" (see the next point). So, Ferrell and we have *different conceptions about the applicability* and we make *different assumptions* in order to derive predictions from the Thurstonian model. These differences are far from trivial. Nevertheless, in our original paper, we should have commented on the similarities and differences between our model and Ferrell's 2AFC(HR) model.

2. The model, that is, a normal distribution of sensory discriminal differences with cutoffs delimiting the intervals corresponding to confidence categories (see Figure 1 in our paper and Figure 1 in Ferrell's paper), does not by itself predict any *general calibration phenomena*, for example, general over- or underconfidence. One interesting prediction, however, follows immediately from the model, namely that there will be more correct than wrong responses when the subject reports "guessing" (Figure 1 in our paper). This holds true with the only qualification that the person uses the guessing category (with the trivial exception of identical stimuli). This "local" underconfidence has been consistently confirmed in a number of experiments (see, e.g., Peirce & Jastrow, 1884, and Fullerton & Cattell, 1892; Garrett, 1922; Griffing, 1895). For a variety of sensory continua and varying overall proportion correct, *when people feel that they are guessing, they are right twice as often as they are wrong*. As documented in our paper, this phenomenon was clearly acknowledged by the early psychophysicists and it made them speculate about "unconscious processes." This phenomenon is better explained by the idea of a discriminal process that creates an asymmetry in the lowest confidence category. Note, however, that since the confidence categories are not expressed as subjective probabilities in the early studies, we cannot tell from these data whether there is an underconfidence bias in the higher confidence categories. The question in our study was: Does the underconfidence bias in the guessing category persist through the following confidence categories ($x_t$ = .6, .7, .8, .9, and 1.0), or does it change into overconfidence or good calibration?

We found it reasonable to *hypothesize* that underconfidence would characterize the entire range of confidence assessments. We noted that if the imbalance in the lowest category is large and subjects use the following categories, the underconfidence bias is likely to generalize to the higher confidence categories:

> In the lowest category of confidence, there will always be more correct than wrong responses.... At this end of the scale, we can safely predict underconfidence, $x_t < c_t$. Furthermore, it is not likely that the disagreement between confidence and proportion correct will disappear in the following categories. If, for example, $c_t$ = .65 for $x_t$ = .5, we will necessarily find $c_t$ > .65 for $x_t$ = .6, and so on. From a value greater than .5 at $x_t$ = .5, one should expect a smooth increase of proportion correct $c_t$ with increasing confidence. Hence, we hypothesize that the imbalance between correct and wrong responses at $x_t$ = .5 extends to the higher categories of confidence with the consequence of an average underconfidence, $\bar{x} < \bar{c}$. (Björkman et al., 1993, p. 77)

What we had in mind was a finding made long before such notions as calibration and over- and underconfidence were established. In our historical review (Björkman et al., 1993, p. 76), we reported that Johnson (1939) and Festinger (1943) had found that the ogival relationship between confidence and stimulus difference was flatter than that for proportion correct and stimulus difference. Confidence in these studies was not defined explicitly to the subjects as expected percentages—that is, that confidence .xx means that .xx of the responses should be correct—but they make one suspect that the subjects made more efficient discriminations than is expressed in their confidence assessments. Also, as noted above, when subjects report that they are "certain," they most often have close to 100% correct decisions. Hence, although these experiments were not concerned with calibration, they tentatively suggest that the calibration curve should fall above the diagonal, not only at $x_t$ = .5 but also in succeeding categories, and end in the point 1.0/1.0.

In his comment, Ferrell (1995) argues that "Björkman et al. (1993) mistakenly generalize the underconfidence at the lowest response value to the entire calibration curve" (p. 249). He goes through elaborate exercises to show that *when the model is not conjoined with the above assumption that the asymmetry in the lowest confidence category generalizes into the higher confidence categories*, we can have both general over- and underconfidence. But this is obvious, and we have not argued otherwise. As should be evident from the quotation above, we are aware of this fact and we arrive at the prediction of general underconfidence by means of a supplementary assumption. To derive strong predictions from the framework provided by the Thurstonian model, additional assumptions are needed, and Ferrell's favorite assumption is to keep cutoffs fixed across different data sets.

Ferrell notes that calibration studies of general knowledge have often shown underconfidence in the lowest category accompanied by overconfidence, or good calibration, in the succeeding categories. This has no bearing on our hypothesis and our results, which very clearly support the hypothesis of underconfidence. Furthermore, the underconfidence in the guessing category observed in cognitive tasks is not even close to the magnitude observed in the sensory tasks. This phenomenon is more properly handled by theories that focus on the role of errors in the process from "objective, environmental probabilities" to overt confidence judgments (see Björkman, 1994; Erev, Wallsten, & Budescu, 1994; Zoll, 1994).

The reason why Ferrell and we make different predictions from the application of a Thurstonian model is simply that we have made different auxiliary assumptions. The issue of what assumptions are reasonable under different conditions is, of course, the real issue that should be discussed. Here we will only comment that, in general, we feel that both formulations are too weak and leave too much to be determined by data. The development of stronger models of confidence in sensory discrimination seems to be an important task for future research.

3. The "hard–easy effect," as originally encountered, refers to the observation of overconfidence for hard tasks (low solution probability) and underconfidence for easy tasks (high solution probability) (see, e.g., Juslin, 1993b; Lichtenstein & Fischhoff, 1977). Ferrell's favorite assumption in terms of providing the signal-detection model with predictive power is to keep cutoffs fixed across different data sets. He claims that, in this way, the signal detection model can account for the hard–easy effect: "The effect is explained in the context of the model as the consequence of the respondent's failing to adjust response criteria appropriately, or doing so insufficiently, when there is a change in task difficulty (Ferrell, 1995, p. 250).

First, we have some difficulty in accepting this as an *explanation* of the hard–easy effect. If we take data known to be characterized by a hard–easy effect and fit the signal-detection model to these data, this *must* appear as an insufficient adjustment of the cutoffs. Similarly, if we have several data sets with the same level of over/underconfidence but different proportions correct, and fit the signal-detection model to these data, the response criteria *must* have changed across these data sets. This amounts to little more than a translation of the hard–easy effect into the lingua of Ferrell's signal-detection model. The real psychological significance lies in explaining *why* the cutoffs are too strict (underconfidence) or too generous (overconfidence) and why the cutoffs are or are not adjusted sufficiently when difficulty is varied across data sets. So far, the signal-detection model provides little guidance when it comes to these issues.

One source of the hard–easy effect in studies using general-knowledge items seems to be the post hoc division of items into hard and easy on the basis of their solution probability, which creates unrepresentatively low cue validities (hard items) or unrepresentatively high cue validities (easy items) in the resulting subsets of items (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1993a). Another factor that contributes to the hard–easy effect is the errors that occur in the process of learning the ecological cue validities which create something like a regression effect between ecological cue validity and overt confidence (see Zoll, 1994). Regardless of which factor dominates in a particular experimental study, once that data are processed in terms of Ferrell's signal-detection model, the result will appear as an insufficient adjustment of response criteria.

The primary purpose in our study was not to test the hard–easy effect but to test the prediction of an underconfidence bias with a number of stimulus sets that provided a reasonable variation of difficulty. Data were aggregated for several stimulus units in order to get stable calibration curves. Since we predicted underconfidence for all stimulus sets with the motivation discussed above, we expected no hard–easy effect in the sense of overconfidence for hard tasks.

The results were $\bar{x} - \bar{c} = -.118$ and $\bar{c} = .885$ for the easy set of weights, $\bar{x} - \bar{c} = -.115$ and $\bar{c} = .782$ for the hard set of weights, $\bar{x} - \bar{c} = -.145$ and $\bar{c} = .747$ for the easy rectangles, $\bar{x} - \bar{c} = -.119$ and $\bar{c} = .667$ for the hard rectangles, where a negative $\bar{x} - \bar{c}$ indicates underconfidence and $\bar{c}$ is the proportion of correct decisions (difficulty). For hard rectangles, mean confidence was .548 on a scale on the interval [.5, 1.0]. Since confidence is unlikely to increase if the stimulus set is made more difficult, these data suggested that there should be underconfidence at all levels of difficulty (although it will have to converge to zero as proportion correct approaches .5, of course). The same conclusion is, indeed, suggested by the observation of a flatter ogival function between confidence and stimulus difference as compared with the function between proportion correct and stimulus difference, implying underconfidence regardless of the stimulus difference (see Festinger, 1943; Johnson, 1939).

Ferrell applies the signal-detection model with the assumption of fixed cutoffs to several experiments by Lichtenstein and Fischhoff (1977), in the paper by Ferrell and McGoey (1980), and to Keren's (1988) perception experiment (notice that Keren's experiment is of the single-stimulus variety and not directly comparable to our pair-comparisons data). However, the assumption of fixed cutoffs cannot predict the *consistent* underconfidence we found for two sensory continua and a range of proportion correct from .67 to .89. In sum, the data we reported were consistent with a prediction of underconfidence for all four stimulus sets, as motivated by the assumption discussed above but not by the assumption of fixed cutoffs across the four stimulus sets.

4. Ferrell claims that there is no need for separate models to account for confidence in sensory discrimination and confidence in one's general knowledge. He argues that the signal-detection model fits not only sensory judgments, but cognitive ones as well. As will be made clear below, "fit" is not sufficient; we also want to understand the processes underlying calibration phenomena. No doubt the scaling of an internal "decision variable" to generate subjective probability assessments is part of the cognitive processes involved in probability assessment. The assumptions about this process that form part of Ferrell's signal-detection model do not seem very controversial. But the real psychological significance lies in *explaining why the cutoffs are where they are* (to speak in the language of Thurstonian models), not in the ad hoc fitting of a model of the scaling process to various data sets, withholding or removing the assumption of fixed cutoffs as required by the data. As already repeatedly noted in the literature (e.g., Gigerenzer et al., 1991; Juslin, 1993a; Keren, 1991), although the signal-detection model is useful in providing a systematic exploration of the signal-detection representation of calibration data for different response formats, the model has so far been of little use for understanding the cognitive processes involved in probability assessment.

Briefly, the reason why different models are needed is the following. Confidence in sensory discrimination is a function of the *subjective distance* between stimuli, an assumption shared by Ferrell's model. In the sensory domain, the uncertainty reflects the *less than perfect reliability (variability) of the coding system*, which creates a discriminal process. The discriminal process as modeled in terms of Thurstone's Case V suggests one interesting difference between cognitive and strictly sensory tasks; the responses by a number of subjects presented with a single sensory stimulus are stochastically independent across subjects, a prediction supported by data (Juslin, Winman, & Persson, in press). Among other things, this independence suggests that it should be difficult to select "misleading items" when uncertainty is strictly sensory (see Juslin et al., in press, for a discussion of the *response-independence model*).

The issue of separate models (or theories) is perhaps the most interesting one in the present discussion. It has been treated in detail by Ferrell (1994) in a comment on a paper by Winman and Juslin (1993). For a more complete discussion, the reader is referred to these papers and to the reply by Juslin and Winman (in press). An important future research task is a more complete exploration of the differences between sensory and cognitive uncertainty when it comes to such phenomena as *calibration*, *coherence* with the rules of the probability calculus (e.g., additivity), and *ambiguity avoidance* (see Einhorn & Hogarth, 1985; Heath & Tversky, 1991).

Ferrell argues that the signal-detection model fits half- and full-range judgments and explains or accommodates essentially all the robust experimental findings for calibration, including the "hard–easy effect" described above and the base-rate effect. However, the model "fits," "explains," and "accommodates" in a typical ad hoc manner. Figure 4, with data from Juslin (1993b), is an illustrative example. The cutoffs were first determined for the random set of items and then used, together with the experimental value of $p(C)$, to compute the calibration curve for the selected set. Ferrell's conclusion is that the difference in calibration for the two sets of questions is well explained by the model as a result of the greater difficulty of the selected set. However, there is a theoretical background to the experiment, namely the ecological model described in detail by Juslin (1993a, 1993b). This model *predicts* overconfidence in the selected set and good calibration in the random set without fitting parameters from data, *something the signal-detection model cannot do*. The calibration curve for the random set represents a remarkable achievement on the part of the subjects (replicated in several experiments), and the ecological model provides an account of this achievement.

The point is *not* to suggest that Ferrell claims that the signal-detection model can account for these issues. He certainly doesn't. The point is that these are questions that *should* be accounted for by any psychological theory of confidence in general knowledge, and that the answers are likely to be *different* from those that are relevant to sensory discrimination.

## REFERENCES

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

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*, **54**, 75-81.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, **93**, 433-461.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, **3**, 519-527.

Ferrell, W. R. (1994). Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology*, **35**, 297-314.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception & Psychophysics*, **57**, 246-254.

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

Festinger, L. (1943). Studies in decision: I. Decision time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, **32**, 291-306.

Fullerton, G. S., & Cattell, J. M. (1892). On the perception of small differences. *Publications of the University of Pennsylvania* (No. 2).

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, No. 56.

GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

GRIFFING, J. H. (1895). On sensations from pressure and impact. *Psychological Review Monograph*, **1**, 1-88.

HEATH, C., & TVERSKY, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk & Uncertainty*, **4**, 5-28.

JOHNSON, D. M. (1939). Confidence and speed in two-category judgment. *Archives of Psychology*, No. 341.

JUSLIN, P. (1993a). *An ecological model of realism of confidence 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., & WINMAN, A. (in press). Reply to William R. Ferrell's paper "Calibration of sensory and cognitive judgments: A single model for both." *Scandinavian Journal of Psychology*.

JUSLIN, P., WINMAN, A., & PERSSON, T. (in press). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Cognition*.

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.

LICHTENSTEIN, S., & FISCHHOFF, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance*, **20**, 159-183.

LUCE, R. D. (1977). Thurstone's discriminal processes fifty years later. *Psychometrika*, **42**, 461-489.

LUCE, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review*, **101**, 271-277.

PEIRCE, C. S., & JASTROW, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences*, **3**, 78-83.

THURSTONE, L. L. (1927a). A law of comparative judgment. *Psychological Review*, **34**, 273-286.

THURSTONE, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, **38**, 368-369.

WINMAN, A., & JUSLIN, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, **34**, 135-148.

ZOLL, J. B. (1994). *Determinants of miscalibration and over/underconfidence: The interaction between random noise and the ecology.* Unpublished manuscript, University of Chicago, Graduate School of Business.

(Manuscript received December 28, 1993;
revision accepted for publication October 17, 1994.)