On the calibration of probability judgements. Some critical comments and alternative perspectives

Citation for published version (APA):

Keren, G. (1997). On the calibration of probability judgements. Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, 10, 269-279. [https://doi.org/10.1002/\(SICI\)1099-0771\(199709\)10:3<269::AID-BDM281>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<269::AID-BDM281>3.0.CO;2-L)

DOI:

[10.1002/\(SICI\)1099-0771\(199709\)10:3<269::AID-BDM281>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<269::AID-BDM281>3.0.CO;2-L)

Document status and date:

Published: 01/01/1997

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

On The Calibration of Probability Judgments: Some Critical Comments and Alternative Perspectives

GIDEON KEREN

University of Technology, Eindhoven, The Netherlands

Calibration of probability judgments has attracted in recent years an increasing number of researchers as reflected by an expanding number of articles in the literature on judgment and decision making. The underlying fundamental question that stimulated this line of research concerns the standards by which probability judgments could (or should) be assessed and evaluated. The most common (though certainly not exclusive) accepted criterion is what has been termed ‘calibration’, the roots of which can be traced in the well-known *Brier score* (Brier, 1950) and subsequent modifications (e.g. Murphy, 1973; Yates, 1982, 1988). Two main criteria that evolved from this line of research are customarily referred to as *calibration* and *resolution*. Calibration (or reliability) supposedly measures the accuracy of probability judgments whereas resolution measures the diagnosticity (or discriminability) of these judgments. The two major substantive and pervasive findings (e.g. Lichtenstein, Fischhoff, and Phillips, 1982; Keren, 1991) are overconfidence and the interaction between the amount of overconfidence and difficulty of the task, the so-called hard–easy effect.

Several problems have been raised with regard to research on calibration, and in this commentary I would like to focus on three of them. First, calibration studies assume (implicitly or explicitly) that probabilities are subjective (e.g. Lichtenstein, Fischhoff, and Phillips, 1982) yet evaluate them by a frequentistic criterion (Gigerenzer, 1991; Keren, 1991). The validity of such a procedure remains controversial.

A second problem concerns the possible tradeoff between calibration and resolution. Yates (1982) noted that calibration and resolution are not completely independent of each other, and Keren (1991) claimed that the requirements for maximizing calibration (i.e. minimizing the discrepancies between probability judgments and the corresponding reality) and achieving high resolution may often be incompatible. A similar point has been recently made by Yaniv and Foster (1995), who studied the evaluation of interval judgments.

A third problem concerns the analysis and interpretation of calibration studies. Specifically, Erev, Wallsten, and Budescu (1994) have eloquently described the importance of regression toward the mean in interpreting calibration studies. Similar conclusions have been reached independently by Pfeifer (1994). In a nutshell, the contribution of the papers by Erev *et al.* and Pfeifer is in pointing out that both overconfidence and the hard–easy effect may, at least to some degree, be an artifact due to regression toward the mean.

In reflecting on the articles in this special volume, I will focus on these three issues and examine how they are treated by the different authors. I will end this commentary by raising the question of what has been learned from thirty years of research on calibration of probabilities, and will offer a brief (and somewhat skeptical) answer to the question.

RANDOM ERROR MODELS

A common underlying thread of several papers to which this commentary is addressed (i.e. Budescu, Erev, Wallsten (Parts I and II); Juslin, Olsson, and Björkman; Wallsten, Budescu, Erev, and Diederich) is the phenomenon of regression-toward-the-mean (or in the more general case, reversion to the mean). They cite, and heavily hinge on, the paper by Erev, Wallsten, and Budescu (1994). Notwithstanding, and certainly not undermining, the importance of the contribution by Erev *et al.* (1994) and Pfeifer (1994), it is important to stress two points.

First, *regression toward the mean is a statistical reality* (but not more than that!) which applies to a wide variety of phenomena (it was originally observed by Galton, 1877, in the context of heredity). Unfortunately, the term has often been misinterpreted as if it implies some causal explanation. The late Louis Guttman (1977) has claimed that 'There is no regression toward the mean . . . The verb "to regress" has no mathematical definition, although the noun "regression" unfortunately is attached to one. A regression is merely a set of conditional averages, usually of arithmetic means' (p. 90). Guttman was clearly not trying to deny that there is a statistical phenomenon referred to as regression toward the mean. My understanding is that he simply wanted to warn researchers that they should leave it at that and not add to it unwarranted interpretations, specifically causal ones. Similarly, in the context of test theory, Kerlinger (1986) pointed out that 'Test scores change as a statistical fact of life: On retest, on the average, they regress toward the mean' (p. 296).

The second point is that while most certainly regression toward the mean may induce, in the technical sense, overconfidence (and under certain conditions underconfidence; e.g. Keren and van Bolhuis, 1995), it does not necessarily rule out the possibility of genuine overconfidence (beyond and above that created by regression toward the mean). This point is also stressed by most of the authors in the present volume.

Any estimation task, not just of probabilities, is exposed to potential effects of regression to the mean (Harrison and March, 1984). Even if estimates are unbiased, they are still subject to random error (or noise), and the larger this random error, the larger the so-called regression effect. While regression effects *per se* are simply statistical facts and as such are technical in nature, the random error can be seen as originating from different sources. Specifically, the random error can arise from stochastic properties of the world, or from problems associated with measurement (Harrison and March, 1984) and at the same time may also reflect differences in expertise. Both Budesu *et al.* and Juslin *et al.* analyze potential sources that may affect the size of the random error.

Juslin *et al.* adopt as their starting point a hybrid framework based on the theory of *Probabilistic Mental Models* (Gigerenzer, Hoffrage, and Kleinbölting, 1991), and *Internal Cue Theory* (Björkman, 1994). According to this framework, people acquire mental representations, termed internal probabilities (P_i) of ecological probabilities (V_i), which are relative frequencies in a reference class of objects (or situations in the environment). The model assumes three necessary requirements for being well calibrated, namely:

- (1) *Cognitive adjustment*, which requires that internal probabilities should equal the corresponding ecological probabilities.
- (2) *Error-free translation*, which assumes that (observed) subjective probability judgments are equal to the internal probabilities. In other words, the translation of the internal state of uncertainty to an overt response should be error-free.
- (3) *Representative design*, which requires selection of tasks such that the probabilities of the events in the task-sample should coincide with the corresponding ecological probabilities. Violation of each of these assumptions, according to Juslin *et al.* will affect the random error and in turn reduce the quality of calibration.

The third assumption is probably the most important one in the conceptual framework of Juslin *et al.* and, at the same time, also the most controversial one. Several researchers have recently advocated a so-called ecological approach to calibration studies in which assumption (3) is central. Elsewhere (Keren and van Bolhuis, 1995) I have elaborated on some of the problems associated with this approach and will reiterate them here only in brief.

To begin with, the approach adopts a frequentistic interpretation in which confidence or probability assessments of unique events are considered meaningless. Such an approach drastically limits the scope of behavioral decision making in studying how people cope with uncertainty and overlooks the fact that most assessments of uncertainty in daily life cannot be justified on frequentistic grounds. In fact, almost all calibration studies, certainly those who use general knowledge questions, essentially elicit subjective probabilities.

Even if we accept for a moment the frequentistic view, the notion of a random or representative sample is problematic in this context (Kahneman and Tversky, 1996). First, when speaking about a representative design, the question is representative with regard to what. If it is meant with regard to difficulty¹ of the items, it would

¹ The notion of difficulty has been central in the calibration literature (specifically with regard to the hard–easy effect) but little attention has been given to the multiple meanings of difficulty and the manner by which it is assessed. Thus difficulty in the context of calibration studies may be assessed by the proportion of correct items, by the amount of required processing (measured, for instance, by reaction times), by the necessary required knowledge or expertise, or by experts' judgments.

obviously be dependent on the population of subjects (difficulty of a set of items may differ for different sub-populations). Alternatively, one may think about a sample that is representative with regard to the relevant domain (e.g. cities in Germany with populations of more than 100,000 inhabitants). Finally, representative may simply imply the requirement of a random sampling procedure.

The notion of a random sample from the relevant reference set may also be problematic. Why necessarily random? In real life, the assessor may not always be confronted with a random sample, and in many cases the assessor may not know in advance whether the sample is random or not. It should be noted that (regardless of the interpretation of a representative sample one is willing to adopt) a sample that was chosen in an informal way, could also occur by random sampling. The normative requirement of probability judgments being well calibrated should apply to any sample of items. There are also difficulties associated with the concept of a relevant reference set. Relevance would depend on the particular context and, in any event, remains a subjective judgment. In the experimental laboratory, there is no insurance that subject and experimenter would necessarily adopt the same reference set. More important, in real life, where one is confronted with unique events (rather than with repeated trials), the relevant reference set will often depend entirely on the subjective judgment of the forecaster.

Juslin *et al.* (Exhibit 2) present data collected from 25 tasks, asserting that item samples have been generated such as to satisfy requirement (3), and conclude that the prediction derived by Gigerenzer *et al.* (1991) and Juslin (1994) 'that representative selection of items should moderate or even eliminate the overconfidence phenomenon thus receives some support in these data'. The support is indeed very limited since the average difficulty of the 25 tasks is very modest and, as the authors themselves notice, there is a clear hard–easy effect. Thus Juslin *et al.* correctly point out that the hard–easy effect across task domains in Exhibit 2 is inconsistent with the conjunction of assumptions (1), (2), and (3). Moreover, as also noted by Juslin *et al.*, an increasing number of studies (e.g. Brenner *et al.*, 1995; Budescu *et al.* in this issue; Griffin and Tversky, 1992; Keren, 1987; Soll, 1996; Suantak, Bolger, and Ferrell, 1996) that have made deliberate and careful attempts to employ a representative design have shown clear overconfidence.² All this evidence casts serious doubt on whether assumption (3) is indeed warranted.³ Notwithstanding, the nature of the task and the specific sample of items may potentially be one of the sources affecting the error variance and hence indirectly the amount of under- or overconfidence.

Another source of error discussed by Juslin *et al.* concerns assumption (1). According to their model, the translation of ecological probabilities (supposedly reflected by the environment) into internal probabilities may be contaminated. Specifically, since according to their approach the cues of the environment have to be assessed in frequentistic terms, and since even in the ideal case the assessor may have a limited sample size, the assessments of the ecological probabilities will not be error-free. Note, however, that both the internal probabilities and the ecological probabilities are constructs that are unknown. The former are unobservable and the latter can at best be assessed under very limited circumstances. Indeed, the question remains of what exactly is meant by ecological probabilities and how should they be assessed. What, for instance, is the ecological probability of a general knowledge question like 'Munich has a larger population than Frankfurt'?

The third source of error variance discussed by Juslin *et al.* (and independently by Budescu *et al.*) concerns the conversion of an internal probability to the overt response of a probability judgment (assumption (2) above). Internal probabilities are unobservable and thus cannot be matched directly with external responses. Nevertheless, insight into this source of error variance can be obtained by examining elicitation procedures and different response scales. Relevant issues here are whether the probability scale is half-or full-range (discussed by several papers to which this commentary is addressed; see also Ronis and Yates, 1987), whether it is a choice or no-choice situation (Ronis and Yates, 1987), whether the judge's propositions are discrete or continuous (Peterson and Pitz, 1988). Though a detailed discussion of these different methods is beyond the scope of the present commentary, suffice it to say that the empirical results are mixed and it is difficult to draw unequivocal conclusions. For example, based on their empirical results, Juslin *et al.* suggest that good calibration or even underconfidence can be obtained

² One wonders how many additional studies are still needed in order to disconfirm the representative design hypothesis.

³ The debate in the recent literature regarding the merits of the ecological validity approach may lead one to wonder whether some authors exhibit what may be termed 'ecological hypocrisy', acting as if random sampling from an almanac is all it takes to achieve ecological validity.

in the half-range method, whereas (using the same material) employing the full-range method resulted in overconfidence. In contrast, Ronis and Yates (1987) report results that favor the full-range method, which yielded significantly less overconfidence. Ronis and Yates (1987), however, noticed that several subjects in the full-range condition gave probabilities below 0.50, which are difficult to interpret. Thus, additional research on elicitation procedures is clearly needed in order to acquire a more transparent picture of the relevant factors. Notwithstanding, one can assert with relative certainty that elicitation procedures may indeed contribute to the error variance as suggested by both Juslin *et al.* and Budescu *et al.*

A different question, regardless of the source of error, concerns the nature of its distribution. Budescu *et al.* discuss alternative models of errors. Note, however, that since the error variance is determined by different factors (the weight of which may differ from one situation to the other), it is probably difficult to talk about the 'ideal' or best-fitting model. For instance, both Juslin *et al.* and Budescu *et al.* propose a binomial model that apparently offers an ideal representation of Björkman's (1994) internal cue theory, which is based on certain assumptions regarding the nature of the error. It is, however, not necessarily the case that this binomial model would provide the best fit for all sorts of data.

Taken together, the articles by Budescu *et al.* (Part I) and Juslin *et al.* offer a useful analysis of the possible sources for the error variance and its implications for interpreting calibration studies. Not undermining these contributions, several questions and issues remain open. First, random error may be affected by more than one factor. There is no direct way to separate these factors and assess their relative contribution. Second, the magnitude of the random error affects the degree of goodness of calibration indirectly by determining the extent to which regression toward the mean will be present. As mentioned above, regression toward the mean remains a statistical fact and should not be interpreted otherwise. Third, and most important, regression toward the mean, even when present, does not rule out genuine overconfidence, a point which is endorsed (implicitly or explicitly) by all the articles in this issue.

Budescu *et al.* (Part II) offer an interesting method for determining whether overconfidence observed in empirical data is solely the result of the presence of error (and consequently due to regression toward the mean) or whether it also reflects a genuine systematic judgment bias. Their analysis introduces many assumptions, the psychological meaning of which should be closely examined. Here I would like to limit myself to two related issues. First, as the authors state, their proposed method attempts to establish the relation between *subjective probabilities* and *objective probabilities* in a particular context while controlling for the effects of random factors. Like many other authors, they overlook the inherent difficulties associated with comparing subjective and objective probabilities. What do the authors mean by subjective and objective probabilities? What are, for instance, the objective probabilities for the most common task in calibration studies, namely, general knowledge questions? Are objective probabilities necessarily associated with frequencies? It requires intricate mind twisting to interpret calibration of general knowledge questions in frequentistic terms.

While Budescu *et al.* (Part II) remain silent with regard to these issues, they impose an explicit requirement for their proposed procedure which is closely related to the above issue. Specifically, for their procedure of separating between genuine effects and those due to random error, they need to estimate the error variance component σ_e^2 . A valid estimation of σ_e^2 entails the availability of independent repeated judgments of the same items. Budescu *et al.* point out that 'unfortunately, and surprising, such data with almanac questions are rarely found'. In my opinion, it is neither unfortunate nor surprising that such data are scarce. Data collection in calibration experiments is often problematic since it entails repetition of a boring and tiresome task, certainly when it involves the popular task of answering general knowledge questions. Indeed, one may question the reliability of data obtained from large sets of general knowledge questions. Is it really reasonable to assume that the amount of attention and processing capacity allotted to the first item is the same as for the 50th or the 100th item? Variability due to a boring and tedious task in which indifference may be escalating as the number of items increases can of course be hidden under the general label of random error but one may then wonder about the random character of the error component. The assumption that repeated items, even if spread within a long list, are independent is highly questionable. The authors' proposal to present repeated items in different forms may also be problematic: subjects may either recognize the similar forms or, alternatively, the different forms may be vulnerable to framing effects. Complementary forms despite their apparent similarity may not always be interchangeable (they are not necessarily symmetrical) since the reference point in each of two complementary forms may be different (Tversky, 1977).

OVERCONFIDENCE — REAL OR ARTIFACT?

Despite the methodological and statistical difficulties in identifying the different sources (real or apparent) for overconfidence, I assert that there is overwhelming evidence to suggest that genuine overconfidence, even if not entirely universal, is a common phenomenon above and beyond the overconfidence due to random error. There are clearly conditions under which overconfidence is moderated and even eliminated, yet the phenomenon is general enough and cannot be dismissed by arguments of artificiality.

Empirical evidence is abundant. Ariely *et al.* (1995) and Budescu *et al.* (Part II) describe empirical data (collected under carefully controlled conditions) that unambiguously exhibits genuine overconfidence. The overconfidence reported by Budescu *et al.* remains after adjustment for quantifiable error components and despite the adoption of a rather conservative and stringent criterion for determining overconfidence.

Budescu *et al.* (Part II) make an important point stating that well-calibrated subjects should use the ends of the extremes of the probability scale very rarely. In fact, the data they analyze (based on an experiment recently completed by Wallsten, Ariely, and Zauberman) clearly shows that subjects deviate from this prescription. Over 24% of the responses fall in extreme categories. Admittedly, regression effects at this point will have the largest impact, but the authors elegantly show that the amount of overconfidence is significantly larger than what would be expected by random error alone. Several other studies have shown a similar pattern in which approximately 20% of all judgments were made with 100% confidence and were only 80% correct (e.g., Dunning *et al.*, 1990; Fischhoff, Slovic, and Lichtenstein, 1977). These clear-cut results are unlikely to be exclusively accounted for by regression effects.

A particularly illuminating study in this context, done in a natural environment and with random sampling of stimuli, has been conducted with bridge players (Keren, 1987). Subjects were exposed to a natural environment of a bridge tournament (with relatively high financial incentives), and asked at each round to assess the likelihood that the final contract will be made. Each subject (player) had to make 28 probability judgments for each of the 28 games played. There were two groups of subjects: *expert* players with international experience, and *amateur* players. It is important to emphasize that the amateur players were not beginners; they all played in a bridge club and had an experience of playing thousand of bridge hands. There was a large and indisputable difference between the two groups: The group of experts exhibited superb calibration whereas the amateurs group was poorly calibrated and clearly showed overconfidence.

A major (though not the only) source of overconfidence exhibited by the amateur group was due to probability judgments of 1.0 (i.e., 100%). Apparently, in more than 29% of their probability judgments, amateur players had a confidence of 100%, for which only 80% of the events were true. In contrast, expert players rarely used a confidence of 100% (these responses constituted a mere 2.5% of all responses), and whenever they used it they were always(!) right.

What are possible reasons for the notable difference in calibration between the two groups? Obviously, the difference cannot be accounted for by arguments of an inadequate representative sample of stimuli since (1) the stimuli (i.e. bridge hands) were randomly chosen (and, unquestionably, from a very large population), and (2) both groups were exposed to exactly the same stimuli. Also, the argument that overconfidence is only applicable to non-frequentistic data does not hold. Even the group of amateur players (not to mention the experts) had previous experience playing thousands of hands, and thus these stimuli presented to them were a random sample from a well-defined reference set. Finally, the random error for the amateurs was indeed larger than that of the experts but certainly cannot account for all the differences. It is quite clear in this study that at least part of the overconfidence exhibited by the amateurs was genuine and caused by cognitive or motivational factors, a discussion of which follows below.

Additional evidence for substantial overconfidence is reported by Griffin and Tversky (1992) and Brenner *et al.* (1995). The studies reported in these two articles demonstrate unequivocal overconfidence that is not eliminated by random selection of items, is equally applicable to both subjective probability judgments and estimates of relative frequency, and as the authors show, cannot be treated merely as regression effects. In addition to the recent studies briefly cited above, which attempted to control for different possible pitfalls, dozens of studies have been conducted earlier which consistently show overconfidence and that at least have to be carefully reanalyzed before it is concluded that they are solely a result of artifacts.

The question that still has to be answered, and has received relatively little attention, concerns potential explanations for genuine overconfidence (it is in this respect that Lichtenstein, Fischhoff, and Phillips (1982) have

characterized the calibration research as 'dust-bowl empiricism'). There are at least several possibilities briefly summarized below.

One class of possible explanations is mainly cognitive in nature. For instance, Koriat, Lichtenstein, and Fischhoff (1980) propose a cognitive strategy by which subjects when evaluating a possible alternative tend to think more in terms of positive (rather than negative) reasons. Consequently, affirmative evidence looms larger than disconfirming evidence. For the limited domain of knowledge assessments (also referred to as the feeling of knowledge), several alternative accounts based on memory have been proposed (e.g. Koriat, 1994).

Dawes (1980) proposed that overestimation of our intellectual abilities is deeply rooted in humans and has occurred consistently since Plato. This overestimation, according to Dawes, has been further reinforced during the last few decades by the enormous scientific and technological progress of that period. Related to this hypothesis is a social norm in which knowledge is highly cherished and ignorance is deplored. Such a norm obviously encourages overconfidence, specifically in general knowledge questions, which have been the most popular in calibration research. The social norm explanation is not necessarily restricted to general knowledge questions. The overwhelming overconfidence exhibited by physicians (e.g. Christensen-Szalanski and Bushyhead, 1981) may be a result of social norms or social expectations. Most patients expect their physicians to provide them with clear diagnoses unqualified by probabilistic terms.

An alternative explanation is a motivational one. Specifically, overestimation may serve as a self-motivating mechanism (admittedly, not a long-lasting one). An example supporting such an account comes from a study by Babad (1987) who employed soccer fans as subjects and asked them to predict the outcome of a game shortly before it started. He had shown that the stronger a person felt affiliated with a team, the more likely this person was to predict a win by that team. Even predictions made at half-time, when the fan's favorite team trailed decisively, were characterized by a pervasive bias in favor of own team. The overconfidence exhibited by the bridge amateur players in the study by Keren (1987) may be accounted for by similar motivational factors. According to this explanation, one of the advantages of experts over amateurs is their ability to adopt a more realistic and sensible perspective leading to more accurate assessments. Other research has questioned the motivational explanation. Specifically, Yates, Lee, and Shinotsuka (1996) found that most people believe that motivation is a major basis for overconfidence, but the authors also note that the evidence for this assertion is rather scant. The issue remains unresolved because the amount of research addressing this issue has been so limited.

While each of these explanations may have an intuitive appeal, they are quite general and their confirmation would require converging evidence from different experiments. Testing such hypotheses within the traditional calibration paradigm may be difficult and would require more genuine methods. The point I am trying to make is that the dominant calibration paradigm has its focus on the psychophysical facets of probability judgments and has neglected the importance of alternative accounts for the overconfidence phenomenon.

PRELIMINARY CONCLUSIONS

Several articles have recently questioned whether the well-known phenomenon of overconfidence in calibration studies is more apparent than real. Specifically, some authors have cast doubt on the ecological validity of calibration studies, whereas others have proposed that the phenomenon may simply be a statistical artifact resulting from regression toward the mean.

A major argument related to the ecological validity issue concerns the random selection of items. Beside theoretical arguments (Kahneman and Tversky, 1996; Griffin and Varey, 1995), there are now sufficient empirical studies demonstrating overconfidence even when items were carefully sampled in a random manner. Being well calibrated and avoiding overconfidence is, normatively speaking, desirable regardless of how items are sampled. It is perhaps worth pointing out that most calibration studies (including those that advocate the ecological view!) have employed general knowledge questions. This is a highly artificial (not to mention boring) task from which generalizations may be at best limited. For those concerned about ecological validity, using different and more realistic tasks should be at least as important as issues regarding the sampling of stimuli.

There is also sufficient empirical evidence to dismiss the claim that overconfidence is entirely a statistical artifact. The articles in this issue to which this commentary is addressed may offer a useful framework for discerning statistical artifacts from genuine overconfidence, although practically separating real (i.e. systematic) from

apparent (random) error in practice remains a complex issue. There is an additional problem underlying the error models, namely, their initial supposition concerning the existence of a well-defined normative yardstick. There are two types of problem associated with this presumption. First, there is an inherent inconsistency in assessing subjective probabilities in frequentistic terms, a point which, as of yet, has not been adequately resolved. Second, as indicated above, the prevailing view encapsulated in calibration studies underscores the accuracy issue and undermines resolution. There is an intrinsic conflict between resolution and calibration (Keren, 1991) for which there is no normative resolution.⁴

The article by Wallsten, Budescu, Erev, and Diederich (this issue) contains an interesting proposal linking the above two problems. Specifically, their proposal is based on the distinction between aleatory (related to external chance factors) and epistemic (related to internal assessment of knowledge) probabilities. They propose that accuracy is dominant in the context of aleatory uncertainty whereas resolution (diagnosticity) is the dominant factor in the context of epistemological uncertainty. Future empirical research should further determine the (descriptive) validity of this conjecture. Although the distinction between aleatory and epistemological uncertainty may be well-defined, in practice it may often be difficult to operationally distinguish these two types. The duality problem (Hacking, 1975) has never been really resolved.

Finally, I turn to an additional question that has been largely overlooked until recently. Do people (as generators and users of probabilistic information) accept the dictates of the normative framework? The question I am raising is not whether people can follow these dictates but rather whether they accept them in the first place. In other words, do subjects and researchers of calibration studies share the same mental model regarding the assessment and evaluation of probability statements? In order to describe how people make probability judgments and where and why they depart from the normatively correct yardstick, one should first study and assess the standards they employ (regardless of whether they correspond to the normative standards). I briefly address this question in the following final section.

WHAT IS CALIBRATION ALL ABOUT?

Given the vast number of studies on calibration, one may pause and wonder what have we learned from this massive line of research. In closing their review article on calibration of probabilities, Lichtenstein, Fischhoff, and Phillips (1982) observed that much of the research on calibration is 'dust bowl empiricism', and in a subsequent review 10 years later (Keren, 1991) I reiterated a similar conclusion. The amount of progress achieved since then remains, in my opinion, questionable.

Calibration studies are mainly psychophysical,⁵ in nature. In this framework, people are envisaged as instruments of measuring uncertainty in which the accuracy of the instruments is central and measured against a normative yardstick. Not only is the proper normative yardstick controversial (as mentioned above), but it is also doubtful whether subjects (serving as the measurement instrument) indeed adopt the same normative perspective (and corresponding assumptions) as do calibration researchers. For instance, one may wonder if the excessive emphasis on calibration, specifically the accuracy dimension, is indeed warranted.

There are several reasons for raising some doubts. First, it is highly questionable whether people really regard calibration as being the essence of probability statements. In a set of innovative studies, Yates *et al.* (1996) have examined what they termed the 'consumer's' view and addressed the question of how those who receive and use probabilistic information appraise these assessments. The most noteworthy result from these studies was the finding that people consider resolution to be much more important than calibration. Indeed, if probability judgments are conceived to be carriers of information, then it follows that resolution is at least as important as calibration. Apparently, most of the articles in this issue (in accordance with most of the calibration literature) treat resolution as either a secondary issue or are silent about it.

⁴ 'Real' experts should be characterized by both good calibration and good resolution (discrimination). It is still questionable whether these two attributes should be conceived as merely dimensions of an overall accuracy construct or whether they imply different types of expertise.

⁵ One of the first occasions in which a calibration task was employed was associated with a visual psychophysical discrimination task (Swets, Tanner, and Birdsall, 1961) in which subjects had to indicate on each trial their confidence in the correctness of their response.

Evidently, there is even more to probability judgment than just calibration and resolution. In a recent (as yet unpublished) study conducted by Karl Teigen and myself, we posed to subjects the following question:

Below you will find weather forecasts made by two meteorologists, which we refer to as A and B, regarding the probability of a rainy (sunny) day. The forecasts were made with regard to 4 different days. As it turned out, it was raining (sunny) on 3 out of the 4 days. Please examine carefully their forecasts and tell us who in your opinion was a better forecaster: A or B? There is no 'correct' answer and we just want to get your opinion.

Forecaster A	Forecaster B
90	75
90	75
90	75
90	75

Following normative considerations, the forecasts of both meteorologists portray the same resolution (namely 0) and B is perfectly calibrated whereas A is clearly miscalibrated. Nonetheless, 25 out of 40 subjects (62.5%) opted for forecaster A. Asked for their reasons, many subjects who preferred forecaster A have noted that forecaster A made clear (and conclusive) predictions and in 3 out of the 4 days on which he was indeed right. In contrast, they claimed, none of the predictions made by forecaster B was sufficiently decisive.⁶ Evidently, as in the Yates *et al.* (1996) study, subjects seem to overweigh correct extreme predictions.

Reflecting on the recent calibration literature in general, and specifically on the articles in this special issue, leads me to two suggestive conclusions. First, students of calibration of probabilities have excessively accentuated the normative facets of probability theory neglecting the layperson's conception of what probability judgments are about and how they should be assessed. In the Yates *et al.* terminology, the perspective of lay people as 'consumers' of probability judgments has been largely neglected. Second, and strongly related, calibration studies have employed a too-narrow psychophysical approach, overlooking other important aspects of how people cope with uncertainty. The psychophysical analysis of probability judgments has indeed an irresistible attraction: It enables easy and fast data collection (what is simpler than asking subjects general knowledge questions and corresponding confidence ratings?), and (despite the somewhat complex analyses that have been developed) apparent simple interpretation. However, the psychophysical account for probability judgments comes at the expense of other ways of understanding how people interpret, use, and assess probability statements.⁷ Not undermining the insights obtained from calibration studies, perhaps we have exhausted its potential and should consider other new directions. The studies conducted by Yates *et al.* (1996) and those reported in the articles by Koehler and Harvey and by Yaniv and Schul (this issue) may offer some possible alternatives for new directions.

REFERENCES

- Ariely, D., Zauberman, G. and Wallsten, T. S. 'Statistical and psychological considerations in calibration research', Paper presented at SPUDUM-15, August 1995, Jerusalem, Israel.
- Babad, E. 'Wishful thinking and objectivity among sport fans', *Social Behavior: An international Journal of Applied Social Psychology*, **4** (1987), 231–40.
- Baranski, J. V. and Petrusic, W. M. 'The calibration and resolution of confidence in perceptual judgments', *Perception and Psychophysics*, **55** (1994), 412–28.
- Björkman, M. 'Internal cue theory: Calibration and resolution of confidence in general knowledge', *Organizational Behavior and Human Decision Processes*, **58** (1994), 386–405.
- Björkman, M., Juslin, P. and Winman, A. 'Realism of confidence in sensory discrimination: The underconfidence phenomenon', *Perception & Psychophysics*, **54** (1993), 75–81.

⁶ Implicit in subjects' reasoning is the idea that a probability for rain of 0.9 is so close to 1.0 that the decision to take an umbrella is unequivocal. In retrospect, that decision was indeed justified in three out of the four cases and the single error is thus underweighted. In contrast, a probability of 0.75 is not sufficiently high for a definitive decision and thus in each case creates ambivalence. The 'average' accuracy of the forecasts seems to be of little importance.

⁷ My comments regarding the psychophysical nature of calibration studies and the associated corresponding limitations draw heavily on similar comments made by Daniel Kahneman (unpublished paper) with regard to the broader field of decision making.

- Brenner, L. A., Koehler, D. J., Liberman, V. and Tversky, A. 'Overconfidence in probability and frequency judgments', *Organizational Behavior and Human Decision Processes*, **65** (1995), 212–20.
- Brier, G. W. 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, **78** (1950), 1–3.
- Christensen-Szalanski, J. J. and Bushyhead, J. B. 'Physicians' use of probabilistic information in a real clinical setting', *Journal of Experimental Psychology: Human Perception and Performance*, **7** (1981), 928–35.
- Dawes, R. M. 'Confidence in intellectual judgments vs. confidence in perceptual judgments', in Lantermann, E. D., and Feger, H. (eds), *Similarity and Choice*, Stuttgart: Hans Huber Publishers, 1980.
- Dunning, D., Griffin, D. W., Milojkovich, J. and Ross, L. 'The overconfidence effect in social prediction', *Journal of Personality and Social Psychology*, **58** (1990), 568–81.
- 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., Slovic, P. and Lichtenstein, S. 'Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, **3** (1977), 552–564.
- Galton, F. 'Typical laws of heredity', *Nature*, **15** (1877), 492–5, 512–14, 532–3.
- Gigerenzer, G. 'How to make cognitive illusions disappear: Beyond "Heuristics and Biases"', *European Review of Social Psychology*, **2** (1991), 83–115.
- Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. 'Probabilistic mental models: A Brunswikian theory of confidence', *Psychological Review*, **98** (1991), 506–28.
- Griffin, D. and Tversky, A. 'The weighing of evidence and the determinants of confidence', *Cognitive Psychology*, **24** (1992), 411–35.
- Griffin, D. and Varey, C. 'Toward a consensus on overconfidence', *Organizational Behavior and Human Decision Processes*, **65** (1996) 227–32.
- Guttman, L. 'What is not what in statistics?' *The Statistician*, **26** (1977), 81–105.
- Hacking, I. *The Emergence of Probability*, Cambridge: Cambridge University Press, 1975.
- Harrison, J. R. and March, J. G. 'Decision making and post decision surprises', *Administrative Science Quarterly*, **29** (1984), 26–42.
- Hogarth, R. 'Generalization in decision research: The role of formal models', *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-16**, 439–49.
- Juslin, P. 'An explanation of the hard–easy effect in studies of realism of confidence in one's general knowledge', *European Journal of Cognitive Psychology*, **5** (1993), 55–71.
- 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.
- Kahneman, D. and Tversky, A. 'On the reality of cognitive Illusions', *Psychological Review*, **103** (1996), 582–91.
- Keren, G. 'Facing uncertainty in the game of bridge: A calibration study', *Organizational Behavior and Human Decision Processes*, **39** (1987), 98–114.
- Keren, G. 'On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies', *Acta Psychologica*, **67** (1988), 95–119.
- Keren, G. 'Calibration and probability judgments: Conceptual and methodological issues', *Acta Psychologica*, **77** (1991), 217–73.
- Keren, G. and Bolhuis, J. van 'On the ecological validity of calibration studies', unpublished working paper, 1995.
- Kerlinger, F. *Foundations of Behavioral Research* (3rd edn), New York: Holt, Rinehart & Wilson, 1986.
- Koriat, A. How do we know that we know? The accessibility model of the feeling of knowing', *Psychological Review*, **100** (1994), 609–39.
- Koriat, A., Lichtenstein, S. and Fischhoff, B. 'Reasons for confidence', *Journal of Experimental Psychology: Human Learning and Memory*, **6** (1980), 107–18.
- Lichtenstein, S., Fischhoff, B. and Phillips, L. D. 'Calibration of probabilities: The state of the art to 1980', in Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, 1982.
- Murphy, A. H. 'A new vector partition of the probability score', *Journal of Applied Meteorology*, **12** (1973), 595–600.
- Peterson, D. K. and Pitz, G. F. Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **14** (1988), 85–92.
- Pfeifer, E. P. 'Are we overconfident in the belief that probability forecasters are overconfident?' *Organizational Behavior and Human Decision Processes*, **58** (1994), 203–13.
- Ronis, D. L. and Yates, J. F. 'Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method', *Organizational Behavior and Human Decision Processes*, **40** (1987), 193–218.

- Soll, J. B. 'Determinants of overconfidence and miscalibration: The roles of random error and ecological structure', *Organizational Behavior and Human Decision Processes*, **65** (1996), 117–137.
- Suantak, L., Bolger, F. and Ferrell, W. R. 'The hard–easy effect in subjective probability calibration', *Organizational Behavior and Human Decision Processes*, **67** (1996), 201–21.
- Swets, J. A., Tanner, W. P. and Birdsall, T. 'Decision processes in perception', *Psychological Review*, **68** (1961), 301–40.
- Tversky, A. 'Features of similarity', *Psychological Review*, **84** (1977), 327–52.
- Wallsten, T. S., Budescu, D. V., Rapaport, A., Zwick, R. and Forsyth, B. 'Measuring the vague meanings of probability terms', *Journal of Experimental Psychology: General*, **115** (1986), 348–65.
- Yaniv, I. and Foster, D. P. 'Graininess of judgment under uncertainty: An accuracy–informativeness tradeoff', *Journal of Experimental Psychology: General*, **124** (1995), 424–32.
- Yates, J. F. 'External correspondence: Decomposition of the mean probability score', *Organizational Behavior and Human Performance*, **30** (1982), 132–56.
- Yates, J. F. 'Analyzing the accuracy of probability judgments for multiple events: An extension of the covariance decomposition', *Organizational Behavior and Human Decision Processes*, **41** (1988), 281–99.
- Yates, J. F., Lee, J. and Shinotsuka, H. 'Beliefs about overconfidence, including its cross-national variation', *Organizational Behavior and Human Decision Processes*, **65** (1996), 138–47.
- Yates, J. F., Price, P. C., Lee, J. and Ramirez, J. 'Good probabilistic forecasts: The “consumer’s” perspective', *International Journal of Forecasting*, **12** (1996), 41–56.

Author's biography:

Gideon Keren is professor of psychology at the Department of Technology Management of the Eindhoven University of Technology. His research interests include probabilistic judgments and forecasts, choice behavior, behavioral game theory, and methodology in the social sciences.

Author's address:

Gideon Keren, Eindhoven University of Technology, POB 513, TEMA building 0.38, 5600 MB Eindhoven, Netherlands.