

Calibration Accuracy of a Judgmental Process that Predicts the Commercial Success of New Product Ideas

THOMAS ÅSTEBRO^{1*} and DEREK J. KOEHLER²

¹*Joseph L. Rotman School of Management, University of Toronto, Toronto, Ontario, Canada*

²*Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada*

ABSTRACT

We examine the accuracy of forecasts of the commercial potential of new product ideas by experts at an Inventor's Assistance Program (IAP). Each idea is evaluated in terms of 37 attributes or cues, which are subjectively rated and intuitively combined by an IAP expert to arrive at a forecast of the idea's commercialization prospects. Data regarding actual commercialization outcomes for 559 new product ideas were collected to examine the accuracy of the IAP forecasts. The intensive evaluation of each idea conducted by the IAP produces forecasts that accurately rank order the ideas in terms of their probability of commercialization. The focus of the evaluation process on case-specific evidence that distinguishes one idea from another, however, and the corresponding neglect of aggregate considerations such as the base rate (BR) and predictability of commercialization for new product ideas in general, yields forecasts that are systematically miscalibrated in terms of their correspondence to the actual probability of commercialization. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS intuitive judgment; forecasting; calibration; bootstrapping; case-based judgment

INTRODUCTION

Psychological research in the 'heuristics and biases' tradition has helped to identify the determinants of intuitive probabilistic judgments that can lead to predictable biases. Judgmental biases such as the tendency to ignore base rates (BRs) have been well established (Tversky & Kahnemann, 1974). However, critics of this field point out that there is relatively little research on the degree to which such biases operate in situations outside the laboratory, where much of the research in this tradition has been conducted (Koehler, Brenner, & Griffin, 2002). A related criticism is that much of the work demonstrating judgmental biases has involved undergraduate participants performing in unfamiliar decision environments, and that real and experienced

* Correspondence to: Thomas Åstebro, Joseph L. Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario M5S3E6, Canada. E-mail: astebro@rotman.utoronto.ca

decision-makers may perform better in familiar decision environments than would be expected based on the laboratory research. A final criticism, most often waged by economists, is that in high-stakes business decisions most biases are attenuated by the fact that there are strong incentives to make optimal judgments (Larrick, 2004).

The present study shows the usefulness of psychological theory even to judgments of highly experienced experts making forecasts in a well-structured judgment task where the stakes are high. The article uses a large set of data on real forecasts made by professionals in a deliberate, highly complex, and uncertain decision-making situation to explore these experts' forecasting abilities. Specifically, we examine experts' forecasts of the commercial potential of new product ideas.¹ Experts at so-called Inventor's Assistance Programs (IAPs) provided by several US and Canadian Universities, Small Business Development Centers, and the like help inventors and entrepreneurs to evaluate a specific new product idea or project before it has reached the market and advise the potential entrepreneur on whether and how to continue efforts (Udell, 1989). The forecasting method used by experts in the program we examine is a judgmental assessment of a large set of cues for a project and, further, an intuitive (rather than a formulaic, statistical) combination of the cue values into an overall assessment.

Provision of a forecast of a project's likely commercialization at an early stage in the development process can be highly valuable for making decisions about investments in new projects. Mansfield et al. (1977), for example, found clear evidence that the earlier the assessment of an R&D project the greater the future technical, commercial as well as financial success (pp. 25–32). But forecasting the commercialization prospects of a new product idea is highly challenging given the substantial uncertainty and lack of relevant data that often characterizes new products (Herbig, Milewicz, & Golden, 1993).

In addition to the inherent difficulties of the forecasting task, reliance on intuitive judgment as the basis for combining the implications of a set of predictive cues may introduce additional unreliability in the forecasting process, producing suboptimal forecasts. A long line of research comparing the performance of intuitive judgmental combination with statistical combination methods supports this claim (for a review, see Dawes, Faust, & Meehl, 1989). Judgmental unreliability also tends to become more pronounced when the uncertainty associated with the outcome being predicted is high, when the number of potential cues is large, and when those cues must be assessed in a manner that relies on pattern recognition or memory processes (Stewart, 2001), all of which is the case for the forecasting task faced by IAP experts.

Research comparing the forecasting accuracy of different types of experts (Ettenson, Shanteau, & Krogstad, 1987; Shanteau, 1992; Shanteau, Grier, Johnson, & Berner, 1991) and the associated prescriptive literature (Fischhoff, 2002; Larrick, 2004; Stewart & Lusk, 1994) would suggest that the IAP experts would be expected to be relatively immune to commonly-observed psychological biases. The reason is that the IAP experts' decision environment promotes unbiased judgments through a number of favorable conditions: the experts are well trained, the decision-making process is highly structured and decomposed into subtasks, and the experts have access to a vast library of past reviews of new product ideas which is routinely used to anchor assessments and judgments. Furthermore, the experts are paid substantial amounts for their reviews and systematically biased judgments would lead to decreased demand for their services, addressing the concerns typically raised by economists. Despite this, we demonstrate that these experts, who are shown by some measures to indeed be highly efficient in encoding and combining predictive cue values, nonetheless produce forecasts that exhibit systematic biases that are predictable from what we know about the psychology of judgment and decision making.

In the type of business environments that we examine it is important not only to rank order projects well but also to carefully calibrate the expected probability of commercialization since the expected probability of

¹Being an expert means 'having, involving, or displaying special skills or knowledge derived from training or experience.' (Webster's Dictionary).

commercialization will drive an evaluation of the return on investment in a given project. That is, the assessments must pass not only the test of internal coherence or consistency with one another, but also the test of correspondence against the external standard of actual commercialization success (Hammond, 1996). For example, the consequences of having a highly-ranked, if its commercialization chances are 75%, new product idea are quite different than if its commercialization chances are only 40%. We find that while the IAP experts rank order projects well, they are not well calibrated in terms of the associated probability of commercialization that their judgments imply. The experts exhibit over-extremity and over-prediction in their forecasts, which—as we elaborate below—is precisely the pattern of miscalibration expected to arise in this kind of environment from case-based reasoning. Over-extreme forecasts are closer to zero and one than they should be; over-prediction is the tendency to consistently over-estimate the probability of the target outcome.

We suggest that existing managerial processes that are available for screening and evaluating new product development projects when there is high uncertainty, such as the Q-sort method (Allen, 2003), are not enough to combat the miscalibration that arises from case-based judgment. To mitigate over-extremity and over-prediction arising from case-based judgment we suggest that new product development decision-makers incorporate class-based probability information in their assessments of projects through a direct link between rank order assessment and the probability forecast.

This article continues with a review of two important methodological approaches that we will use to evaluate the experts' decision accuracy: statistical bootstrapping and probability calibration. A short review of the decision-making context follows. An assessment of the context linked to psychological theory allows us to make two predictions: (a) that the experts are likely to be reliable and provide valid rank ordering of projects, and (b) that they are not likely to provide well-calibrated judgments and will tend to exhibit both over-extremity and over-prediction in their forecasts. In the Section on Results we use statistical bootstrapping and an analysis of probability calibration to test the hypotheses. We end with a discussion on the validity of psychological theory for analysis of decision-making in real settings by experienced managers and with implications for managers involved in evaluations and screening decisions of new product development projects.

STATISTICAL MODELS AND INTUITIVE JUDGMENT

The main conclusion from a large number of studies is that statistical models are at least equal and most often superior to intuitive judgmental methods for combining cue values in forecasting outcomes (Camerer, 1981; Dawes et al., 1989; Grove & Meehl, 1996).² This conclusion holds true across a number of decision-making contexts, in both real and experimental settings and for both experts and novices. In fact, Dawes (1979) reports that equal weights of cues often yield as accurate performance as statistically derived weights. The apparent reason is that the likelihood function is often relatively flat over many different combinations of weights (Edwards & von Winterfeldt, 1986) possibly due to high cue redundancy. Even randomly assigned weights to cues yielded greater accuracy than experts' judgments in five out of five samples (Dawes, 1979). Dawes and Corrigan (1974, p. 105) conclude, 'The whole trick is to decide what variables to look at and then to know how to add.'

Research on the accuracy of decision-makers' forecasts in the domain of new product evaluation is scarce.³ Zacharakis and Meyer (2000) compared the decision accuracy of 51 highly experienced Venture

²There is a related literature on combining statistical and intuitive forecasts. See, for example, Blattberg and Hoch (1990) and Hoch and Schkade (1996).

³There is, however, a larger related literature on non-probabilistic forecasting of new product dollar sales (Armstrong, Brodie, & McIntyre, 1987; Blattberg and Hoch, 1990; Herbig, Milewicz, & Golden, 1993; Tull, 1967).

Capitalists (VCs) to the forecasting accuracy of a bootstrap statistical model of the VCs predictions of seed and early-stage projects. The classification accuracy of a bootstrap model with equal weights was 60%, surpassing the VCs who had average decision accuracies between 17.1% and 39.5%. Only one VC had an accuracy equal to the bootstrap model, a result which is within the margin of error. These results indicate that substantial improvements are possible in the screening stage of investment decisions by using simple statistical models.

One reason why intuitive judgments may be inferior to those derived from statistical combination is that intuitive judgments are subject to unreliability while statistical combination is not. Humans, even expert humans, are unlikely to consistently encode and weight the predictive cues in arriving at a forecast; such unreliability will attenuate the accuracy of the resulting forecasts. Experts who perform a forecasting task in a highly structured, routinized manner, on this account, would be expected to perform less poorly in comparison to a corresponding statistical model (Murphy & Winkler, 1984). The expert forecasts examined in the present study are in fact derived from a highly structured, routinized forecasting process. (The extent to which the process is structured and thus likely to yield accurate forecasts is presented in the section following the next.) As such, we might expect any advantage of a statistical model over the expert forecasts in this study to be fairly modest.

CALIBRATION OF INTUITIVE JUDGMENT

Research comparing the performance of expert intuitive judgment against that of statistical models has as its focus the overall, correlational accuracy achieved by experts versus that of the model. Typically, the experts will do well according to these measures as long as they are able to accurately rank-order the cases they evaluate (e.g., in terms of likelihood) relative to one another. In the present research, we go beyond this global evaluation to develop and test hypotheses about specific biases in intuitive judgment, derived from psychological research in the heuristics and biases tradition, concerning the absolute level of accuracy or correspondence between judgments and outcomes. Specifically, we start with the assumption that the judgments of the IAP experts are primarily case-based, and test predictions that follow regarding the judgments' expected level of calibration, that is, their correspondence to (or systematic deviation from) the outcome of interest. This approach may provide better guidelines for improving the accuracy of intuitive expert judgments, which could be particularly helpful in cases where the barriers to the introduction of statistical models are high.

In developing the 'heuristics and biases' approach to the study of judgment under uncertainty, Kahneman and Tversky (1973, 1979; Tversky & Kahnemann, 1974, 1983) posited that intuitive judgments and predictions tend to be driven primarily by characteristics of the specific case at hand (e.g., the details of a new renovation project to be undertaken by a contractor) and tend to neglect characteristics of the broader class or category to which the specific case belongs (e.g., characteristics of residential renovations in general). The focus on case-specific characteristics and neglect of class-based aggregate properties leads to predictable judgmental biases, including BR neglect and non-regressive prediction (Tversky & Kahnemann, 1974).

The effects of such judgmental biases on the predictive accuracy of probability assessments (i.e., correspondence between predictions and actual outcomes) can be depicted using a calibration plot, which aggregates the assessments into probability categories and then plots the objective outcome probability for the set of cases assigned to a particular category against the mean subjective probability associated with that set. For example, all cases for which the judged probability of the outcome falls in the range between 0.30 and 0.45 might be aggregated, then the proportion of cases falling in that category for which the outcome holds would be calculated, and plotted against the mean probability assigned to the set of cases falling into that category (which, by definition, would be somewhere between 0.30 and 0.45).

If the judgments are perfectly calibrated, then all the points in the calibration curve should fall on the identity line, representing perfect or ideal calibration. A tendency to consistently over-estimate the probability of the target outcome, by contrast, results in a calibration curve that falls below the identity line. A tendency toward under-estimation results in a calibration curve that falls above the identity line. Judgments that are overly extreme (i.e., closer to zero or one than is justified) yield a calibration curve that is flatter than the identity line; a tendency toward under-extremity results in a calibration curve that is steeper than the identity line.

Specific predictions regarding patterns of systematic miscalibration can be derived from the notion of case-based intuitive judgment as instantiated in a formal model of subjective probability calibration. Random support theory (RST; Brenner, 1995, 2003) is particularly useful for this purpose, as it has free parameters that are interpretable as reflecting (in)sensitivity to important class-based characteristics (Brenner, Griffin, & Koehler, 2006; Koehler, Brenner, & Griffin, 2002). We offer a brief overview of RST here; for more details, see Brenner (2003) and Brenner et al., 2006.

RST is one of several stochastic models (Ferrell & McGoey, 1980) that have been developed to account for calibration data. RST is based on support theory (Tversky & Koehler, 1994), which represents the judged probability $P(A, B)$ that focal hypothesis A rather than alternative hypothesis B is correct as the balance of evidence, or *support*, for A relative to that for B . In the present application, as an illustration, we take the focal hypothesis to represent commercialization of the project and the alternative hypothesis to represent failure to commercialize the project.

RST treats the support for the focal and alternative hypotheses as random variables, with its free parameters representing characteristics of the (log-normal) support distributions. Specifically, it invokes two conditional support distributions for focal hypothesis A and alternative hypothesis B , one representing cases for which A is correct (i.e., for which the focal hypothesis holds) and the other representing cases for which B is correct (i.e., for which the alternative hypothesis holds). The two conditional distributions are sampled in proportion to the overall probability (or BR) of the focal hypothesis relative to the alternative hypothesis in the judgment environment. As in signal detection models, a discriminability parameter (α) represents the separation between support distributions for correct and incorrect hypotheses, that is, the extent to which the correct hypothesis tends to receive greater support from the available evidence than does the incorrect hypothesis. This parameter reflects the predictability of the outcome variable from the available cues (qualified by the judge's ability to use the cues effectively). Both outcome BR and discriminability (α), then, are free parameters that can be viewed as characterizing aspects of the judgment environment.

The model's remaining two free parameters, by contrast, can be viewed as characteristics of the judge's 'policy' (or strategy). The focal bias parameter (β) represents the extent to which the focal hypothesis is accorded systematically greater (or less) support than the alternative hypothesis (i.e., the separation between support distributions for the focal and alternative hypotheses) and therefore is favored in the probability judgment. The extremity parameter (σ) reflects the common standard deviation of the support distributions in the model; consequently, greater values produce more variable support values and hence are associated with more extreme probability judgments.

For any judgment environment characterized by the BR of the outcome variable (BR) and the predictability of the outcome from the available cues (α), there exist unique values of the focal bias parameter β and the extremity parameter σ such that the model will produce perfectly calibrated judgments (Brenner et al., 2006). Specifically, the focal bias parameter β must be set to reflect the outcome BR, with a higher value of β set to match higher BRs; and the extremity parameter σ must be set to reflect the predictability of the outcome from the available cues (as measured by α), with higher values justified by a more predictable outcome.

According to the notion of case-based judgment, however, the overall outcome BR and general predictability of the outcome from the available cues (i.e., the overall diagnostic value of the cues) are precisely the kinds of aggregate, class-based considerations that are typically neglected in intuitive

predictions. Within the RST framework, then, case-based judgment can be implemented as constraints on its free parameters, with β and σ being insufficiently sensitive to characteristics of the judgment environment (BR and α , respectively) to maintain well-calibrated judgments.

The case-based version of RST predicts that systematic patterns of miscalibration should arise in different judgment environments. Over-prediction, indicated by a calibration curve falling consistently below the identity line in the calibration plot, is expected when the outcome being predicted has a low BR; similarly, under-prediction (calibration curve above the identity line) is expected when the outcome BR is high. Over-extremity, indicated by a calibration curve that is flatter than the identity line, is expected when the predictability of the outcome from the available cues is low; under-extremity (calibration curve steeper than the identity line) is expected when outcome predictability is high. Precisely this pattern of results has been documented both in on-the-job expert judgments (Koehler et al., 2002) and in laboratory studies in which the relevant characteristics of the judgment environment were experimentally manipulated (Brenner et al., 2006). To make predictions regarding the accuracy and calibration of judgments we now examine the specific judgment environment faced by experts in the present study.

THE DECISION-MAKING CONTEXT

An IAP using the system developed by Udell (1989) was launched at the Canadian Innovation Centre (CIC) in Waterloo in 1976. Since 1982 the Canadian IAP has used full-time, in-house analysts and continuously revised and improved its evaluation method. The Canadian IAP evaluated more than 13 000 projects between 1976 and 2000.

The Canadian IAP evaluates the new product idea on 37 different cues and provides a recommendation to the potential entrepreneur. The service is provided for a fee that was approximately US \$185 in 1995. These fees at the time covered about half of the program's expenses, the rest being covered by the Canadian government. The average accumulated out-of-pocket R&D expenditures for the inventors at the time of evaluation are Cdn. \$6,625 (1995 value). The cues and their definitions employed during the study period are described in Appendix A. To have an idea evaluated, the entrepreneur fills out a questionnaire. In addition to background information about the entrepreneur, the questionnaire asks for a description of the idea and supplementary documentation such as patent applications, sketches, and test reports. The in-house analyst compares the idea with other similar ideas in their library of previous reviews and searches various databases. Personal contact with the entrepreneur beyond the provided documentation is avoided by the analyst. A particular salient feature is the use of a vast library of past reviews to do case-based comparisons. The analyst typically retrieves a few 'comparable' new product ideas from the library to anchor cue assessments on.

The analyst uses these data to subjectively rate the project on the 37 cues. There are three possible scores on each cue: A, very good; B, moderate; and C, a critical flaw. After rating all cues and recording the ratings on a sheet, the analyst determines an overall score for the project using intuitive judgment. Since the method of integrating the cues is intuitive rather than statistical, the overall assessment might differ across evaluations and evaluators even though data are identical. The judgment is an ordinal ranking, not an explicit probabilistic forecast, though it is supposed to be informative with regard to the idea's commercialization prospects. The analyst conducts the forecast without specific knowledge of the BR probability of commercialization.⁴ The judgment can be completely ignored by the client and the review does not necessarily provide any particular benefit in terms of preferred treatment from third parties.

⁴Data on the BRs were first presented to the IAP in 1997 (Åstebro, 1997). Before being presented with these data the senior analyst indicated that the IAP expected 'less than 10%' to be commercially successful.

Interviews with the senior analyst at the IAP indicated that the overall assessment is based on a mixture of two decision rules. If a project is critically flawed on one or more cues, either the lowest or next to lowest overall score is provided: D or E (i.e., non-compensatory weighting).⁵ If, however, there are no (or few) critical flaws, then the scores on the cues are usually assessed in some additive fashion. In addition to the review by a single expert, a group meeting is conducted where the evaluating expert presents a summary and a final overall score is agreed upon. The evaluation process typically takes 5–7 hours and may stretch over several weeks as the analyst collects information from various sources. A report is delivered to the entrepreneur consisting of scores on the 37 cues and a recommendation on commercialization options.

The decision situation contains a large number of cues that can only be qualitatively assessed and the outcome to be predicted is extremely uncertain. Decision-relevant information is uncertain and cannot easily be quantified. Many of the cues are themselves forecasts. The average time to event outcome (i.e., commercialization) is approximately 1.5 years. The IAP collects information about new product ideas that have gone through the review by clipping newspaper articles that they may find. Outcome feedback⁶ pertaining to the prediction is therefore biased and spotty. All these conditions are potential obstacles to making accurate predictions (Goldberg, 1968). On the other hand there is ample time to form an opinion. The IAP uses a standardized procedure where all cues are scored and a record is kept of all scores. The IAP employed the same senior analyst between 1981 and 2000. During that period all analysts were trained by the senior analyst in the evaluation procedure—the initial training took about 2 days followed by close supervision over 2 weeks. Analysts typically are engineers. The IAP is paid significant amounts, encouraging considerable deliberations. A group meeting at the end of the process where the analysts presents the evaluation and receives criticism from fellow analysts and the senior analyst also mitigates erroneous judgments. It appears that the process is reliable in terms of cue assessments. Baker and Albaum (1986) test the reliability of cue assessments across 86 judges and six products and find Cronbach alphas ranging from 0.84 to 0.96, implying high reliability.

HYPOTHESES

Conditions of the IAP are such that we would expect their experts to be efficient in their use and integration of cues to arrive at an overall evaluation of a project that can be used to accurately rank order the cases. The extensive individual experience of the IAP experts and their access to a library of past cases would be expected to promote efficient encoding of relevant cues and to provide some guidance as to how they should be integrated. By its decomposition of tasks, the standardized method of scoring cues, the careful deliberation and group review, the formalized evaluation process in place at the IAP would be expected to provide further safeguards on the reliability of judgments. We therefore hypothesize that:

H1: The IAP experts are expected to encode and integrate cue values in a valid, reliable manner in arriving at an overall assessment of a particular new idea.

Addressing this hypothesis, we first directly compare the forecasts made by the experts with observed outcomes to arrive at an overall measure of their classification accuracy. This test does not rely on the measurement of cues. We then construct a linear additive statistical model that examines the correlation between cue values and forecasts. This model is often referred to as the ‘bootstrap’ model. A comparison

⁵The overall score D is typically assigned to projects that have little or no novelty value (i.e., where similar products are already available on the market). The score E is reserved for those with obvious technical flaws that the IAP believe cannot be corrected.

⁶While other kinds of feedback could be provided even in the absence of outcome information, such as feedback regarding the correlation between the various cues and the judge’s forecast, task information feedback regarding the correlation between the cues and the outcome variable appears to be most helpful in improving prediction performance (Balzer, Sulsky, Hammer, & Sumner, 1992).

between the actual forecasts and the forecasts made by the bootstrap model will indicate the degree to which the experts produce reliable evaluations (Camerer, 1981). To the extent that the experts produce either cue value assessments or forecasts with high unreliability, the bootstrap model will so indicate through a poor fit to the actual forecasts. We then construct a linear additive statistical model that examines the correlation between cue values and observed outcomes. This model will be referred to as the 'prediction' model.⁷ This model indicates the informational value of the cues. A comparison between the actual forecasts and the forecasts made by the prediction model will indicate the degree to which the experts use cue information appropriately to produce valid evaluations that correctly rank order the cases in terms of commercialization probability (Blattberg & Hoch, 1990).

The case-specific evaluation upon which the forecasts are assumed to be based, however reliable and valid, is not by itself sufficient for well-calibrated probabilistic forecasts. Good calibration requires the mapping from the case-based evaluation to the probability scale to be sensitive to class-based considerations, that is, characteristics of the market or judgment environment to which the case at hand belongs. We believe that the IAP process is insufficiently designed for the experts to take such information into consideration. One obvious reason is the lack of systematic feedback on the commercialization of cases judged. In addition, while the forecasting process used at the IAP does seem to be constructed in a manner that promotes efficient use of case-specific cues in arriving at an overall evaluation of the strengths of the idea, it is less apparent that it does anything to encourage explicit consideration of class-based factors such as the overall BR or predictability of the target outcome (commercialization). In other words, though the judgment process is arguably more deliberative and highly structured than that of the typical 'intuitive'⁸ judgment task studied in the heuristics and biases literature (and so, for instance, might fall closer to the analytical end than to the intuitive end of Hammond's cognitive continuum of judgment modes; see Hammond, 1996, for a review), the IAP evaluation process is inherently focused on case-specific evidence giving rise to an evaluation of the overall strength of the idea (relative to other ideas) rather than its commercialization probability per se. An extensive library of past cases evaluated by the IAP is available but is generally searched for a specific case (or a few) that resembles the current case being evaluated, rather than as a source of aggregate information about the entire set of ideas evaluated by the IAP taken as a whole. We therefore predict:

H2: The experts' forecasts are expected to be systematically miscalibrated due to their insufficient sensitivity to class-based characteristics such as outcome base rate and predictability.

The case-based RST model can be used to derive specific predictions regarding expected patterns of calibration in light of the characteristics of the prediction task faced by experts in the present study. Two key features of the task, which are common in many important societal and business problems, are a relatively low outcome BR (i.e., only a small fraction of the new ideas are commercialized) and low outcome predictability (i.e., the available cues are far from perfectly predictive of which new product ideas will eventually succeed). According to case-based RST, then, we would expect the calibration curve representing the accuracy of the expert predictions to fall below the identity line (over-prediction, as expected given a low outcome BR) and to be less steep than the identity line (over-extremity, as expected given low outcome predictability).

More refined predictions can be derived, furthermore, if the set of cases (new product ideas) under evaluation can be meaningfully segregated into subsets that vary on the key dimensions of outcome predictability or BR. Because case-based judgment leads to neglect of aggregate, class-based characteristics of the set to which the case at hand belongs, we would expect insufficient adjustment for these characteristics. So, for example, if the cases can be segregated into two groups that vary in terms of how predictable the

⁷Brunswik (1955) refers to this as the 'ecological' model.

⁸In the heuristics and biases approach, a judgment is said to be *intuitive* if it is reached "...by an informal and unstructured mode of reasoning, without the use of analytical methods or deliberate calculation." (Kahneman & Tversky, 1982, p. 124).

outcome (successful commercialization) is from the available cues, we would expect some insensitivity to this difference between the two subsets and hence a flatter calibration curve for the subset of cases for which the outcome is relatively less predictable.

METHOD AND DATA

The sample frame for our development sample consisted of all 8797 valid records of new product ideas submitted to the IAP for evaluation from 1976 to 1993. We obtained 1091 usable responses from 1465 randomly sampled IAP clients who could be reached by telephone and asked to participate in a survey, representing an adjusted response rate of 75%. (For details on sampling plan and sampling bias tests see Åstebro, 2004.) The data set for analysis was ultimately reduced to 559 projects containing 499 failures and 60 commercialized ideas spanning the period 1989–1993.⁹ We further conducted a second telephone survey of all IAP clients between 1994 and 2001. This survey had a response rate of approximately 61%. (For details on sampling procedure see Åstebro, Jeffrey, & Adomdza, in press.) After merging survey information with administrative records on ratings, there remained complete data on 465 ideas, of which 425 were failures and 40 were successes. This second, hold-out, dataset will be used for model validation purposes.

Evaluation information in the IAP record included ratings for each of the 37 cues as well as the ideas' overall rating. Data on the independent variables were consequently collected before outcomes were observed and independent of this study. We therefore avoid any potential methods bias (Campbell & Fiske, 1959; Fischhoff, 1975). Three cues had too many missing data to be included. Missing data on the 34 remaining covariates were imputed assuming data are missing at random (MAR). The majority of cues had no missing data, while a few cues had up to 3.8% of observations missing. We converted the scores on the cues into numerical data according to the following: $A = 1$, $B = 0$, and $C = -1$. Table 1, columns (2) and (3), reports the frequency distribution of the responses over the IAPs' overall rating for the development sample. A majority of new product ideas (73% rating D or E) are advised to terminate efforts. Five per cent receive the most favorable overall score (A), 7% are advised to conduct additional market or technical analysis (B), and 15% are advised the idea is suitable to launch as a limited (i.e., part-time) effort (C). Hold-out sample data are also listed in Table 1. It appears that the judgments shifted over time away from the more extreme positions (A and E) and more towards a 'doubtful' evaluation.

The survey interview script contained the following question: 'Did you ever start to sell <NAME> or a later, revised or improved version of this invention?' Responses define a binary variable that takes unity if a new product idea ever obtained sales revenue, and zero otherwise. Follow-up questions with respect to how the invention was commercialized and the presence of revenues allowed us to verify an affirmative response as valid. We refer to this outcome as *successful commercialization* and use this to calibrate the experts' forecasts.

Commercialization is a necessary but not sufficient condition for financial success. See Åstebro (2003) for an analysis of the financial success of the ideas. In this study, we chose not to relate cues to the financial outcomes of the ideas because data on financial returns are much more difficult to estimate and contain some significant measurement uncertainty. Another benefit of using successful commercialization to calibrate the experts' forecasts is that data are readily observable for the whole sample whereas the financial rate of return is only observable for those reaching the market, a much smaller sample. It should however be noted that the overall ratings correlate well with both the idea's internal rate of return conditional on reaching the market

⁹Twenty observations were dropped for the regression analysis as they had no data on the predictors, and two observations were dropped because outcome data were uncertain at survey time. Further, data spanned two submission periods with somewhat different evaluation procedures, with the first period from 1976 to 1989 (early July), and the second from July 21, 1989 to 1993. Because both evaluation criteria and scales differed substantially across the two periods, we decided to use only data from July 1989 and onward.

Table 1. Projects undertaken by independent inventors

Rating	Development sample: 1989–1993			Hold-out sample: 1994–2001			
	Sample total	Per cent of all	Probability of commercialization (%)	Sample total	Per cent of all	Probability of commercialization (%)	Median return among commercial* (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A-recommended for development.	28	5	71	2	0	0	26.0
B-may go forward, but need to collect more data	38	7	24	36	8	17	26.0
C-recommended to go forward, returns likely modest	82	15	20	38	8	21	-13.2
D-doubtful, further development not recommended	341	61	4	383	83	7	-28.5
E-strongly recommend to stop further development	70	12	1	6	1	0	N/A
Weighted Average	559	100	11	465	100	9	-7.3
Total							

N/A: Return on investment IS not applicable, or alternatively, negative infinity.
 *Data for inventions rated A and B not possible to compute separately. Numbers on returns for A and B combined. Data from Åstebro (2003).

and the probability of commercialization (Table 1, columns 4 and 8). The median rate of return for those reaching the market is +26% for those rated A and B, it is -13.2% for those rated C and -28.5% for the few rated D that reach the market. These results suggest that the advice dispensed by the IAP to inventors has considerable predictive value both in terms of predicting whether they will reach the market and in terms of their financial success conditional on reaching the market. That the IAP provides net valuable advice to the inventors has indeed been shown by Åstebro and Bernhardt (1999) and Åstebro and Gerchak, (2001).

A concern is that the outcome data may be affected by a self-fulfilling prophecy. The advice provided by the IAP may affect inventors' efforts unduly. Hypothetically, if the cues are completely uninformative of commercialization likelihood while the recommendation turns out to be highly correlated with commercialization efforts (for example, due to affectation), we would observe positive and biased correlations between cues and the likelihood of commercialization when the correlations, in fact, should be zero. We therefore investigate this potential bias in depth.

RESULTS

Experts' decision accuracy

We first tabulate the decision accuracy of the experts by assuming that an idea rated 'A,' 'B,' or 'C' can be classified as a predicted 'commercialization' and those rated 'D' or 'E' can be classified as a predicted failure to commercialize. This information is then compared to the actual commercialization outcome as described above.

Table 2, columns (2) and (3), shows that, by this analysis, experts at the IAP predicted 441 out of the 559 outcomes correctly (78.9%). This breaks down into predicting 45 of 60 commercial ideas correctly (75.0%)

Table 2. Predictive accuracies

	Experts' judgments*		Bootstrap model ⁺		Prediction model ⁺	
	Failure	Success	Failure	Success	Failure	Success
Development sample						
Failed	396	103	392	107	407	92
Succeeded	15	45	15	45	17	43
	#	%	#	%	#	%
Overall predictive accuracy	441	78.9	437	78.2	450	80.5
Correctly predicts success (Sens.)	45	75.0	45	75.0	43	71.7
Correctly predicts failure (Spec.)	396	79.3	392	78.6	407	81.6
False positive's	103	69.6	107	70.4	92	68.1
False negative's	15	3.6	15	3.7	17	4.0
Correlation with outcome	0.38		0.40		0.41	
LR (χ^2)	97.78				78.82	
$p > \chi^2$	0.000				0.000	
Area under the ROC curve (AUC)**	0.813				0.815	
Pseudo-R ^{2***}	0.257				0.205	
Hold-out sample						
Overall predictive accuracy	81.1		72.0		66.2	

*Rating A, B, and C classified as 'success,' rating D and E classified as 'failure.' Test statistics are for a regression model with dummy variables that exactly replicates the experts' decisions.

**Hand (2001).

***Domencich and McFadden (1975).

⁺Only variables significant at $p < 0.05$ were included using stepwise backward variable elimination.

and predicting 396 of the 499 failures correctly (79.3%). There is a surprisingly even ability of the experts at predicting both failures and commercializations. Consider that the baseline probability of commercialization is low, around 0.11. Therefore, even though experts see approximately one commercialization in 10 reviews and thus have significantly less opportunity to obtain training on reviewing these, they are almost as able to make correct judgments on the high quality as the low quality ideas. It should be recognized, however, that the experts make a significant number of false positive predictions. Out of 148 new product ideas predicted to be commercial, 103 (69.6%) were actually failures.

The classification accuracy of the experts must be analyzed within the decision-making context. One important contextual feature is the low probability of commercialization: 0.11, suggesting that a classification rule based on predicting all ideas as failures would be high. Indeed, 499 out of 559 ideas (89.3%) would then be correctly classified as failures while the percentage of correctly classified commercializations would be 0% (0 out of 62). While accurate in the aggregate, this rule yields non-diagnostic advice that completely defeats the purpose of the IAP to identify and encourage potentially commercial new product ideas and would not provide a useful service to inventors. Another simple rule (probability matching), which also yields non-diagnostic predictions, would be to use the BR of 11% commercializations to forecast 60 randomly chosen ideas to be commercial. This rule correctly classifies 444 of 499 failures (89.0%) and 7 out of 60 commercializations (11%) for an overall classification accuracy of 80.7%. To compare the experts' forecasting accuracy to that of appropriate base-line models, given the goal of producing diagnostic predictions, we next calibrate the statistical models such that the proportion of correctly classified commercializations is maintained at approximately 75%.

Bootstrap model

A bootstrap model will give a sense of how well the experts' 'policy' as captured by a model of the relationship between cue information and judgments performs, assuming for a moment that the experts' 'policy' contains only main and linear effects on the odds of commercialization. A logistic regression model is fitted using stepwise backward variable elimination with the overall rating as the outcome variable, assuming ideas rated 'A,' 'B,' or 'C' are forecasted as commercialized (=1) and those rated 'D' or 'E' are forecasted as failures (=0), using the cues as independent variables.¹⁰ Using a *p*-value of 0.05 to determine inclusion of predictors, the resulting bootstrap model has a pseudo-R² of 0.59 [LR $\chi^2(11) = 378.14$] and contains 10 of the possible 34 explanatory cues. The cues predicting experts' forecasts are, 'Profitability,' 'Functional Performance,' 'Protection,' 'Appearance,' 'Duration of Demand,' 'Size of Investment,' 'Tooling Cost,' 'Development Risk,' 'Potential sales,' and 'Function.' The bootstrap model is a good descriptor of the experts' judgments, correctly classifying 503 of the 559 decisions (90.0%). This result indicates that even without knowing the experts' decision rules an appropriate statistical meta-model of these rules can be created.¹¹ The bootstrap model correctly classifies 78.2% of the outcomes within-sample and 72.0% out-of-sample. The within-sample accuracy measures are approximately equal to those for the experts' judgments themselves [Table 2, columns (4) and (5)]. The relatively high agreement between the simple bootstrap model and the experts' judgments suggests that the expert judgments are made with fairly high

¹⁰An ordinal logit model was also estimated that did not provide greater classification accuracy than that here reported.

¹¹This model overestimates the reliability of the experts' policy by using within-sample validation. The model's out-of-sample prediction accuracy of decisions is lower, 80.9%. The decrease in model accuracy is traced in part to the apparent change in decision policy between the two sample periods, as evident from Table 1. Further, the bootstrap model is sensitive to data and alternative model reduction techniques, indicating a flat maximum likelihood and high cue redundancy. Much more work could be done to further optimize the bootstrap model given these conditions. For such work see Åstebro and Elhedhli (2006). Since in this article the bootstrap model is not used to predict future decisions but simply to describe the experts' within-sample policy, we were content with presenting a simple linear additive model representation, as long as its descriptive accuracy was reasonably high.

reliability; the trivial difference in predictive accuracy between the bootstrap model and the expert judgments, in turn, suggests that the cost of any unreliability in the expert judgments is relatively low.

The experts and the bootstrap model typically agree in their predictions, with 503 cases on the diagonal. The experts' deviations from the bootstrap model are spread between 26 predicted commercial successes when the model predicts failure, and 30 predicted failures when the model predicts commercial success. The prediction accuracies of the off-diagonal elements are about equal showing neither an informational advantage nor bias by the experts. On the ideas where the experts disagree with the bootstrap model are the probabilities of success, 0.09 and 0.11, respectively, none statistically significantly different from the BR. Apparently, any ability of the experts to exploit nonlinear cue-outcome relations is offset by the bootstrap model's greater reliability.

Prediction model

The prediction model describes the 'true' relationship between cue information and outcomes and is estimated to benchmark the calibration accuracy of the experts' judgments. Typically the prediction model is reduced in complexity to linear and additive effects, which will capture a majority of the variance (Dawes et al., 1989). For the purpose of obtaining robust statistical prediction, a backward stepwise variable elimination procedure is used including only predictors that are significant at the 5% level and the logistic specification, which is linear in effects on the odds of commercialization. Results are robust to alternative elimination procedures such as forward elimination and to alternative model specifications such as the probit.¹² The cut-off for classification is calibrated such that the proportion of correctly classified successes is similar to the judges' accuracy at approximately 75%.

The statistical procedure selects the following four cues as predictors: 'Profitability,' 'Development risk,' 'Functional Performance,' and 'Protection.' These together correctly predict 450 out of 559 outcomes (80.5%), while the out-of-sample forecasting accuracy is reduced to 66.2%. Although arriving at a slightly higher classification accuracy, the prediction model used fewer cues than did the bootstrap model. Columns 6 and 7 of Table 2 indicate that the model correctly predicts 43 of 60 successes (71.7%) and 407 of the 499 failures (81.6%) within sample. The rates of false negatives and false positives are very close to the experts' rates.¹³ The incorrect classifications by the model could be attributable either to the difficulty of predicting the outcome from the cues or the difficulty of correctly coding the cues themselves in the first place.

We cannot reject the hypothesis that the experts come close to an optimal use of cue information when the basis of comparison is linear additive effects. The experts' judgments are close in accuracy to a linear bootstrap model of their policy and the bootstrap model is, in turn, a close representation of the (linear)

¹²The likelihood function for this model is steeper than that for the bootstrap model leaving less room for alternative model specifications based on slight changes to data or estimation procedure. One cue, 'Functional Performance' gets interchanged for 'Function' between the two methods while the pseudo-R² changes only on the third decimal when using forward instead of backward variable elimination. Reviewing Appendix A it is seen that the two cues are almost identical.

¹³It might be argued that the backwards elimination and use of only main and linear effects does not capture all useful information. We examined this argument using split samples, one covering 1989–1992 (estimation sample) and one covering 1993 (prediction sample). First, it should be noted that the calibration accuracy of the experts was 83.8% in 1993, and the forecasting accuracy of a prediction model estimated on the 1989–1992 sample was 82.6% for 1993. Using all 34 predictors we find a better fitting model with a within-sample pseudo-R² of 0.283 and an area under the ROC curve of 0.851. However, when applying this model to the 1993 sample it had an overall prediction accuracy of merely 71.9%. This indicates over-fitting. Another method is to include predictors with higher *p*-values, but not all 34. We explore this using stepwise regression with an inclusion criterion of $p < 0.10$. In addition to the previous predictors, this allows entry of five more cues for a within-sample pseudo-R² of 0.21 and an area under the ROC curve of 0.820 for the 1989–1992 pool. Applied to the 1993 pool, the model correctly predicts 12 successes (70.6%) and 122 failures (81.3%) for an overall forward prediction accuracy of 80.2%. This model does not improve prediction accuracy. Finally, including all two-way interactions among the four significant predictors does not increase the 1989–1992 within-sample pseudo-R² or the area under the ROC curve and the out-of-sample predictions are identical to the simple linearly additive model (82.6%).

predictive value of available data. Predictions by the experts deviating from the bootstrap model are neither biased nor particularly effective, but simply random. The experts pay attention to the cues indicated by the statistical model as most important in predicting commercialization and, in addition, use several cues that are not predictive when making their judgments.

To get a sense of the small differences in accuracy between expert judgments, the bootstrap model and the prediction model, compare these results to the summary of five similar comparisons in Dawes (1979). The average marginal improvement in forecast accuracy of the bootstrap model over experts' judgments across those five studies is 23% while in the current study the experts perform 1% better. The average marginal improvement in forecast accuracy of a prediction model over experts' judgments across the five studies is 73% while it is 2% in this study.

Given that this is not an experimental study with randomized assignment but an analysis of real decisions there is a concern that the data may be affected by a partial self-fulfilling prophecy that may affect results. The self-fulfilling prophecy can affect our results in two ways. It may bias upwards the estimated predictive ability of the judges. It also poses a problem when one would like to estimate the relationship between cues and outcomes. The latter relationships are potentially contaminated by the effect that judgments have on outcomes. The self-fulfilling prophecy, however, does not bias reported results on the relationship between cues and judgments. That is, our bootstrap analysis that examines whether the experts' decision rules are reliable remains unaffected. Åstebro and Chen (2004) estimate the potential bias, defined as the average effect of the judgment on the probability of commercialization, controlling for the expected commercial quality of the idea. The expected quality of the invention is estimated econometrically by an index measuring the likelihood of reaching the market, while controlling for the selection bias. A more colloquial way of saying this (and a restricted case) is that they estimate the degree to which there is a bias in the relationship between cues and outcomes if the outcomes are 100% determined by the IAP recommendation, and not at all by the underlying quality of the ideas. The authors find that most of the efforts by the inventors are driven by the underlying quality of their ideas and that the IAP advice accurately reflects this quality. The most likely bias is a rather small increase (decrease) in the inventor's expectation of the probability of success as a function of a positive (negative) review by the IAP, while controlling for the expected quality of the idea. The detected bias is not large enough to invalidate the conclusion that the statistically estimated model describing the relation between the cues and the probability of commercialization is relatively unbiased for the majority of new product ideas. Therefore, the tests of the validity of the experts' forecasting accuracy remain valid.

Whereas the experts continue to produce accurate judgmental forecasts for the period 1994–2001 it becomes evident that both the bootstrap and prediction models developed on data for the 1989–1993 period deteriorate in out-of-sample tests. There can be many reasons for this deterioration. The policies of the experts appear to change over time, as noted above, and so the bootstrap model would need to be revised. The distribution of outcomes is also different for the hold-out sample, but much of this may be due to random variation, and the sample means of success are not significantly different. In any case, our analysis does not rely on strong out-of-sample prediction accuracies. The bootstrap model is merely intended to replicate the experts' decisions and the prediction model is intended to identify valid predictors within sample. The key pieces of analysis are within-sample comparisons of accuracy across these two models and with the judgment accuracy of the experts.

Calibration

Here we investigate the calibration of the IAP experts' judgments, that is, their correspondence to the actual commercialization outcomes of the ideas. In our dataset, analysis of the calibration of expert predictions is somewhat complicated by the nature of the IAP classification system, which does not directly elicit subjective probability assessments from the experts. The bootstrap model discussed above, however, usefully captures

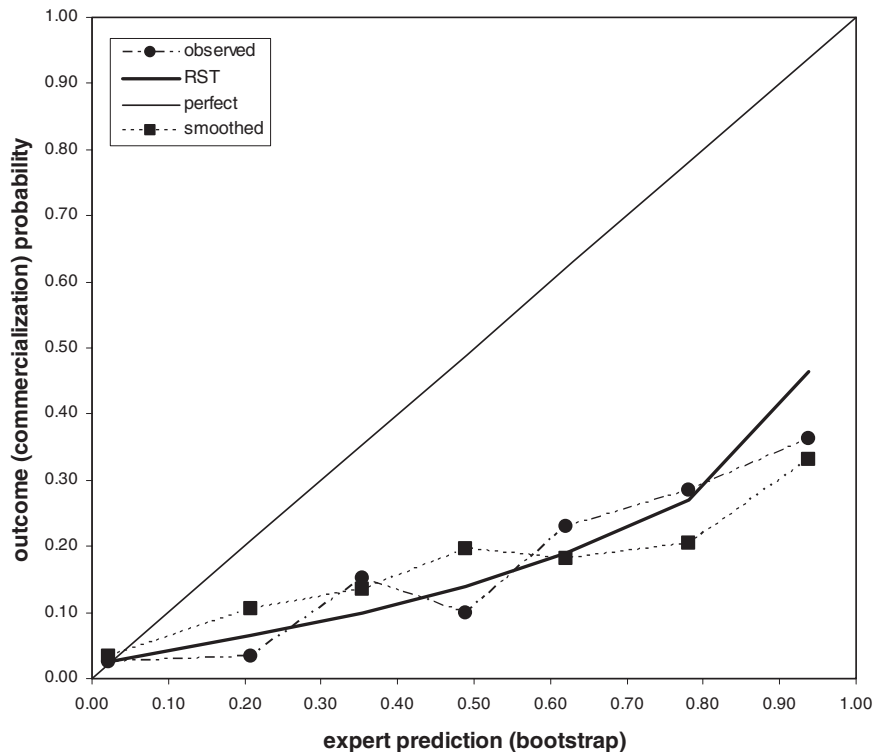


Figure 1. Calibration plot showing correspondence between IAP expert predictions of commercialization probability derived from the bootstrap model and the actual probability of commercialization (as derived both from observed commercialization probabilities and 'smoothed' estimates of commercialization probabilities estimated from the prediction model). Fit of RST model to the smoothed expert calibration curve is also shown

the predictive policy of the experts and produces, for each case, a predicted commercialization probability that can be subjected to conventional calibration analysis. In the analysis reported below, then, the probability assigned to a case by the bootstrap model of the experts' judgments, along with an outcome measure associated with that case, is the dependent variable.

In Figure 1, expert predictions (in the form of a probability derived from the bootstrap model of the experts' judgment policy) were aggregated into 7 probability categories uniformly spanning the unit interval; outcome probability (proportion of ideas commercialized) is then plotted for each probability category associated with the expert predictions (curve labeled 'observed,' round dots). The use of 7 probability categories, while somewhat arbitrary, was chosen as it is large enough to provide a sufficient number of points for an informative characterization of calibration performance but small enough that each point is based on a reasonably large number of observations. As expected if the expert judgments are case based, as elaborated in the introduction, the calibration curve falls consistently below (over-prediction, expected given the low outcome BR) and is flatter than (over-extremity, expected given the relatively low level of predictability of the outcome from the cues) the identity line.

The observed calibration curve in Figure 1 was constructed using the dichotomous outcome variable (commercialized vs. not commercialized). For smaller sample sizes, as in analyses to be reported below, it is useful to construct a less noisy variant of the outcome variable. To do this, the prediction model is used to

estimate the probability of commercialization for each idea, and then the aggregate estimated probability of commercialization is calculated for each set of ideas assigned to each judged probability category and used as the outcome variable in the calibration analysis (instead of simply using the proportion of those cases that were commercialized). For comparison, the calibration curve for the full dataset that results from this approach is also shown in Figure 1 (curve labelled 'smoothed,' square dots). The figure shows that the observed pattern of calibration is quite similar to that found using the more conventional method of analysis, but with fewer pronounced non-monotonicities.

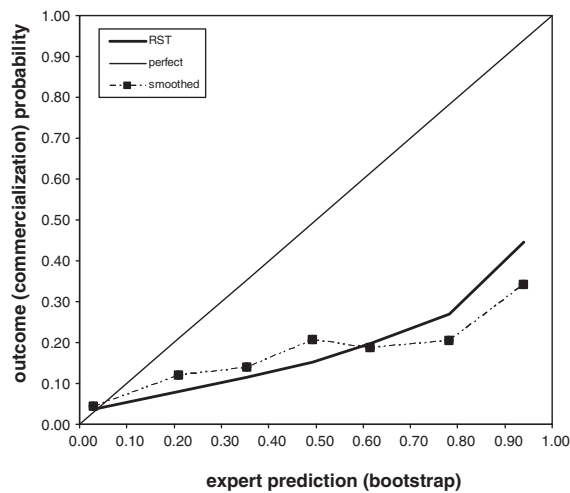
The RST model was fit to the smoothed expert calibration data (curve is labeled RST in Figure 1), which associates each case (idea) with a forecast probability (from the bootstrap model of the judge) and an outcome probability (from the prediction model). Best-fitting parameter values¹⁴ were $\alpha = 0.806$, $\beta = -0.364$, $\sigma = 1.281$, with the sample outcome BR set to 0.107 (see Table 1, bottom row). The observation that $\sigma > \alpha$ confirms that the predictions are too extreme (as indexed by σ) relative to the predictability of the outcome (as indexed by α). The observed negative focal bias parameter β indicates some accommodation of the very low outcome BR, but not nearly enough; a much more negative value of β (-2.626) would be required to maintain good calibration. As a result, the judgments show a distinct over-prediction bias. Fitting the RST model to the noisier observed expert calibration curve (instead of the 'smoothed' curve) yielded largely similar results: $\alpha = 0.979$, $\beta = -0.271$ (ideal $\beta = -2.163$), $\sigma = 1.228$.

As a further investigation of the consequences of intuitive case-based judgment, the new product ideas under evaluation were split into two subsets differing in the overall level of uncertainty associated with their future prospects. This split was based on two cues, demand predictability ('How closely will it be possible to predict sales?') and development risk ('What degree of uncertainty is associated with complete successful development from the present condition of the innovation to the market ready state?'). Recall that each idea was rated as an A, B, or C on each cue which, as noted above, was translated for our analyses into a score of 1, 0, or -1 , respectively; higher scores on demand predictability and development risk indicate lower uncertainty. New ideas with a total score on these two cues of 0 or higher ($n = 358$) were assigned to the low uncertainty subset, and those with a score of -1 or lower ($n = 201$) were assigned to the high uncertainty subset. Broadly speaking, the cues characterizing a particular new idea would be expected to be less predictive of its eventual commercialization likelihood under conditions of greater uncertainty. If intuitive expert judgments of a new idea's commercialization prospects are primarily case based, we would expect them to be insufficiently sensitive to aggregate characteristics such as the overall predictability of the outcome variable from the available cues. Because they are not sufficiently 'corrected' for the overall level of predictability, therefore, we would expect the calibration curve for the expert predictions regarding the high uncertainty ideas to be flatter than that for the low uncertainty ideas.

To test this possibility, separate calibration curves were constructed for the low and high uncertainty ideas, and the RST model fit to each. Figure 2 shows the results. As predicted, the calibration curve for the high uncertainty ideas is flatter than that for the low uncertainty ideas. For the low uncertainty ideas, the estimated RST parameter values were $\alpha = 0.706$, $\beta = -0.049$, and $\sigma = 1.299$, with an outcome BR of 0.147 and ideal $\beta = -2.489$ required for good calibration. For the high uncertainty ideas, the estimated RST parameter values were $\alpha = 0.473$, $\beta = -2.181$, and $\sigma = 0.873$, with an outcome BR of 0.37 and ideal $\beta = -6.924$ required for good calibration. The lower value of alpha for the high uncertainty ideas confirms our assumption that the commercialization outcome is less predictable than it is for the low uncertainty ideas. The corresponding difference in sigma indicates that the experts did attenuate the extremity of their predictions for the high uncertainty ideas, but not sufficiently in light of the lower predictability of the outcome in these cases relative

¹⁴The model was fit to the conditional probability distributions (i.e., distribution of forecast probabilities given commercialization and given no commercialization), such that alpha and beta reproduced the means of the distributions and sigma their pooled variance. See Brenner, Griffin, & Koehler (2006) for RST model-fitting details.

Low Uncertainty



High Uncertainty

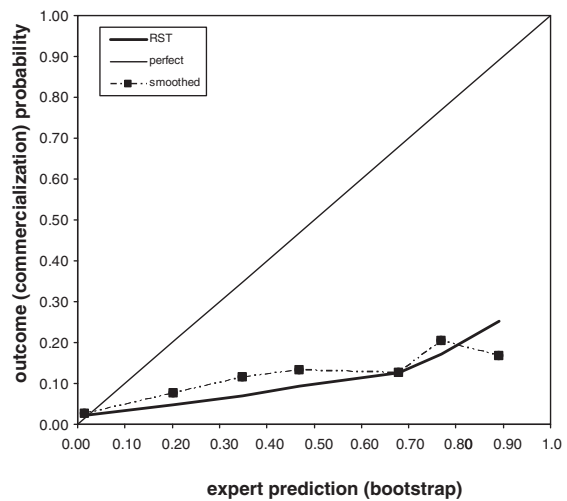


Figure 2. Calibration plots showing correspondence between IAP expert predictions of commercialization probability derived from the bootstrap model and the actual probability of commercialization ('smoothed') for low and high uncertainty ideas. Fit of RST model to each calibration curve is also shown

to that for the low uncertainty ideas. The lower outcome BR for the high uncertainty ideas indicates that not only was the outcome less predictable, but also the overall chances of commercialization were lower than for the low uncertainty ideas. The estimated value of beta for the two subsets of ideas indicates that experts did adjust substantially for the poorer prospects of the high uncertainty ideas, but again this adjustment does not appear to have been sufficient, with a tendency toward overly optimistic predictions (i.e., over-prediction) being present in both subsets and somewhat more pronounced for the high uncertainty subset in particular.

CONCLUDING REMARKS

Experts at the Canadian IAP are reasonably accurate forecasters of the future commercialization of new product ideas. Over a period of 5 years the experts were able to, *ex ante*, correctly classify 79% of the ideas. Such performance is impressive in light of the inherent uncertainty in predicting such notoriously unpredictable outcomes in a setting where feedback on their decisions is not readily available and the BR of commercialization success was not precisely known.

The most notable strength of the IAP experts is their highly effective use of the available predictive cues. The experts are marginally better at predicting outcomes than a (linearly additive) bootstrap model and achieve 98.0% of the accuracy achieved by an optimal linear statistical prediction model, showing high predictive validity and support of H1. The IAP experts appear to be significantly better at using the available predictive information than other experts, such as clinicians predicting psychosis, graduate admission officers predicting graduate student' performance, faculty members predicting the performance of other faculty members, and VCs predicting the likelihood that new ventures will succeed (Dawes, 1979; Zacharakis & Meyer, 2000).

Consistent with the notion that intuitive judgments (even those of experts) tend to be case-based, however, the commercialization forecasts of the IAP experts exhibited systematic miscalibration of the form that would be expected to arise from insensitivity to aggregate, class-based characteristics of the judgment environment. Specifically, their forecasts are biased in ways typical for settings with low BRs and high uncertainty, exhibiting both over-prediction and over-extremity, showing support for H2. Miscalibration can be costly in this context, even when the assessments accurately rank order the ideas in terms of their prospects, as a highly-ranked idea may have a substantially lower probability of commercialization (even if it is higher than that of many of the other ideas under assessment) than the assessment suggests.

Potential explanations for the experts' performance

In searching for explanations to the IAP experts' forecasting performance as well as their biases, we consider the application (or lack thereof) by the IAP of decision-making processes that have been recommended by researchers as 'good practices' (Fischhoff, 1982, 2002; Larrick, 2004; Stewart & Lusk, 1994).

A first observation is that a fair amount of training is provided to the experts. Second, the procedure is standardized and decomposed, something that has been argued to reduce the cognitive burden and increase reliability (Dawes & Corrigan, 1974; Einhorn, 1972). Except for forecasting weather (Murphy & Winkler, 1984), this particular process seems to be more standardized than most other real-world forecasting tasks described in the literature (Fischhoff, 2002). Although the cues are in many cases forecasts themselves (such as the degree of competition from new firms expected to enter after launch), and thus subject to error in measurement, Armstrong, Brodie, & McIntyre (1987) and York, Doherty, & Kamouri, (1987) demonstrate that measurement error in cues may not be critical, especially if there are many redundant cues. Further analysis of the data indeed revealed large amounts of multicollinearity across the whole matrix—if an analyst tends to assign an A to one cue she tends to assign an A to a cue of a related dimension. Selecting the 'wrong' cue for making the overall judgment is therefore not that detrimental to accuracy.¹⁵

Third, the group meeting at the end of the review may encourage careful evaluation of the idea by the IAP expert that could promote accurate predictions. Hagafors and Brehmer (1983), for example, suggest that reliability increases if forecasters are asked to verbally justify forecasts, especially when no outcome feedback is available. Larrick (2004) argues that the principal mechanism by which such accountability improves decision-making is pre-emptive self-criticism. In addition, the length of the deliberation regarding the assessment decision suggests that the analysts would not tend to make snap judgments, which otherwise are prone to be associated with greater decision-making biases than deliberate choices (Frederick, 2002; Slovic, Finucane, Peters, & MacGregor, 2002).

Fourth, the experts take steps to remain personally detached in their evaluations, for example, by not talking to the inventor. They see the primary value of their service as providing an impartial, outsider's view of the idea's commercialization prospects that the inventor, potentially prone to an overly optimistic assessment, may otherwise have difficulty obtaining.

Finally, the IAP experts' method of comparing a specific new product idea to a 'similar' new product idea archived in their extensive library of reviews may have both benefits and drawbacks. Edwards and von Winterfeldt (1986) argue that such comparisons may work well. For example, diamond evaluators reduce the decision problem to assessing similarities and differences with other remembered or currently available diamonds on key criteria. The IAP analysts use such case-based comparisons to sort a new product idea into

¹⁵Note that the overall forecasting accuracy which we document the IAP experts to possess does not rely on cue measurement. It is only when we construct the bootstrap and prediction models that we rely on cue measurement. To the extent that there is unreliable cue measurement, the corresponding bootstrap and prediction models would account for that by increased standard errors of estimates.

an ordinal ranking scheme without considering BRs. It is likely that some judgment error is avoided by this simplifying scheme. However, similar decision-making heuristics that focus on the case at hand rather than on class-based data has been shown in experiments to produce biased judgments (Kahneman & Tversky, 1973; Tversky & Kahnemann, 1983). We indeed find that the experts still exhibit over-prediction and over-extremity, the expected result of case-based reasoning in this judgment environment, despite their apparently effective use of the predictive cues. This situation is possible since accurate rank-ordering into five bins, each with a relatively wide probability-range, allows for considerable over (or under) prediction of true probabilities, while still conserving correct rank order.

Managerial implications

When the goal of an evaluation process such as the one reported here is simply to rank order cases (e.g., as part of an initial screening), intuitive expert forecasts are likely to do a reasonably good job, given circumstances such as those of the IAP: extensive experience, structured decision process, library of past cases. Indeed, under these conditions, the introduction of linear statistical models may not offer much in the way of gains in accuracy, though it might still produce savings in time and effort. In this respect our results echo those found in marketing, reviewed by Armstrong et al. (1987). But when the goal of the evaluation process is to associate a probability of commercialization with a particular case, rather than just rank ordering, the insensitivity of intuitive expert forecasts to class-based considerations in the mapping from the case-based evaluation to the probability scale can be costly. Indeed, in the type of business environments that we examine it is important not only to rank order ideas well but also to carefully calibrate the forecast probability of commercialization since the expected probability of commercialization will drive an evaluation of the return on investment in a given idea.

We suggest that existing managerial processes that are available for screening and evaluating new product development projects when there is high uncertainty, such as the Q-sort method (Allen, 2003), are not enough to combat the typical biases that can arise from case-based judgment. Prescriptions for avoiding miscalibration biases include the use of statistical models to carry out the mapping from experts' evaluations to the probability scale, as well as changes to the structured forecasting process that encourage explicit consideration of outcome BR and predictability. Using statistical models for decision support is advocated by many (Armstrong et al., 1987; Blattberg & Hoch, 1990; Hoch & Schkade, 1996). However, research has shown that there is resistance to the use of computer aids for decision making (Wierenga, Van Bruggen, & Staelin, 1999). If an organization can develop a decision environment that is as efficient as the IAP, the introduction of a regression model for decision support may not be necessary. Instead, in its simplest form the mapping from rank ordering to the probability scale could be accomplished by a table, such as that presented in Table 1, that takes into account the relevant outcome BR and predictability from the available cues. Such tables are not difficult to develop for most organizations as they are based on historical aggregate new product development data. And they fit the recommendation by Larrick (2004) of being simple to use and therefore more likely to be adopted.

When predictions are being made for cases that differ in the classes from which they are drawn (i.e., in aggregate characteristics such as outcome BR and predictability), the judgment process could also be structured in ways that encourage their explicit evaluation. It is likely that the predictability of cue information varies across stages in the development process or across various groups of projects such as small-scale enhancements versus large-scale projects. In addition, Herbig et al. (1993) suggest that predictability is greater for consumer product outcomes than for industrial products. An evaluation of the degree of predictability of the cue information is also possible using historical aggregate data. The same is true for the outcome BR, which may differ across product categories (e.g., consumer versus industrial products). Changing the decision process such that these variables are explicitly encoded for each new case may encourage experts to place greater weight on class-based characteristics in their forecasts.

APPENDIX A

Brief explanations of cues used by the IAP from 1989 and onward

Cue name	Explanation
Technical feasibility	Is the technical solution sound and complete?
Functional performance	Does this innovation work better than the alternatives?
Research and development	How great a burden is the remaining research and development required to bring the innovation to a marketable stage?
Technology significance	How significant a contribution to technology or to its application is proposed?
Technology of production	Are the technology and skills required to produce the new product idea available?
Tooling cost	How great a burden is the cost of production tooling required to meet the expected demand?
Cost of production	Does production at a reasonable cost level appear possible?
Need	Does the innovation solve a problem, fill a need or satisfy a want for the customer?
Potential market	How large and how enduring is the total market for all products serving this function?
Trend of demand	Will the demand for such an innovation be expected to rise, remain steady, or fall in the lifetime of this idea?
Duration of demand	Is the demand for the innovation expected to be 'long term'?
Demand predictability	How closely will it be possible to predict sales?
Product line potential	Can the innovation lead to other profitable products or services?
Societal benefits	Will the innovation be of general benefit to society?
Compatibility	Is the innovation compatible with current attitudes and ways of doing things?
Learning	How easily can the customer learn the correct use of the innovation?
Visibility	How evident are the advantages of the innovation to the prospective customer?
Appearance	Does the appearance of the innovation convey a message of desirable qualities?
Function	Does this innovation work better than the alternatives?—or fulfill a function not now provided?
Durability	Will this innovation endure 'long usage'?
Price	Does this innovation have a price advantage over its competitors?
Existing competition	Does this innovation already face competition in the marketplace that will make its entry difficult and costly?
New competition	Is this innovation likely to face new competition in the marketplace from other innovations that must be expected to threaten its market share?
Marketing research	How great an effort will be required to define the product and price that the final market will find acceptable?
Promotion cost	Is the cost and effort of promotion to achieve market acceptance of the innovation in line with expected earnings?
Distribution	How difficult will it be to develop or access distribution channels for the innovation?
Legality	Does the new product idea meet the requirements of applicable laws, regulations and product standards and avoid exposure to product liability?
Development risk	What degree of uncertainty is associated with complete commercialization of development from the present condition of the innovation to the market ready state?
Dependence	To what degree does this innovation lose control of its market and sales due to its dependence on other products, processes, systems or services?

(Continues)

Appendix A (Continued)

Protection	Is it likely that worthwhile commercial protection will be obtainable for this innovation through patents, trade secrets or other means?
Size of investment	Is the total investment required for the project likely to be obtainable?
Potential sales	Is the sales volume for this particular innovation likely to be sufficient to justify initiating the project?
Payback period	Will the initial investment be recovered in the early life of the innovation?
Profitability	Will the expected revenue from the innovation provide more profits than other investment opportunities?

ACKNOWLEDGEMENTS

Åstebro acknowledges partial financial support through a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, partial support from the MINE program, Ecole Polytechnique, and in-kind support from the Canadian Innovation Centre. Koehler also acknowledges support from a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

REFERENCES

- Allen, K. R. (2003). *Bringing new technology to market*. Upper Saddle River, NJ: Prentice Hall.
- Armstrong, J. S., Brodie, R. J., & McIntyre, S. H. (1987). Forecasting methods for marketing: Review of empirical research. *International Journal of Forecasting*, 3, 355–346.
- Åstebro, T. (1997). *The economics of invention and inventor's assistance programs*. Waterloo: University of Waterloo.
- Åstebro, T. (2003). The return to independent invention: Evidence of risk seeking, extreme optimism or skewness-loving? *The Economic Journal*, 113, 226–239.
- Åstebro, T. (2004). Key success factors for technological entrepreneurs' R&D projects. *IEEE Transactions on Engineering Management*, 51, 314–321.
- Åstebro, T., & Bernhardt, I. (1999). The social rate of return to Canada's Inventor's Assistance Program. *The Engineering Economist*, 44, 348–361.
- Åstebro, T., & Chen, G. (2004). Statistical decision-making models and treatment effects. Manuscript, available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=578525.
- Åstebro, T., & Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, 52, 395–409.
- Åstebro, T., & Gerchak, Y. (2001). Profitable advice: The value of information provided by Canada's Entrepreneur's Assistance Program. *Economics of Innovation and New Technology*, 10, 45–72.
- Åstebro, T., Jeffrey, S. A., & Adomdza, G. K. (in press). Inventor perseverance after being told to quit: The role of cognitive biases. *Journal of Behavioral Decision Making*. DOI: 10.1002/bdm.554
- Baker, K. G., & Albaum, G. S. (1986). Modeling new product screening decision. *Journal of Product Innovation Management*, 32–39.
- Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, 53, 35–54.
- Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36, 887–899.
- Brenner, L. (1995). A stochastic model of the calibration of subjective probabilities. Unpublished doctoral dissertation, Stanford University.
- Brenner, L. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior and Human Decision Processes*, 90, 87–110.

- Brenner, L., Griffin, D., & Koehler, D. J. (2006). Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97, 64–81.
- Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.
- Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Decision Processes*, 27, 411–422.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the Multi-Trait-Multi-Method Matrix. *Psychological Bulletin*, 56, 81–105.
- Dawes, R. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571–582.
- Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95–106.
- Dawes, R., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Domencich, T., & McFadden, D. (1975). *Urban travel demand*. Amsterdam: North-Holland.
- Edwards, W., & von Winterfeldt, D. (1986). Cognitive illusions and their implications for the law. *Southern California Law Review*, 59, 225–276.
- Einhorn, H. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7, 86–106.
- Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better? *Psychological Reports*, 60, 227–238.
- Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, 26, 32–53.
- Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 288–299.
- Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Fischhoff, B. (2002). Heuristics and biases in application. In T. Gilovic, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 730–762). Cambridge, UK: Cambridge University Press.
- Frederick, S. (2002). Automated choice heuristics. In T. Gilovic, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 548–558). Cambridge, UK: Cambridge University Press.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgment. *American Psychologist*, 23, 483–496.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2, 293–323.
- Hagafors, R., & Brehmer, B. (1983). Does having to justify ones judgments change the nature of the judgment process? *Organizational Behavior and Human Decision Processes*, 31, 223–232.
- Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.
- Hand, D. J. (2001). Measuring diagnostic accuracy of statistical prediction model. *Statistica Neerlandica*, 55, 3–16.
- Herbig, P., Milewicz, J., & Golden, J. E. (1993). The do's and don'ts of sales forecasting. *Industrial Marketing Management*, 22, 49–57.
- Hoch, S. J., & Schkade, D. A. (1996). A psychological approach to decision support systems. *Management Science*, 42, 51–64.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 234–251.
- Kahneman, D., & Tversky, A. (1979). Intuitive prediction: biases and corrective procedures. *Management Science*, 12, 313–327.
- Kahneman, D., & Tversky, A. (1982). On the study of statistical intuition. *Cognition*, 11, 123–141.
- Koehler, D. J., Brenner, L., & Griffin, D. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. In T. Gilovic, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 685–715). Cambridge, UK: Cambridge University Press.
- Larrick, R. P. (2004). Debiasing. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–334). Oxford, UK: Blackwell.
- Mansfield, E., Rapaport, J., Romeo, A., Villani, E., Wagner, S., & Husic, F. (1977). *The production and application of new industrial technology*. New York: W.W. Norton.
- Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79, 489–500.
- Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, 81, 75–86.

- Shanteau, J., Grier, M., Johnson, J., & Berner, E. (1991). Teaching decision making skills to student nurses. In J. Baron, & R. Brown (Eds.), *Teaching decision making to adolescents*. Hillsdale, NJ: Erlbaum.
- Slovic, P., Finucane, M., Peters, E., & MacGregor, D. (2002). The affect heuristic. In T. Gilovic, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). Cambridge, UK: Cambridge University Press.
- Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81–106). Norwell, MA: Kluwer.
- Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting*, *13*, 575–599.
- Tull, D. (1967). The relationship of actual and predicted sales and profits in new-product introductions. *Journal of Business*, *40*(3), 233–250.
- Tversky, A., & Kahnemann, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
- Tversky, A., & Kahnemann, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *91*, 293–315.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567.
- Udell, G. (1989). Invention evaluation services: A review of the state of the art. *Journal of Product Innovation Management*, *6*, 157–168.
- Wierenga, B., Van Bruggen, G. H., & Staelin, R. (1999). The commercialization of marketing management support systems. *Marketing Science*, *18*, 196–207.
- York, K. M., Doherty, M. E., & Kamouri, J. (1987). The Influence of cue unreliability on judgment in a multiple-cue probability learning task. *Organizational Behavior and Human Decision Processes*, *39*, 303–317.
- Zacharakis, A., & Meyer, D. (2000). The potential of actuarial decision models: Can they improve the Venture Capital investment decision? *Journal of Business Venturing*, *15*, 323–346.

Authors' biographies:

Thomas Åstebro is associate professor of strategic management at University of Toronto. Prior to that he held the University of Waterloo Associate Chair of Management of Technological Change. Dr Åstebro conducts research in technological change and entrepreneurship and is particularly interested in judgment and decision-making in this context.

Derek J. Koehler is associate professor of psychology at the University of Waterloo. His research investigates the intuitive assessment of uncertainty involved in everyday planning, prediction, and decision-making. With Nigel Harvey, he recently edited the *Blackwell Handbook of Judgment and Decision Making*.

Authors' Addresses:

Thomas Åstebro, Joseph L. Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario M5S3E6, Canada.

Derek J. Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada.