PMM theory, overconfidence, and representative sampling of items: A review of data

Arndt Bröder
Zusammenfassung

PMM theory and overconfidence: A review of data

Abstract

So-called “ecological models”, for example the theory of Probabilistic Mental Models (Gigerenzer, Hoffrage & Kleinbölting, 1991) assume that the phenomenon of “overconfidence” is not a cognitive bias per se, but a result of a biased selection of items in typical calibration studies. These models predict the elimination of overconfidence when items are representative for a defined reference class. Furthermore, the well-documented “hard-easy-effect” is expected to be eliminated by representative sampling of items. Studies examining these predictions often use random sampling of items to achieve the goal of representative item pools. Some studies and their results are described in this paper. Results are equivocal: Some studies seem to support ecological models, whereas others do not. The apparently supporting results can possibly be explained by the confounding of sampling procedure with item difficulty. Therefore, it is argued that at least three conditions must be fulfilled in further studies to reach at definitive conclusions concerning ecological models: (1) a “manipulation check” or rational conventions have to be established to test, whether a “representative” set of items was used in a specific study, (2) the effects of difficulty and sampling procedure have to be disentangled, and (3) unsystematic response errors that can mimic apparent overconfidence bias have to be explicitly incorporated into the models and assessment procedures.
PMM theory and overconfidence: A review of data

Contents

1. Introduction: Overconfidence .......................... 1

2. PMM theory and other ecological models .......... 2
   2.1 Basic concepts and processing assumptions 3
   2.2 Predictions of PMM theory .................... 5
   2.3 Similar and extended ecological models .... 6

3. A review of studies with representative sampling 8
   3.1 Studies using random sampling of items 8
   3.2 Studies not included in the review .......... 14

4. Discussion ............................................. 15

5. References ............................................ 18
1. Introduction: Overconfidence

A long tradition in the analysis of human judgment and decision making has been concerned with a multitude of “cognitive biases” that appear to demonstrate systematic deviations of human judgment from so-called rational norms or normative models. A cornerstone of this research programme is the book *Judgment under uncertainty: Heuristics and Biases*, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky who demonstrated apparently irrational biases such as the “conjunction fallacy” (Tversky & Kahneman, 1982a), the “hindsight bias” (Fischhoff & Beyth, 1975), the “base rate neglect” (Tversky & Kahneman, 1982b) among numerous others that gained some attention outside the psychological laboratories. Another stable phenomenon in the literature is “overconfidence” that can be demonstrated by comparing subjective confidence judgments of having solved an item correctly to the actual percentage of correctly solved items (Lichtenstein, Fischhoff & Phillipps, 1982; see also Keren, 1991, for a review). Usually, the subjective ratings exceed the actual percentage correct. This overestimation of one’s decisional accuracy has been interpreted as a self-serving motivational bias or as a cognitive “confirmation bias” (Koriat, Lichtenstein & Fischhoff, 1980). Whether one favors the motivational or the cognitive interpretation does not really change the basic conclusion: people are biased in a systematic and “obviously irrational” way.

However, in recent years this conclusion has been challenged by researchers who favor the notion of “bounded rationality” as opposed to “irrationality” (see Gigerenzer, 1991, 1993, 1996, 1997). According to this “ecological view”, people are well adapted to their environment and adopt decision strategies that are effective (albeit simple) within this environment. Somewhat simplified, this view states that the so-called cognitive biases are no fundamental errors in judgment, but merely an artifact of laboratory settings that systematically distort the environmental structure. As McClelland and Bolger (1994) put it: “The pessimists believe that biases are in people – the optimists believe that biases are in research” (p. 455). The battle between pessimists and optimists is partly documented in Gigerenzer (1996) and Kahneman and Tversky (1996).

Another stable phenomenon connected to the calibration of subjective probabilities has been termed the *hard-easy-effect* (Lichtenstein & Fischhoff, 1977). That is, overconfidence is most pronounced when to-be-judged items are hard in terms of a low solution probability. With easy items, good calibration or even underconfidence can be observed. Overconfidence and hard-easy-effect have been demonstrated in experts and laypersons, even if some exceptions exist that will be discussed later (see Fischhoff & Beyth, 1975; Keren, 1987; Murphy & Winkler, 1984). Despite the long tradition of calibration studies, Lichtenstein et al. (1982) have criticized the “dust-bowl empiricism” (p.333) which
was confined to demonstrating the phenomena in an increasing number of settings without stimulat- ing the formulation of precise process theories. Gigerenzer, Hoffrage and Kleinbölting (1991) state that “these robust phenomena still await a theory” (p. 506). They also criticize that the models presented before (some of which are summarized in McClelland & Bolger, 1994) do not account for all observed data patterns or do not show an integrative theoretical perspective. Their theory of probabilistic mental models (PMM theory) was designed to fill this gap. PMM theory is an ecological model which attributes apparent overconfidence to a biased selection of items and predicts the elimination of overconfidence as well as some other effects when this selection bias is eliminated through “representative sampling” of items from appropriate reference classes (the terms will be defined below). Very similar models and predictions have been proposed by Justlin (1993; 1994) and Björkman (1994).

In this paper, I will concentrate on predictions of PMM theory concerning the representative sampling of items and the predicted elimination of overconfidence. First, PMM theory and its predictions will be described. Second, several studies will be reviewed that realized representative sampling and assessed whether overconfidence was still observed. Not all studies claiming to have realized a representative sampling process will be discussed, however, because some failed to realize random sampling despite claiming so. The review will show that some studies were successful in eliminating overconfidence while others were not. It will be argued that the overall reduction of the magnitude of overconfidence due to representative sampling is no argument in support of ecological models, even if this might appear to be the case at first glance. In a final discussion section, I will try to figure out possible critical factors of “successful” studies and it will be argued that an “attenuated” version of these ecological models including some unsystematic error variance might be fruitful in further investigations that try to find out whether overconfidence is “real” or merely “artifactual”.

2. PMM theory and other ecological models

In a typical calibration study, participants are often confronted with general knowledge almanac items of the type “Who was born first: Aristotle or Buddha?”. After guessing the answer, a numerical confidence judgment is assigned to that item which is the subjective probability of having chosen the correct alternative. When participants have completed a large list of such items, the data are grouped according to the confidence responses, and for each confidence category \( x_j \) the relative frequency of items answered correctly \( p(C)_j \) is determined. Judgments are said to be well calibrated if the \( p(C)_j \) correspond well to the
For illustrative purposes, data are often presented in a calibration curve which is a graph depicting the \( p(C) \) for all \( x_i \). Good calibration is represented by a curve lying roughly on the diagonal, whereas overconfidence (underconfidence) is demonstrated by curves lying below (above) the diagonal. Several indices to quantify calibration have been developed, some of which are summarized in Yates (1990). The easiest measure, however, is the difference between mean confidence and mean percentage correct. This index (hereinafter CAL) can be expressed in percentage points. Negative values indicate underconfidence while positive values indicate overconfidence.

Most models trying to explain the common findings of overconfidence and the hard-easy-effect concentrate on the processes that generate the confidence judgment. PMM theory is an integrative model covering the inference process (choosing an answer) and the confidence judgment, and hence, McClelland and Bolger (1994) evaluate PMM theory to be “the most complete model for the calibration of subjective probabilities that has so far been produced.” (p. 470). First, the main concepts of PMM theory will be outlined, second the specific predictions will be presented, and third some extended versions of ecological models will be introduced briefly.

2.1 Basic concepts and processing assumptions

The authors of PMM theory see themselves in the tradition of Egon Brunswik’s probabilistic functionalism (Brunswik, 1955; see Gigerenzer et al., 1991). According to Brunswik’s view, inferences concerning distal objects have to be drawn from proximal cues that are available to the observer. These cues are statistically related to the to-be-judged distal object and thus bear some predictive power that is called their “ecological validity”. One of Brunswik’s assumptions is that people are “well-adjusted” or adapted to the environmental structure and therefore efficient in utilizing these cues for inferences (Brunswik, 1956). This assumption is explicitly incorporated to PMM theory by Gigerenzer et al. (1991). From this assumption follows Brunswik’s claim for representative designs in psychology: In order to investigate cognitive functions, the stimulus situation must be representative for the environment which the human mind is adapted to. Otherwise, our picture of these processes will be flawed and artifactual (Brunswik, 1943).

According to PMM theory, a typical two-alternative choice item like “Which city has more inhabitants: Heidelberg or Bonn?” in most cases requires an inductive inference, i.e. a “best guess” (In the case of some knowledge that allows for a deduction of the correct answer, Gigerenzer et al. refer to a Local Mental Model instead of a Probabilistic Mental Model). Gigerenzer et al. (1991) assume that best guesses of course are no random guesses
which is indeed plausible because a species drawing only random inferences in case of lacking definite knowledge would presumably not have been very successful. According to PMM theory, a best guess can be made by activating several probability cues in memory. These are features covarying with the target variable (number of inhabitants in the example) which has to be judged. In the Heidelberg-Bonn-case these cues might be pieces of information one has about both cities (e.g. whether they have an international airport or not). The cues have a certain predictive value concerning the target variable which is called their ecological validity. In PMM theory, the cue validity is defined as the conditional probability of drawing a correct inference when one of the to-be-chosen objects possesses the critical cue feature (e.g. an airport) whereas the other object does not. Other formal representations of cue validity are also possible (see Martignon & Hoffrage, in press). This validity is defined relative to a certain reference class of objects (e.g. German cities) that can be compared with regard to the target variable.

In PMM theory, it is assumed that repeated exposition to reference classes has established an internal representation of ecological cue validities that corresponds well to the ecological validities. This assumption resembles Brunswik’s notion of a “cue family hierarchy” that is a well-adjusted representation of ecological validities. When confronted with a probabilistic inference task, the cue values have to be integrated to arrive at a decision and a confidence judgment. According to PMM theory, this cue integration follows a simple heuristic that has been labeled “Take The Best” and is a special case of the lexicographic rule (see Fishburn, 1974; Svenson, 1979) formulated for binary choices with dichotomous cues. The Take-The-Best-heuristic is a simple cognitive model: The participant will activate the most valid cue from memory and test whether this cue discriminates between the objects (e.g. one city has an international airport, the other has not). If the cue discriminates, the object favored by the cue is chosen (the city having an airport), and search for further information is terminated. If the cue does not discriminate, the next best cue is activated and so on. As a confidence judgment, participants will specify the validity of the cue that determined the choice. Despite being simple, this apparently irrational heuristic which deliberately ignores information is surprisingly successful in “natural” environments as extensive simulation studies have shown (Gigerenzer & Goldstein, 1996; Czerlinski, Gigerenzer & Goldstein, in press). This heuristic and its descriptive adequacy as a cognitive model of human inference have been discussed elsewhere (Bröder, 1999).

Altogether, the formulation of PMM theory presented above accounts for the cognitive processes of a particular choice and the confidence judgment that is just a translation of the internal representation of cue validity into a numerical confidence judgment.
2.2 Predictions of PMM theory

The predictions derived from these processing assumptions are clearcut and identical to predictions from similar models as presented by Justlin (1993, 1994) and Björkman (1994). Why does overconfidence occur in typical calibration studies? Gigerenzer et al. (1991) argue that in typical studies items are selected for difficulty (see also Justlin, 1993). That is, no representative sample of items is drawn from a reference class. This will lead to a situation where “misleading” items are overrepresented in the item sample. The sample cue validities are lower than the cue validities in the reference class. That is, cues will lead to a lesser percentage of correct inferences in the sample than in the reference class. The confidence judgments, however, are based on the reference class validities, and hence, they appear to be overconfident because in the item sample the percentage correct will be less than those judgments. All predictions of PMM theory are variations of “representative item sampling”.

Prediction 1: Elimination of overconfidence with representative sampling

If the generation of items leads to an item sample which maintains the cue validities of the reference class (representative sampling), then the percentage correct as well as the confidence judgments based on the cues will coincide and thus, no overconfidence will be observed.

Prediction 2: Elimination of the Hard-easy-effect

The difficulty of items in terms of their solution probability does not per se enhance overconfidence. More important is their difficulty relative to the corresponding reference class. When items are sampled in a representative way from two reference classes of varying difficulty, no overconfidence will be observed with any of the item samples, regardless of the different solution probabilities. That is, in the harder item sample, decisions will be made on the basis of less valid cues (and therefore result in less correct inferences), but this will be accompanied by lower confidence judgments.

Prediction 3: Reversal of the Hard-easy-effect

Applying the same logic as in prediction 2, the hard-easy-effect can even be reversed. If “hard” items are selectively sampled from an easy reference class while representative items are sampled from a hard reference class, overconfidence is to be expected for the former items while good calibration is expected for the latter ones. This is expected to be the case even if the latter items are harder than the former ones in terms of solution probability.

Prediction 4: Confidence-frequency-effect

Another prediction which is not central to this paper is concerned with the comparison of confidence judgments with the estimated number of correctly solved items (frequent judgments). Gigerenzer et al. (1991) assume that both kinds of judgments are based on different reference classes, and therefore, these judgments are expected to differ in a
systematic way. The reference class for frequency judgments might be something like “general knowledge items” that are usually perceived as difficult. Hence, in selected item sets, frequency judgments are expected to be well calibrated, while confidence judgments should show the normal overconfidence. In representative sets, however, frequency judgments are expected to be underconfident. In addition, this confidence-frequency-discrepancy is expected to be reduced when the sample of items to be judged in a frequency judgment becomes smaller. In this paper, I will not deal with this confidence-frequency effect. Suffice it to say that it has repeatedly been shown that frequency judgments are more conservative than confidence judgments (Brener, Koehler, Liberman & Tversky, 1996; Schneider, 1995; Treadwell & Nelson, 1996). However, some easy alternative explanations are possible to account for this phenomenon. Furthermore, the shrinking discrepancy between confidence and frequency judgments when item sets get smaller that was demonstrated by Gigrenzer et al. (1991, Exp. 2) could not be replicated in two studies by Treadwell and Nelson (1996).

With the exception of prediction 4, all predictions concern the representative sampling of items which is expected to eliminate an apparent “cognitive bias” which is seen as an artifact of “unfair” item selection in typical calibration studies.

2.3 Similar and extended ecological models

Similar ideas have been developed by Juslin (1993, 1994) and Björkman (1994). However, as Björkman (1994) points out, the PMM assumption of a Take-The-Best-heuristic is not necessary to derive the predictions outlined above. In his Internal Cue Theory Björkman does not formulate strong assumptions concerning the actual inference process. To derive the predictions it is only necessary to assume the “(a) correspondence between ecological and internal validity, (b) perfect translation of internal validity into a confidence assessment, and (c) consistent utilization of cues.” (p. 386). Any efficient heuristic or decision process based on well-adapted cue-utilization will therefore be sufficient to predict the abovementioned effects.

The predictions outlined above seem straightforward and easy to test. However, three problems must be mentioned with respect to the empirical examinations intended to test the assertions. First, it is not obvious how a “representative” sample of items from a reference class should be generated. In most cases, a random sample of items is drawn from a prespecified reference class (e.g. the world countries, see Juslin, 1993). A conceptual problem with this procedure has been mentioned by Keren (1997) who notes that the exposition of people to “natural” reference classes need not be “random”. That is, even the learning environment in which cue validities were acquired might be systematically biased in
favor of some objects. Thus, a random sample from a reference class need not necessarily maintain the cue validities that were learned by the participant beforehand. Second, all cases of sampling from a population entail the possibility of sampling error, so there is no guarantee that reference class validities are actually maintained in the sample. I know of no experiment in which a manipulation check has been performed in order to ensure that the sample was really representative of the reference class in that respect. However, unless it is not clear whether this prerequisite was achieved in a specific experiment, it is hard to interpret the data. There are no generally accepted criteria that allow to decide whether a sample was truly representative in a specific experiment, even if this problem has been acknowledged by some authors (Juslin, Olsson & Winman, 1998). Unless these criteria are established, one has to rely on the rule of thumb that large item samples are more trustworthy than small ones. Third, many authors have argued that unsystematic error variance in the judgments can contribute to an apparent overconfidence effect due to the phenomenon of regression towards the mean (Budescu, Erev & Wallsten, 1997; Budescu, Wallsten & Au, 1997; Erev, Wallsten & Budescu, 1994; Juslin, Olsson & Björkman, 1997; Juslin, Olsson & Winman, 1998; Soll, 1996). Two sources of this error variance can be separated: One part of unsystematic error might stem from sampling errors in learning the reference class (Juslin et al., 1997; Soll, 1996), that is, the learned cue validities are based on an error-prone sample of the reference class and might therefore differ from reference class validities. Another source is the possibility of unreliable confidence judgments. It has repeatedly been shown by simulations that these unsystematic errors can mimic an apparent overconfidence bias despite unbiased “true” judgments. The consequences for testing the predictions of ecological models are not yet clear. If a representative sample of items has been drawn and overconfidence is observed in the data, this apparent bias might be due to unsystematic error variance. This data pattern could be consistent with a weaker version of the ecological models that allow for unreliable judgments. To evaluate the impact of random error, quantitative estimates of this error variance must be available and appropriately “partialled out” from the apparent bias in order to assess whether there is a “true” bias in addition to the effect due to unsystematic error. The only model I know of which explicitly makes an effort to achieve this goal was proposed by Budescu, Erev & Wallsten (1997).

Obviously, as long as the three problems mentioned above are not solved (or at least somehow settled by convention), studies testing the implications of “representative sampling” must be interpreted with caution with respect to PMM theory or other ecological models. Despite these hurdles, in the following section some calibration studies with representative sampling are reviewed in order to get a first impression of the empirical status of PMM theory and other ecological models.
3. A review of studies with representative sampling

Some calibration studies have tested PMM theory’s claim that good calibration is to be expected when items are sampled in a representative manner from a reference class. Other studies that used representative sampling without explicit reference to PMM theory or other ecological models of calibration will be reviewed, too. Some studies claim to have used random sampling, but in fact they did not (Suantak, Bolger & Ferrell, 1996, Experiment 2; Soll, 1996) or the item samples were so small that they are probably contaminated by a large sampling error (Griffin & Tversky, 1992, Study 5). In their experiment 2, Suantak et al. (1996) asked questions about mortality rates of several diseases and accidents. However, in every trial, one disease with a high mortality rate had to be compared to one with a low mortality rate. Hence, the selection of items from the reference class was not random (For a detailed discussion, see Justin et al., 1998). In Soll’s (1996) study, the items reflected “every possible configuration of cues” (p.128), regardless of their base rates in the reference class. Therefore, his item sample cannot be considered representative for the ecological reference class (American cities), either. These problems will be reported along with the studies. As has been noted above, no criteria exist to make sure that random sampling actually resulted in a representative item sample. Two aspects allow for a crude evaluation of this question: First, sampling error, of course, should be less severe the larger the to-be-judged item samples are. Second, studies with random sampling of items for each participant are expected to be less sensitive with respect to sampling error than studies, in which one item sample was generated and administered to all participants. The latter aspect is not always clear from the experimental descriptions.

For every study reviewed, the mean overconfidence measure CAL (as defined in section 2) will be reported if available from the description of the data. When this is not the case, a qualitative description of the data will be given. Every description of a study will be accompanied by a short comment.

3.1 Studies with random sampling of items

Gigerenzer, Hoffrage and Kleinbölting (1991)

Experiment 1: 80 participants answered 350 items each and specified confidence judgments for every item. The 350 items consisted of a “representative” set of 300 items and a “selected” set of 50 items. In the selected set, typical almanac items were used (“Who was born first: Aristotle or Buddha?”). The representative set was constructed as follows: 25 of the 65 West German cities with more than 100,000 inhabitants were selected at random. All 300 paired comparisons between these cities were judged by every participant. The target
variable was the number of inhabitants of the cities. The items were the same for all participants. The CAL score was 13.8% for the selected set, whereas the score was -0.9% for the representative set, thus showing the usual overconfidence for almanac items and near perfect calibration for the randomly chosen city comparisons. However, the selected items were much harder than the city items (52.9% vs. 71.7% correct).

Experiment 2: 97 participants judged 260 items each, 50 of which were the almanac items of experiment 1. The representative set was constructed the same way as in experiment 1, but only 21 cities were selected to construct the 210 paired comparisons. The items were the same for all participants. Overconfidence was 15.4% in the selected set and only 2.8% in the representative set. Again, the difference in difficulty was 56.2% vs. 75.3%.

In both studies, the calibration of confidence judgments was near to perfect in the random sets, whereas overconfidence was demonstrated for the representative set, thus supporting PMM theory.

Juslin (1993)
20 participants judged 240 items each, 120 of which were randomly sampled and 120 of which were selected for difficulty by 12 selectors. The random sample was constructed as follows: In every trial, a pair of the 179 world countries was drawn randomly for each participant anew. Participants were asked to compare the countries on one of six target variables (population density, life expectancy etc.). The mean overconfidence score for the random sample was -0.1% which “is among the lowest ever reported for almanac questions” (Juslin, 1993, p. 64/65).

Juslin (1994)
20 participants judged 120 items similar to the representative set in Juslin (1993). However, this time all participants received the same list of items. The judgments showed slight but significant underconfidence (-4.0%). For three additional studies, no experimental details are reported, but the random samples yielded underconfidence values of -4.0%, -1.6% and -1.6%, respectively. Hence, the randomly selected items tended to produce slight underconfidence.

Hoffrage (1995)
Experiment 3: 56 participants judged 200 items each. The target variable was city population. 100 items were comparisons of randomly drawn American cities while the remaining 100 items were comparisons of German cities. Items were identical for all participants. The rationale behind the comparison of two reference classes was to show the elimination of the hard-easy-effect when items are randomly drawn from reference classes differing in
difficulty, assuming that American city items are harder for German students than German city items. However, the manipulation failed because the percentage correct was slightly higher for American cities (76.04% vs. 75.67%). There was moderate underconfidence for the American set (-3.7%) and moderate overconfidence for the German set (+3.8%).

Experiment 4: 60 participants judged 100 items. For half of the participants, items were selected from the 75 most populous American cities, for the other participants, items were sampled from the 32 most populous cities of the USA. Calibration was -1.6% and +12.1%, respectively. Thus, the latter set produced an overconfidence effect of a size usually observed in typical calibration studies with selected item sampling. Obviously, the random generation did not have the intended effect in this set.

Experiment 5: Because the intended manipulation of difficulty did not work in Experiment 3, an attempt was made to replicate the study with other material. 100 participants judged 100 items each. Items were comparisons of “randomly selected famous persons”. The target variables were the age at death (hard) and the date of birth (easy). The manipulation was successful (57.09% correct vs. 73.49% correct). However, slight overconfidence was produced in both conditions (5.18% and 3.34%, respectively). Contrary to the prediction, the hard-easy-effect was significant (although small).

The pattern of results reported by Hoffrage (1995) is not very convincing in favor of PMM theory. The prediction of an eliminated overconfidence and hard-easy-effect through random sampling was not confirmed. However, it is unclear why two conditions produced slight underconfidence while all others produced overconfidence. Taken together, Hoffrage’s data pattern is consistent with a post hoc explanation which assumes that confidence judgments are familiarity-based and therefore “metacognitive”. However, “familiarity” is not a good predictor of actual success. German participants should judge American cities to be relatively unfamiliar, while German cities are relatively familiar. Hence, confidence judgments are lower for the former than for the latter, producing under/overconfidence when in fact success is equally good in both conditions (Experiment 3). In Experiment 4, most of the 32 most populous American cities sound familiar while a lot of the 75 most populous cities sound unfamiliar (e.g. Tampa, Anaheim, Shreveport). When confidence judgments are based on item familiarity, the results of Experiment 4 were to be expected.

Justlin, Olsson & Winman (1998)

Justlin et al. (1998, Appendix) summarize 44 data sets from their own laboratory in which items were randomly sampled from a wide variety of reference classes (some of which are discussed above: Justlin, 1993, 1994). Although details of the studies are not supplied, the overall evaluation is impressive: Mean overconfidence across all studies is only +1.0%. However, there is considerable variation between studies with some showing...
underconfidence and others overconfidence. Whether this variation is due to sampling error in every study (cancelling out across studies) or whether it is due to some other uncontrolled factors cannot be answered. However, the reviewed studies disclose a pattern which is very critical for ecological models, contradicting prediction 2 of PMM theory (see above): Across the 44 data sets, the correlation between the calibration measure and percentage correct turns out to be $r = -0.80$ which demonstrates a clear hard-easy-effect.\footnote{The value $r = -0.7998$ was computed from the Table in the appendix of Justlin et al. (1998). Justlin et al. admit that there is a correlation, but they do not report its magnitude because “The magnitude of the latter effect is difficult to estimate, though, due to the linear dependency between proportion correct and over-/underconfidence.” (p. 20). However, if this argument were taken seriously, it would be meaningless to speak of a hard-easy-effect, anyway! When overconfidence is defined as the difference between mean confidence judgments and percentage correct, the question whether they correlate remains an empirical one, and of course, there is no formal implication of a correlation between both variables. Another question is, whether this index or definition is a good one that captures what we mean by the term “overconfidence”.}

The reduction of overconfidence fits nicely to ecological models, even if an elimination of overconfidence was not achieved in all studies. The hard-easy-effect observed across studies clearly contradicts PMM theory. However, Justlin et al. (1998) show by simulation studies, that a weakened version of ecological models incorporating unsystematic judgment errors can account for the finding of a hard-easy-effect. PMM theory can be seen as a “limiting case” of the ecological models.

Schneider (1995)

\textit{Experiment 1:} 73 participants judged 80 items each. The items were handwriting samples collected on a campus flea market. Participants had to judge whether a sample was written by a male or a female. In addition to the confidence judgments, participants had to supply frequency estimates of percentage correct in order to test the confidence-frequency-effect. Mean overconfidence was a significant value of 5.3% despite “representative sampling”. Furthermore, the data showed a clear hard-easy-effect when items were grouped post hoc according to solution probability. The frequency judgments were underconfident as apparently predicted by PMM theory (but see discussion below). Schneider found out that some of the items were “misleading” showing a solution probability less than $p = 0.5$. When these items were eliminated, overconfidence disappeared.

\textit{Experiment 2:} 56 participants judged the same 80 handwriting samples as in Experiment 1. Overconfidence was 7.6%, there was a large hard-easy-effect.

The post hoc elimination of “misleading” items in Experiment 1 is not tenable. In every reference class, there are misleading items. The rationale of representative sampling is to ensure that these items are not overrepresented in the sample. If these items are deleted
post hoc, the resulting elimination of overconfidence cannot be interpreted in favour of PMM theory. Schneider’s apparent corroboration of the confidence-frequency-effect cannot be readily interpreted in terms of PMM theory, either. PMM theory assumes that confidence judgments and frequency judgments are based on different reference classes (e.g. “City populations” vs. “General knowledge items”). However, it is hard to imagine that different reference classes will be activated in the case of handwriting samples. The observed “confidence-frequency-effect” therefore is compatible with the simple alternative explanation that frequency judgments are in general more conservative than confidence judgments (see Treadwell & Nelson, 1996 and Bröder, 1999, for a discussion).

**Dunning, Griffin, Milojkovic and Ross (1990)**

*Study 1:* 38 Judges “predicted” the answers of 8 randomly chosen participants to 20 items of a behavior questionnaire. Beforehand, the judges were allowed 30 minutes to interview the “target” persons. Mean overconfidence was 11.5%.

*Study 2:* 60 students predicted the answers of their roommates to the behavior questionnaire. Mean overconfidence was 9.7%.

*Study 3:* 50 Judges predicted the answers of 10 randomly chosen target persons. Mean overconfidence was 8.7%.

These studies showed consistent overconfidence in predicting other persons’ behavior, even if these target persons were selected randomly. However, the samples were small in all cases. Furthermore, the *items* to be judged were not sampled randomly from a reference class. Dunning et al. (1990) state that “the degree of overconfidence found on particular items was virtually uncorrelated (...) with an undergraduate panel’s ratings of how ‘artificial’ versus ‘everyday and natural’ the items were.” (p. 572/573). This is an argument (although weak) that these items were not selected for difficulty.

**Griffin and Tversky (1992)**

*Study 4:* 28 participants predicted the behavior of a randomly chosen other participant in each trial of a modified prisoner’s dilemma game. Mean overconfidence was 15%.

*Study 5:* 298 participants judged 30 items of randomly chosen American states that were to compare on target variables of varying difficulty (population, voting behavior, education). Overconfidence was 6.5%, 8.5%, and 15.8%, respectively, thus showing overconfidence and a hard-easy-effect.

The overconfidence in Study 4 cannot be accounted for by a selected sampling of items. However, the sample size was relatively small and therefore sensitive to sampling error. The same argument applies to Study 5. Only 30 items were to be judged, and they
were the same for all 298 participants! This sample is likely to be biased, and therefore, the result is not very trustworthy.

**Suantak, Bolger and Ferrell (1996)**

*Study 1*: Two participants judged 1200 (!) randomly selected paired comparisons between states of the USA. The target variables differed in difficulty (e.g. population, number of unemployed persons). Suantak et al. report neither percentage correct nor mean confidence, but the calibration curves show overconfidence and a clear hard-easy-effect for both participants.

**Budescu, Wallsten and Au (1997)**

As Erev et al. (1994) and Budescu, Erev & Wallsten (1997) have shown, apparent overconfidence can be explained in representative item samples even if an ecological model is valid in principle. Representative sampling is only a *necessary* condition for perfect calibration, but not sufficient, when the additional assumption of unreliable judgments (unsystematic error variance) is made (For an extensive description, see Budescu, Erev & Wallsten, 1997). Whether substantial overconfidence (meaning a cognitive bias) can be inferred from the data depends on the question if the magnitude of overconfidence exceeds the amount that can be explained by the unreliability of judgments. Therefore, Budescu, Erev and Wallsten (1997) developed a model to achieve estimates of this error variance, determine the amount of overconfidence expected with perfect calibration plus error and compare the actual overconfidence value with this expectation. This test is possible at the individual level. *Experiment*: 32 participants judged 100 items, randomly selected from the 50 most populous cities in the USA. Each item had to be judged four times in order to allow for an estimation of unsystematic error variance. The target variable was population. Mean overconfidence was 12.9%. However, this value might be due to random error despite perfect calibration. Therefore, it was tested for each individual, whether the observed amount of overconfidence exceeded the expected amount. This was the case for 26 of the 32 participants. Two were classified as underconfident, two were perfectly calibrated, and for two others, the procedure could not be applied.

The study by Budescu, Wallsten and Au (1997) is remarkable in one respect. First, by including the possibility of random error and “partially out” the contribution of this error to apparent overconfidence, their test of the ecological models is quite conservative. Their data, however, suggest a refutation of the ecological models even in this weaker formulation incorporating unsystematic judgment error. One puzzling fact is that their descriptive overconfidence-value of 12.9% is quite large as compared to other studies that used representative sampling. So, a replication of the results is desirable to rule out the
possibility of some idiosyncratic sampling error in this study. Another problem with Budescu et al.’s procedure must be mentioned: To estimate error variance, items have to be judged repeatedly. Of course, effects of consistent answering cannot be ruled out and therefore, the “real” error variance might be underestimated by the procedure.2

Other studies
Two often-cited studies are the ones reported by Keren (1987) and by Murphy and Winkler (1984). Keren reported slight underconfidence (-4.6%) of 16 expert bridge players who predicted whether a “contract” was made during the game. In a second study, 24 amateur players played the same games, but showed considerable overconfidence (10.7%). The selection procedure of the stimuli in Keren’s study is not quite clear (see Keren, 1987, p.101/102), and therefore we do not know whether the items (games) were truly representative of typical bridge games. But as experts and amateurs received the same stimuli, the calibration difference between the groups is a fact that cannot be explained by the ecological models unless they incorporate random error in the learning process with the reference class (Soll, 1996). Murphy and Winkler (1984) demonstrated excellent calibration of experienced weather forecasters who predicted precipitation (see their figure 2).

3.2 Studies not included in the review

Four studies explicitly aiming at testing PMM theory were not included in the review for different reasons. Kühberger (1995) attempted to apply PMM theory to so-called “framing effects”. In the predictions “derived” from PMM theory, I see no relation to the theory as it has been described in the relevant literature (Gigerenzer, 1993, 1997; Gigerenzer et al., 1991).

Brenner, Kochler, Liberman and Tversky (1996) report an experiment intended to test the confidence-frequency-effect. Their participants specified frequency estimates for different populations as well as confidence ratings. The judgments did not differ reliably which was interpreted as a refutation of PMM theory. However, PMM theory does not predict a confidence-frequency-difference when proportion estimates are compared to confidence judgments. A difference is only to be expected when both judgments rely on different reference classes which was not the case in Brenner et al.’s study.

2 Budescu et al. (1997) evaluate this problem to be “negligible”. However, they supply no justification for that statement.
Soll (1996) extended the ecological model in incorporating sampling error during the learning phase. However, his experiment is unsuitable for testing the model because his items were not representative for the reference class.

A study by Harvey and Rawles (1992) is only published as a short abstract from which their method and results cannot be evaluated. For these reasons, the abovementioned studies were not included in the review.

4. Discussion

As the results described above show, the representative sampling was relatively successful in some studies but failed to eliminate overconfidence in others. As Mellers, Schwartz and Cooke (1998) comment, “reasons for these discrepancies remain unclear” (p. 464). Keren (1997) is far more critical in stating that “One wonders how many additional studies are still needed in order to disconfirm the representative design hypothesis” (p. 271, footnote 2). The initial studies by Gigerenzer et al. (1991) and Justlin (1993, 1994, 1995) demonstrate near-to-perfect (but no perfect) calibration and thus seem to support the ecological models. The excellent calibration of weather forecasters (Murphy & Winkler, 1984) and expert bridge players (Keren, 1987) also fit nicely into the ecological view: In both cases, stimuli were representative for reference classes with which participants had extensive experience. However, even the amateur bridge players in Keren’s study showed overconfidence despite having a lot of experience with the reference class and optimal learning conditions (immediate feedback). 3 The review of 44 tasks by Justlin et al. (1998) showed that overconfidence was not eliminated in all cases, and in some studies, underconfidence was observed. The strong prediction of an elimination of overconfidence therefore cannot be maintained even if the aggregation across studies demonstrates nearly perfect calibration. The other studies mentioned clearly failed to eliminate overconfidence with representative sampling. However, the studies of Dunning et al. (1990) and Griffin & Tversky (1992) must be interpreted with care because of relatively small sample sizes.

None of the studies intended to eliminate the hard-easy-effect was successful (Hoffrage, 1995, Experiment 5; Suantak et al., 1996, Experiment 1; Schneider, 1995, both experiments). The high correlation between overconfidence and item difficulty across the studies reported in Justlin et al. (1998) clearly demonstrates the existence of a hard-easy-effect even with randomly sampled items. The magnitude of the correlation between overconfidence and solution probability even suggests that difficulty is the major

3 Keren (1987) notes that all of his “amateur” players “had been players for a long time” (p.106).
determinant of overconfidence. The prediction of an eliminated hard-easy-effect is a direct implication of the predicted elimination of overconfidence, and therefore, the data disconfirm both predictions.

Some authors argue that the overall reduction in the magnitude of overconfidence compared to typical calibration studies is an argument in favor of the ecological models (e.g. Hoffrage, 1995). As Griffin and Tversky (1992) have pointed out, this argument is not justified because randomly selected items are (trivially) easier than items selected for difficulty, and therefore, this observed reduction might just be another demonstration of the hard-easy-effect.4

To summarize: The reported results seem sufficient to refute the strict versions of the ecological models if one accepts the assumption that a “representative” sampling has been realized in the critical studies. Obviously, a fair test of the models can only be achieved when this condition is met. As has been pointed out above, none of the studies reported a manipulation check to make sure that this prerequisite was fulfilled. Furthermore, no conventions exist to allow for a rejection of studies that do not meet some reasonable criteria, e.g. large item samples (a good candidate would be study 5 by Griffin & Tversky, 1992). In my opinion, this manipulation check is necessary to establish a firm falsification of the ecological models. If this objective is not met, there is always the possibility to “explain away” the critical findings in terms of sampling error in particular studies. However, there has repeatedly been shown the persistence of overconfidence and hard-easy-effect with large random item samples that are unlikely to be biased (Budescu, Wallsten & Au, 1997; Suantak et al., 1996), and so, the preliminary summary of evidence speaks against the model. At least, the results show that representative sampling is not sufficient for the elimination of the observed overconfidence bias.

Some authors have acknowledged the potential role of random errors in confidence judgments that might mimic an apparent overconfidence bias while “true” judgments are unbiased (see above). This possibility has been demonstrated by computer simulations, and it has been argued that the ecological models are valid in principle while the imperfect results (remaining overconfidence and hard-easy-effect) could be attributed to judgment unreliability. But the assertion that remaining apparent overconfidence despite representative sampling might be attributable to unsystematic error variance is not sufficient to show that it actually is. The only attempt I know of in which a separation of “apparent bias” and “true

4 Griffin and Tversky (1992) write: “The difficulty effect can also explain a main finding by Gigrenzer, Hoffrage, and Kleinbolting [sic]”. Of course, an effect cannot explain another effect. Only theories can explain effects. However, the argument remains valid that studies with representative sampling might produce smaller overconfidence effects because of the use of easier items -- whatever the correct explanation of the hard-easy effect might be.
bias” has been aspired, is the study by Budescu, Wallsten and Au (1997). Their data suggest that a considerable amount of “true overconfidence” is still observed when “apparent overconfidence” is partialled out from the data. But before this is interpreted as a falsification of ecological models, this data pattern has to be replicable.

Concluding remarks
The ecological models and especially PMM theory are preferable to older notions of imprecisely formulated heuristics as the “anchoring and adjustment heuristic” (Tversky & Kahneman, 1974) or the “confirmation bias” because they are more precise and allow for testable predictions. However, the predictions have not been confirmed in many instances. The supporting evidence, on the other hand, is questioned by Griffin and Tversky’s (1992) notion that the apparent reduction of overconfidence with representative sampling might just be another demonstration of the hard-easy-effect. This interpretation is strongly supported by the examination of Justlin et al.’s (1998) data.

The supplemental incorporation of random error to account for apparent bias is perhaps promising as a sensible extension of ecological models and may help to account for the data. To evaluate this hypothesis, a research strategy analogous to Budescu, Wallsten and Au’s (1997) approach seems indispensable. In order to assess, whether there is a “true” cognitive bias beyond the statistical artifact due to unreliability, there must be a possibility to estimate this error variance and its impact on the data. Second, it has to be shown that representative sampling of items per se reduces overconfidence and is not just another demonstration of the hard-easy effect. That is, the confounding between item sampling and difficulty has to be effectively eliminated in further studies. At the moment, the possibility cannot be ruled out that the reduction of overconfidence with representative sampling is almost completely determined by item difficulty!

Even if the ecological models will turn out to be false or have to be grossly modified, their heuristic value in stimulating research and formulating precise theories has been -- and hopefully continues to be -- great.
5. References


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


Berichte aus dem Psychologischen Institut der Universität Bonn


**Band 19 (1993)**

**Band 20 (1994)**

**Band 21 (1995)**

**Band 22 (1996)**

**Band 23 (1997)**

**Band 24 (1998)** (Kein Heft erschienen)

**Band 25 (1999)**