University of Pennsylvania
ScholarlyCommons

Marketing Papers | Wharton School

2001

Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/marketing_papers

Part of the Marketing Commons

Recommended Citation
Armstrong, J. S. (2001). Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting. Retrieved from http://repository.upenn.edu/marketing_papers/150

Suggested Citation:
Armstrong, J.S. Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting. In *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Ed. J. Scott Armstrong). Kluwer, 2001.

Publisher URL: http://www.springer.com/business+%26+management/business+for+professionals/book/978-0-7923-7930-0

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/marketing_papers/150
For more information, please contact libraryrepository@pobox.upenn.edu.

Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting

Abstract
Judgmental bootstrapping is a type of expert system. It translates an experts' rules into a quantitative model by regressing the experts' forecasts against the information that he used. Bootstrapping models apply an experts' rules consistently, and many studies have shown that decisions and predictions from bootstrapping models are similar to those from the experts. Three studies showed that bootstrapping improved the quality of production decisions in companies. To date, research on forecasting with judgmental bootstrapping has been restricted primarily to cross-sectional data, not time-series data. Studies from psychology, education, personnel, marketing, and finance, showed that bootstrapping forecasts were more accurate than forecasts made by experts using unaided judgment. They were more accurate for eight of eleven comparisons, less accurate in one, and there were two ties. The gains in accuracy were generally substantial. Bootstrapping can be useful when historical data on the variable to be forecast are lacking or of poor quality; otherwise, econometric models should be used. Bootstrapping is most appropriate for complex situations, where judgments are unreliable, and where experts' judgments have some validity. When many forecasts are needed, bootstrapping is cost-effective. If experts differ greatly in expertise, bootstrapping can allow one to draw upon the forecasts made by the best experts. Bootstrapping aids learning; it can help to identify biases in the way experts make predictions, and it can reveal how the best experts make predictions. Finally, judgmental bootstrapping offers the possibility of conducting ?experiments@ when the historical data for causal variables have not varied over time. Thus, it can serve as a supplement for econometric models.

Disciplines
Business | Marketing

Comments
Suggested Citation:
Armstrong, J.S. Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting. In *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Ed. J. Scott Armstrong). Kluwer, 2001.

Publisher URL: http://www.springer.com/business+%26+management/business+for+professionals/book/978-0-7923-7930-0

This book chapter is available at ScholarlyCommons: http://repository.upenn.edu/marketing_papers/150

Principles of Forecasting: A Handbook for Researchers and Practitioners,
J. Scott Armstrong (ed.): Norwell, MA: Kluwer Academic Publishers, 2001.

Judgmental Bootstrapping: Inferring Experts= Rules for Forecasting

J. Scott Armstrong
The Wharton School
University of Pennsylvania

ABSTRACT

Judgmental bootstrapping is a type of expert system. It translates an expert=s rules into a quantitative model by regressing the expert=s forecasts against the information that he used. Bootstrapping models apply an expert=s rules consistently, and many studies have shown that decisions and predictions from bootstrapping models are similar to those from the experts. Three studies showed that bootstrapping improved the quality of production decisions in companies. To date, research on forecasting with judgmental bootstrapping has been restricted primarily to cross-sectional data, not time-series data. Studies from psychology, education, personnel, marketing, and finance, showed that bootstrapping forecasts were more accurate than forecasts made by experts using unaided judgment. They were more accurate for eight of eleven comparisons, less accurate in one, and there were two ties. The gains in accuracy were generally substantial. Bootstrapping can be useful when historical data on the variable to be forecast are lacking or of poor quality; otherwise, econometric models should be used. Bootstrapping is most appropriate for complex situations, where judgments are unreliable, and where experts= judgments have some validity. When many forecasts are needed, bootstrapping is cost-effective. If experts differ greatly in expertise, bootstrapping can allow one to draw upon the forecasts made by the best experts. Bootstrapping aids learning; it can help to identify biases in the way experts make predictions, and it can reveal how the best experts make predictions. Finally, judgmental bootstrapping offers the possibility of conducting ?experiments@ when the historical data for causal variables have not varied over time. Thus, it can serve as a supplement for econometric models.

Keywords: conjoint analysis, expert systems, protocols, regression, reliability

---

In the early 1970s, I was flying from Denver to Philadelphia. Some fit young men were on the flight. Wondering who they were, I turned to the person sitting next to me to see if he knew. He did. His name was Ed Snider, and he owned the Philadelphia Flyers, the hockey team whose players were on this flight. Before we were out of Colorado, I realized that this was my big chance. I would convince him to use my services to select hockey players by developing judgmental bootstrapping models. Sports writers would learn about me. Other teams would then flock to my door. I would become rich and famous. So, after a suitable interval, I asked, “Tell me Ed, how do you select your players?” He told me that his managers had recently adopted a new procedure. Originally he was the only one in the Flyers’ organization who thought it would work. He said that it worked for the Dallas Cowboys, and many people thought they made the best draft picks in football. His managers resisted, but after a two-year experiment, they agreed that the new approach was better. What was this new method? It was judgmental bootstrapping. So much for my visions of glory.

“So, Ed, if this procedure works so well and with you telling other people about it, aren’t you afraid that the other teams will start using it?” “No,” he replied “they have too much confidence in their own judgment.”

People routinely use judgment to make important forecasts in many jobs (e.g., lawyers, , parole officers, doctors, production schedulers, loan officers, bankers, investors, and marketers). Many of these predictions are poor because of various biases and a lack of reliability in judgment. Sometimes it is difficult to find competent experts to make judgments. Often, making judgmental forecasts is time consuming. For example, Schneidman (1971) took four months to examine data from 25 subjects to decide who was most likely to commit suicide. This would not be an option for someone working a suicide hotline.

Judgmental bootstrapping (also called policy capturing) addresses shortcomings in judgment. It can help to identify and reduce biases, improve reliability, make the predictions by the best experts available for use by others with less expertise, reduce costs of forecasting, and provide forecasts rapidly.

In judgmental bootstrapping, the reasoning of experts is converted into a set of explicit rules. Judgmental bootstrapping infers what an expert did when making a set of predictions. It is a type of expert system, but it is based only on an expert=s predictions and cues (information the expert used). In contrast, expert systems are not limited to data used by an expert, nor by ways in which one might represent expertise (Collopy, Adya and Armstrong 2001).

Although a judgmental bootstrapping model is not as comprehensive or flexible as an expert, it applies the expert=s rules consistently. That is, it improves the reliability of judgment. This is advantageous because judgmental forecasts are typically unreliable, and this is a major cause of inaccurate forecasts (Stewart 2001).

HISTORY OF JUDGMENTAL BOOTSTRAPPING

Frederick Winslow Taylor called for scientific management in the early 1900s. He claimed that by observing jobs in a systematic way, one could determine how to do them better. Taylor concluded that this would be applicable to low-level occupations such as pig-iron handling. The ideal worker for such jobs, Taylor (1911, p. 59) said, “is so stupid that . . . he must consequently be trained by a man more intelligent than he . . .”

Taylor did not extend scientific management to jobs involving thinking. However, not long after, Wallace (1923) concluded that it was possible to also study jobs involving thinking.[1] He based his conclusions on studies by Hughes (1917). At that time, experts rated the quality of corn in the springtime to predict the eventual crop size. Hughes had experts rate 500 ears of corn. The experts agreed substantially with one another. Hughes then developed a bootstrapping model for a typical expert. Although the bootstrapping model correlated closely with the judges= predictions, it had only a small correlation with the actual yield. The bootstrapping model revealed that the experts put too much weight on the length of the ear; thus, the model provided useful feedback to judges. Hughes did not report on the accuracy of the experts= predictions of crop size.

In the 1960s, researchers in a variety of fields studied judgmental bootstrapping. They were not aware of each other’s work until Dawes (1971) reviewed the research. Dawes also coined the term *bootstrapping*. The term suggests that forecasters can lift themselves by their own bootstraps. It is an unfortunate name because it is used by statisticians to mean something else. As a result, the term *judgmental bootstrapping* is often used. However, I will use the term *bootstrapping* for the remainder of this paper.

DEVELOPING A BOOTSTRAPPING MODEL

In bootstrapping, experts make predictions about real or simulated situations. A statistical procedure can then be used to infer the prediction model. Bootstrapping starts with the expert's forecasts and works backwards to infer rules the expert *appeared to use* in making these forecasts. This contrasts with the more common approach to expert systems, where one attempts to determine what rules were actually used, and then perhaps what rules should be used. Bootstrapping uses the expert's forecasts as the dependent variable, and the cues that the expert used serve as the causal variables. The model is typically estimated by ordinary least squares regression analysis:

[1] Henry Wallace went on to a long political career, including being vice-president of the U.S. and entering the presidential race. Ironically, many people seemed to regard “thinking” as his weak area.

$$Y' = a + b_1X_1 + b_2X_2 + ... + b_nX_n$$

Bootstrapping models resemble econometric models (or linear models as psychologists sometimes call them), except that Y' represents the expert's forecasts, rather than actual outcomes. For example, one could provide a doctor with data about a sample of 50 patients, asking her to diagnose the patients and make predictions about the outcomes of various treatments. One would then regress the data on the explanatory variables against the doctor=s predictions.

The principles for developing bootstrapping models are based primarily on expert opinion and on commonly accepted procedures in the social sciences and econometrics (Allen & Fildes 2001). I discuss them in the sequence one might use in developing a bootstrapping model.

- **Include all of the variables that the expert might use**

Using non-directive interviewing, one could ask an expert what information she used to make the forecasts and why she used that information. Sometimes, however, experts may not be aware of the variables they are using. For example, an interviewer might believe that she is focusing on job skills when conducting personal interviews with candidates, yet such factors as the interviewee's weight, height, looks, and accent might be important. So the model should include variables that might have an important influence.

To ensure that all key factors have been included, it is helpful to ask a variety of experts what information they use. Furthermore, one might try to assess what other information the experts have about the situations and whether it might influence their forecasts. There may also be literature concerning how people make these or similar decisions.

While it is important to include all important variables, the number of variables should be small. Studies that use regression analysis to infer relationships can seldom deal effectively with more than five variables, and often three variables can tax the system.

To develop a model that is simple yet includes all important variables, analysts should narrow the potential number of variables to a manageable number, then control their entry into the regression analysis. In judging which variables to include, analysts should depend upon experts, prior literature, and available data. When analyzing the data, discard any variable whose sign is contrary to the belief of the expert and ensure that the magnitudes of the relationships look reasonable.

- **Quantify the causal variables**

Bootstrapping consists of running a regression against the variables used by the expert. To do this, one must quantify the variables with a reasonable degree of accuracy. To the extent that causal variables are difficult to quantify, one would expect bootstrapping to be less accurate.

One way to quantify variables is to have the experts make ratings. For example, in trying to assess whether job candidates would be successful researchers, one might rate the extent to which their papers contain ?important findings.@ Objective measures, such as the number of citations, would improve upon subjective ratings. Hamm (1991), in a study of highway engineers, found that bootstrapping models based on objective measures were more accurate than those based on subjective measures.

The formulation of a causal relationship is not a trivial step, especially when the effects are not linear. Consider, for example, the task of hiring newly minted PhDs. One of the best predictors of whether someone will publish is whether they have already published. Zero is a bad score for publications. A large number of publications is also likely to be bad as the evaluator might doubt the quality. This leads to a closer examination of the papers, which often serves to confirm the suspicion that the papers are of poor quality. So the best score is probably a small number of publications. This relationship can be reflected in the way the variable is scaled. Two dummy variables would make sense in this case: "Did the candidate publish?" which is good, and "Did the candidate publish more than six papers?," which would be bad. Or one might focus on impact instead. "How many important findings has the candidate made?" or "Did the candidate publish an important paper?"

- **Use the most successful experts**

The analyst should draw upon the expertise of those who have been most successful at the forecasting task. Ideally, these experts understand the true relationships involved. This advice is based on common sense. Assume that you have the option of using a model based on predictions by the world=s best heart doctor. Your alternative is to use a model based on an average heart specialist. Which model would you use?

Roebber and Bosart (1996) show that because experienced weather forecasters receive good feedback about the accuracy of their forecasts, they use wider sets of cues than do inexperienced weather forecasters. These additional cues would be likely to produce a more accurate bootstrapping model.

What if you cannot find a good expert? Say that you are asked to develop a bootstrapping model of someone who lacks expertise. Bootstrapping can help here also, as long as the expert's forecasts have some validity.

- **Ensure that the variables are valid**

If the experts receive good feedback, an analyst might be able to identify valid variables by using the most successful experts. In addition, experts might have some awareness of invalid variables. In this case, the analyst should ask experts to choose the desired variables. Finally, the analyst might draw upon prior research to identify variables that are valid.

Experts may use invalid cues. An improvement in the consistency of an expert=s judgments might make things worse in such cases. For example, a number of experiments have demonstrated that academic reviewers are biased against new findings when they review papers for journal publication, especially if the findings are surprising and important (Armstrong 1997). Researchers with new findings should hope that the system is unreliable so that they might eventually have their work published.

- **Study more than one expert (or more than one group of experts)**.

The analyst can improve accuracy by developing bootstrapping models based on more than one expert, or, if working with groups, more than one group. Although little research has been focused on this topic, I generalize from the literature on the use of experts in forecasting which was based on Hogarth (1978), Libby and Blashfield (1978), and Ashton (1985). The analyst should study at least five and perhaps ten experts. Presumably, one would want to use more than five experts if their models differed substantially.

When working with group rather than individual predictions, reliability is less of a problem. Assuming that the group follows good processes, one would need few group bootstrapping models, perhaps three.

Analysts should develop a model for each individual (or group). Comparisons could then be made among the models. Do the models agree on the key variables and directions of relationships? Do they agree, roughly, on the size of the relationships? The analyst must resolve differences. I expect that a good first step would be to use median coefficients from various experts= bootstrapping models. (The analyst would have to recalculate the constant term). Alternatively, the analyst could use a few bootstrapping models and combine their *forecasts*; this strategy improved decisions in Moskowitz and Miller's (1972) study of a simulated production system.

- **Use experts who differ.**

If all experts use the same process, then it is sufficient to develop a model for only one expert. That situation would be unusual. Generally, experts differ, and it is desirable to seek experts who differ. Their models may differ because they use different data or relationships. For example, in forecasting sales for a proposed retail outlet, marketing researchers might know about the target market, store managers might understand customers= shopping habits, real estate people could generalize from similar stores in similar areas, and local retailers might have general knowledge about the market.

The extent to which experts differ might not be obvious initially. The analyst could develop bootstrapping models for five experts, examine their level of agreement, and then decide whether to use more experts.

- **Use a large enough sample of stimulus cases.**

The required number of stimulus cases varies depending on such factors as the complexity of the problem (the number of cases should increase as the number of cues increases), the expert=s experience and knowledge of relationships (fewer cases are needed if experts have good knowledge), the extent to which the causal variables are intercorrelated (more cases are needed if they are), the amount of measurement error in the variables (more cases are needed if there is more error), and the need for accurate forecasts. On the other hand, one would not want to overburden experts by giving them too many cases.

For simulated data (where multicollinearity can be eliminated), I suggest that analysts use at least 20 cases. For actual data, where collinearity and measurement error are common, analysts should use more than 100 cases. These are only rough estimates based on my inferences from studies to date and from discussions with a few researchers who have used bootstrapping models. Goldberg (1970), in analyzing actual data on mental patients, used 123 cases to develop bootstrapping models that proved to be more accurate than 79 percent of the clinicians. This was better than the 72 percent score when he developed models using 86 patients; however, an increase to 215 clinicians produced no further gain.

- **Use stimulus cases that cover most reasonable possibilities.**

Bootstrapping models allow one to make forecasts for a variety of possible situations. To this end, the analyst should ask the expert to make predictions for a wide range of possibilities. This will make it easier to obtain reliable estimates of the relationships. One way to do this is to find historical cases in which the causal variables differed substantially.

If the historical data showed little variation, the analyst can construct simulated experiments to ensure examining a wide variety of situations. *It is particularly important to introduce variations for factors that have been constant in the past but might change in the future.*

- **Use stimulus cases that display low intercorrelations yet are realistic**.

If the causal variables have been highly correlated with one another, it will be difficult to determine relationships. In such cases, the analyst can use simulated data. For example, the analyst could describe situations in which the prices of a brand were increased substantially, while the prices of competing brands decreased, stayed the same, or increased. At the same time, the analyst could simulate situations in which consumer income increased and some in which it decreased. One restriction is that these cases should all seem realistic to the experts. The design procedures are similar to those used for conjoint analysis (Wittink and Bergestuen 2001).

- **Use simple analyses to represent behavior.**

Researchers have tried to find the best procedure to capture the complexity of experts= rules. For example, Cook and Stewart (1975) examined seven different ways to obtain weights for variables. This included asking experts to divide 100 points among the variables, rate variables on a 100-point scale, make paired comparisons, and estimate ratios. They found that the procedures yielded similar results when the criterion was to match the expert=s decisions. Schmitt (1978) replicated this study and obtained similar findings, and further support was provided by Goldberg (1968, 1971), Heeler, Kearney and Mehaffey (1973), Slovic, Fleissner and Bauman (1972), and Wiggins and Hoffman (1968). Since different methods produce similar results, one might focus on choosing simple procedures. Simple procedures imply a simple functional form, such as an additive model with few causal variables and no interaction terms.

The predictive validity of the bootstrapping model is not highly sensitive to the type of regression analysis used. More surprisingly, it is typically not sensitive to the estimates of the magnitudes of the relationships. The key steps then are to (1) obtain the proper variables, (2) use the correct directions for the relationships, and (3) use estimates of

relationships that are approximately correct. For these principles, I have generalized from research on econometric models. Evidence is provided by Dawes and Corrigan (1974), who reanalyzed data from four studies: Yntema and Torgerson (1961), Goldberg (1970), Wiggins and Kohen (1971), and Dawes (1971). In this study, unit weight models (where the causal variables are transformed to standard normal deviates and deviations are then weighted equally) for cross-sectional data did better than bootstrapping models. Armstrong (1985, pp. 225-232) summarizes additional evidence.

Simester and Brodie (1993) developed a model for the sentencing of sex offenders in New Zealand. Thirty-eight judges did the sentencing. The models examined 23 features of the offenders and their offenses (which is more variables than I recommend). They developed a bootstrapping model based on 67 offenders and tested it on 22 of them. The bootstrapping model was about as accurate as a forecast that placed equal weights on the variations in each of the causal variables the experts used.

☝ **Conduct formal monitoring.**

If data become available on actual outcomes, the bootstrapping model can be recalibrated to improve the estimates. This information might also lead experts to reexamine their use of information. For example, Werner et al. (1984) examined predictions by 15 psychiatrists as to which of 40 mental patients might become violent. The experts= weights of factors and the weightings from an econometric model using actual data differed substantially. The experts thought that “suspiciousness” was related to violence (r = +.49), but it was not (r = -.03 against actual assaults). According to the econometric model, previous assaults and hostility judgments were related to assaults committed during the seven days following admission to the mental institution.

PREDICTIVE VALIDITY OF BOOTSTRAPPING

My search for evidence included checking references in key papers and using suggestions by researchers. I had little success with computer searches. The term “judgmental bootstrapping” yielded no hits from1988 to 2000 in the *Social Science Citation Index*, while “bootstrapping and forecasting” yielded only two relevant studies. I have included all published studies that contained empirical evidence on the accuracy of bootstrapping relative to other methods.

Evidence from Related Areas

Camerer (1981) summarized empirical evidence showing a close correspondence between experts= predictions and those from their bootstrapping models. This does not imply that the bootstrapping forecasts are more accurate, but it does suggest that bootstrapping models can capture key elements of an expert=s decision process. He concluded that bootstrapping should improve judgments under “almost any realistic task conditions.”

Grove and Meehl (1996) completed an extensive review of the empirical evidence on econometric models and concluded that they are equal to or more accurate than unaided judgment in most settings. If econometric models are superior to judgment, and if accuracy is not highly sensitive to the estimates of a model=s coefficients, one would expect bootstrapping models to be more accurate than unaided judgment.

The concern in this paper is with inferring *expert* judgment about the behavior of others. This procedure is similar to conjoint analysis, in which *consumers* report on their preferences when presented with hypothetical data in which product features vary jointly. (Following this line of reasoning, bootstrapping might be called exjoint analysis.) Wittink and Bergestuen (2001) discuss how models of potential customers= judgments about hypothetical products are used to forecast behavior. They provide evidence that these models provide good forecasts of consumers= decisions.

Bootstrapping is a type of expert system, therefore the performance of other expert systems is pertinent. Collopy, Adya and Armstrong (2001) summarized evidence from eight comparisons and concluded that expert systems improve forecast accuracy over that provided by expert judgment.

Three studies compared bootstrapping against decision makers in production problems. These studies required managers to make production decisions over time, using sales forecasts as one of the inputs. Bowman (1963) examined ice cream, chocolate, candy, and paint companies. A regression analysis of management=s decisions on production and the work force would have led to improvements over the decisions actually made in three of the four situations. Kunreuther (1969) developed a bootstrapping model for short-range production forecasting in an electronics firm. The model, developed partly from direct questions and partly from bootstrapping, was a simple two-variable model. According to Kunreuther, this model would have enabled the firm to carry a 25 percent smaller inventory while improving service. Moskowitz (1974) presented 86 managers with a simulated production problem. The managers made production and work-force decisions for one and three periods in the future. The goal was to make decisions that reduced costs for situations where the forecasting error varied from low to medium to high. The bootstrapping models led to better decisions than the managers had made for both forecast horizons and for all three levels of forecast error. In no case was a manager superior to his model. Moskowitz et al. (1982) added further support in a follow-up study.

Predictive Validity: Direct Evidence on Bootstrapping

Studies from various fields show that bootstrapping generally improves upon the accuracy of an expert=s forecasts. In some comparisons, experts appear to have more information than the bootstrapping model. However, bootstrapping=s gain from added consistency seems to outweigh the fact that it sometimes relies on less information.

To ensure that I had interpreted the studies correctly, I sent copies of my codings to each of the researchers (with the exception of the authors of one paper, whom I could not locate). Replies from the authors of eight papers led to some revisions. The evidence is impressive and it comes from such diverse areas as psychology, education, personnel, marketing, and finance. If you are not interested in the details, you can skip to the summary (Table 1).

Psychology: Yntema and Torgerson (1961) provided pictures of 180 ellipses to six subjects. The ellipses were various combinations of six sizes, six shapes, and five colors. They had been constructed so that their ?worth@ always increased with size, thinness, and brownness, although these were not linear relationships. Subjects were asked to judge the worth of each ellipse. They were trained with 180 ellipses on each of 11 days, with the order of the ellipses varying each day. Subjects were told the correct worth of the ellipses after each trial. On the twelfth day, the judges evaluated all 180 ellipses with no feedback. Yntema and Torgerson created a bootstrapping model for each judge. The average correlation between the judge=s evaluation and the true worth was .84, while the average correlation between the bootstrapping model=s prediction and true worth was .89. The researchers also constructed a model by asking the judges what weights they placed on size, shape, and color; this model did as well as the bootstrapping model (average correlation was .89). In other words, accuracy was not sensitive to the way the coefficients were estimated.

Goldberg (1971) asked 29 experts to use scores from a psychological test (the MMPI) to differentiate between psychotics and neurotics in a sample of 861 patients. He also presented the experts with scores on 11 variables from the MMPI. He developed bootstrapping models for each expert using part of the sample and tested them on the rest of the sample. He used various calibration samples. In one series of tests, he took seven samples of 123 cases each to develop the bootstrapping models and tested each model on validation samples of 738 cases each. The bootstrapping models proved to be more accurate for 79% of the experts.

Education: Wiggins and Kohen (1971) asked 98 graduate students in psychology to forecast first-year grade-point averages for 110 students entering graduate school. Bootstrapping models, developed for each expert, were superior to all 98 experts; furthermore, most of the bootstrapping models were more accurate than the best of the 98 experts and also more accurate than the combined forecasts by the 98 experts.

Dawes (1971) examined the admission decisions for the University of Oregon=s PhD program in psychology. The six categories for rating applicants were: (1) reject now; (2) defer rejection but looks weak; (3) defer; (4) defer acceptance but looks strong; (5) accept now; and (6) offer fellowship. The committee used scores on a quality index of the schools awarding the undergraduate degree, the Graduate Record Examination, grade point averages, letters of recommendation, and a record of work experience. A simple regression of admission committee decisions against these variables yielded a bootstrapping model that correlated well with the committee=s decisions for 384 applicants

(r= .78). None of the applicants who were rated in the lower 55% by the bootstrapping model were admitted by the committee. After 19 of the accepted students had been in the program for a year, Dawes found that the model=s predictions for their success were more accurate than the committee=s predictions. The correlation for the bootstrapping model=s predictions was roughly twice that for the committee=s predictions.

Personnel: Roose and Doherty (1976) developed bootstrapping models for each of 16 sales managers by asking them to predict the success of 200 life insurance sales agents, given information on 64 variables (more than what I see as a manageable number). They reduced this to five variables, unfortunately using stepwise regression to do so (which also violated my principles). They then used the models for a validation sample of 160 salespeople. Bootstrapping yielded small gains over the forecasts by individual managers. A consensus bootstrapping model did not improve upon the combined forecast from the managers. The bootstrapping model was less accurate than a unit-weights model with variables selected by a regression on actual outcomes. This suggests that the managers were not using the best variables.

In a study conducted in a corporate setting, Dougherty, Ebert and Callender (1986) developed bootstrapping models for three interviewers whose experience ranged from 6 to 26 years. They each saw the same 120 taped interviews, and rated the applicants on eight dimensions using nine-point scales. Their models matched their direct judgments rather well (average correlation was .9). Predictions by the experts and by their models were each compared with supervisors= ratings of performance after about ten months on the job for the 57 applicants who were eventually hired. As with other studies of interviews, the validities were low (the correlation for individual predictions was about .06). The bootstrapping models were much better than two of the three experts, and tied with the third.

Ganzach, Kluger and Klayman (2000) used 116 interviews of 26,197 male conscripts for the Israeli military. They made global predictions of the interviewee's "probability of success" from low to high (1 to 5). They then developed a judgmental bootstrapping model for each interviewer. The success of the model was judged using "disciplinary transgressions, such as desertion or imprisonment" over a three-year period. Judgmental bootstrapping was slightly less accurate than the global judgments (r of .216 versus .230).

Marketing: Ashton, Ashton and Davis (1994) developed bootstrapping models for 13 experienced managers to forecast the number of advertising pages *Time* magazine sold annually over a 14-year period. They gave managers data for one, two, or three quarters, and asked them to forecast total annual sales of advertising pages. The managers, who made a total of 42 forecasts (three forecasts for each year), were not previously familiar with the *Time* data. Interestingly, the researchers presented the data out of time sequence; they told the managers which quarter was involved but not which year. This eliminated information that managers would have had in a real situation. The bootstrapping model=s errors were smaller than the manager=s forecast errors for 11 comparisons, there was one tie, and in one case the model=s error was larger. On average, the bootstrapping model reduced the error by 6.4%. Besides reducing the average error, bootstrapping was less likely than the judge to make large errors. The largest errors in the bootstrapping forecasts were 80% as large as those in the managers= judgmental forecasts.

Finance: Ebert and Kruse (1978) developed bootstrapping models for five analysts who forecasted returns for 20 securities using information on 22 variables. The large number of variables violates the principle of simplicity and is risky because the number of variables exceeds the number of cases. To compound the problem, Ebert and Kruse used stepwise regression. They tested the models on samples of 15 new securities. Given that the models violated guidelines for developing bootstrapping models, it is surprising that the bootstrapping models were more accurate than analysts for 72% of the comparisons.

In a study by Abdel-khalik, Rashad and El-Sheshai (1980), bootstrapping models were as accurate as 28 commercial-bank lending officers in predicting defaults on loans. The savings here would be in reduced costs and reduced likelihood of bias in awarding loans.

Libby (1976), in a study concerning the prediction of bankruptcy for 60 large industrial corporations, concluded that experts were more accurate than their models. However, Goldberg (1976) showed that Libby=s results were due to severe skewness in the causal variables. When the data were transformed and reanalyzed, the percentage of times that the model beat the expert increased from 23% to 72%.

Summarizing Direct Evidence on Predictive Validity

Table 1 summarizes the eleven studies. The use of bootstrapping models is not risky. It generally improved accuracy, even when the researchers violated principles for developing bootstrapping models. The column on accuracy gain represents my judgments on the results of the comparisons. Overall, the gains have been consistent and substantial.

Table 1
Comparisons of Bootstrapping and Expert Forecasts

Area & Study	Task	Experts	Cases Calibration/ Testing	Cues tried/used	Criterion**	Gain in Accuracy
Psychology						
Yntema & Torgerson (1961)	artificial task	6	180*	3/3	correlation	small
Goldberg (1970)	psychotics or neurotics	29	123/738	11/11	% improved	modest
Education						
Dawes (1971)	Ph.D. candidates	1	384/19	4/4	correlation	large
Wiggins & Kohen (1976)	student grades	98	110*		% improved	large
Personnel						
Roose & Doherty (1976)	insurance salesmen	16	200/160	64/5	correlation	negligible
Dougherty et al. (1986)	white collar jobs	3	120/57	8/8	correlation	large
Ganzach et al. (2000)	military conscripts	116	26,197	6/6	correlation	small *loss*
Marketing						
Ashton et al. (1994)	advertising pages	13	42*	5/5	MAD	large
Finance						
Goldberg (1976)	bankrupt companies	43	60*	5/5	% improved	large
Ebert & Kruse (1978)	stock returns	5	35/25	22/?	% improved	large
Abdel-khalik et al. (1980)	bank loan defaults	28	32*	18/?	hit rate	none

Notes: * The same observations were used for development as for validation.
** MAD is the Mean Absolute Deviation; % improved refers to the percentage of forecasts that were more accurate than the experts.

CONDITIONS FAVORING THE USE OF BOOTSTRAPPING

The conditions favoring the use of bootstrapping vary depending upon whether the alternative is judgment or econometric methods.

Conditions Favoring Bootstrapping over Judgment

Four conditions favor the use of bootstrapping over judgment: (1) the problem is complex, (2) reliable estimates can be obtained for bootstrapping, (3) valid relationships are used, and (4) the alternative is to use individual inexperienced experts. These are discussed here.

- **Problem is somewhat complex.**

If the problem is simple enough, it may be unnecessary to develop a bootstrapping model because the judgmental process may be obvious. As complexity increases, the experts may not be able to use relationships consistently and efficiently. In addition, complexity makes it difficult for experts to use feedback effectively. In such

cases, bootstrapping, with its consistent approach, is likely to be more accurate than judgmental forecasts. If the problem is too complex, it might not be possible to structure it.

- **Reliable estimates can be obtained for the bootstrapping model.**

One way to judge the reliability of a bootstrapping model is to ask judges to make repeated predictions on the same data (or similar data), preferably at two points in time. The two time periods should be far enough apart that the judges cannot remember their earlier predictions. For tasks of moderate complexity, a week is probably sufficient. Separate bootstrapping models would then be developed for each set of estimates. Comparisons would be made for the judges' predictions or for the relationships. Einhorn, Kleinmuntz and Kleinmuntz (1979) used such a procedure in a task involving ratings of the nutritional quality of cereal. Their expert made three different judgments for twenty situations, a total of 60 forecasts. This allowed the researchers to examine the reliability of the judgments.

Libby (1976) tested reliability by repeating 10 of the 60 cases he had presented to the experts. Some experts made all the ratings at one sitting, while others rated the firms a week later. The judgments were the same for 89 percent of the ratings.

- **Valid relationships are used in the model.**

Bootstrapping is more useful when the expert=s judgments are valid, which occurs when the expert receives good feedback. If the experts use the wrong factors or incorrect relationships, their bootstrapping models will be of limited value and may produce less accurate forecasts than the experts. Their models would be applying the wrong rules more consistently.

Ganzach, Kluger and Klayman (2000), in their study of conscripts to the Israeli military, found that "independence" was positively related to the experts' global ratings on the probability of "success." However, one would not expect this to be related to the criterion they use for validation, which was disciplinary problems. The bootstrapping model consistently applied the wrong rule in this case.

- **The alternative is to use unskilled individual judgments.**

The bootstrapping model is perfectly consistent; given the same information about a situation, it will always produce the same forecast. As a result, it should be especially useful in comparison with unaided judgmental forecasts that are inconsistent. This often occurs for unskilled forecasters making individual judgments

As noted by Stewart (2001), forecasting skill depends on many things. Perhaps most important is that the expert needs well-summarized and accurate feedback. Without it, experts may be unskilled even after working in an area for two decades. When the experts are highly skilled, there is less potential for bootstrapping to help.

Group judgments are more accurate than those of the typical member. Part of the gain can be attributed to improvements in consistency. Thus, bootstrapping is likely to have less value when it is based on well-structured group processes (such as Delphi). Still, Dawes (1971), in his study on graduate admissions, found that bootstrapping improved accuracy when he developed it using the average group ratings, where the median number of raters was five. He found that bootstrapping was more accurate than the group average.

Conditions Favoring Bootstrapping over Econometric Methods

Bootstrapping offers advantages relative to econometric methods when no data are available on the criterion (the dependent variable) and causal variables have displayed little historical variation.

- **No criterion data (or lack of variation).**

When data are available for the dependent variable, one would expect that an econometric model would be more accurate than a bootstrapping model. (This assumes that there is much variation in the dependent variable.) After all, knowing what actually happened should be more informative than merely knowing what

was forecasted to happen. Ashton, Ashton and Davis (1994), in their study on predicting advertising pages, found an econometric model to be more accurate than their bootstrapping models.

Bootstrapping allows one to develop a model when no actual data exist for the dependent variable. Examples include predicting the success of new products, the effects of treatments for someone with lower back pain, the results of proposed organizational changes, or the outcomes of new government social programs. In such cases, analysts can create simulated data and ask experts to make predictions for these artificial cases.

- **No data on the causal variables (or lack of variation).**

If there are no data on the causal variables, regression analysis is of no value. If the data for a causal variable did not vary, a regression analysis will conclude that the variable has no effect. For example, if the price of a product has remained constant over a long period, statistical analyses of the data would show that price is statistically insignificant. Bootstrapping offers a way around this problem because the analyst can create artificial cases in which price varies substantially. This can be done using an experimental design to ensure large uncorrelated variations in the causal variables. For example, this would allow one to infer price elasticity from the sales predictions that experts make for these situations. While this procedure is a standard feature of conjoint analysis, it has not been examined in the bootstrapping literature.

LIMITATIONS

Bootstrapping has been used primarily for cross-sectional prediction problems. There has been little study of its use with time-series data.

Bootstrapping models could be expected to do poorly when encountering what Paul Meehl referred to as “broken leg cues;” that is, cases where the future goes beyond the experience of the model. For example, in looking at the characteristics of a race horse before betting, knowing it had a broken leg would be important. If such a variable were not included, the model would do poorly. (Of course, a good analyst would have provided a variable to represent the horse=s health.) In contrast, broken leg cues might be obvious to a person looking over the field. One suggestion is to use the model as long as no substantial changes have occurred. When relevant factors not included in the model change, the analyst could override the model or reformulate it. Although it seems obvious that bootstrapping will be less successful if sudden and large changes occur, no researchers have found this problem to be serious.

IMPLICATIONS FOR PRACTITIONERS

Bootstrapping is inexpensive when many forecasts are needed. It also aids learning. With bootstrapping, an analyst can formulate experiments. Nevertheless, it suffers from problems with acceptability.

Inexpensive and Rapid Forecasts

Bootstrapping models are inexpensive to develop compared to other types of expert systems. Once developed, bootstrapping models are inexpensive to use. Thus, bootstrapping is especially cost-effective when an expert must make many forecasts, as in situations faced by lawyers, stockbrokers, and university administrators. For example, Johnson (1988) describes the process for selecting interns for hospitals. Twelve physicians examined the folders for 200 applicants, a task that required about eight minutes each, after which they participated in two all-day sessions to select the interns. This represents an investment of about 64 physician-days. To obtain these forecasts from a bootstrapping model, one would need less than a day for a clerk to enter the data.

Aid to Learning

Experience often fails to provide people with adequate feedback about their forecasts. For example, Roose and Doherty (1976) found that the more experienced of 16 selectors were no more accurate in the selection of successful

new employees than were the less experienced 16 selectors. The experienced personnel selectors were consistent but inaccurate. Thus, bootstrapping models should be useful for learning in situations where experts do not receive good feedback on the accuracy of their predictions.

Bootstrapping may highlight the use of invalid factors. Assume, for example, personnel selectors favor those who are tall and good looking, although these traits are not relevant to the job. A bootstrapping model of their predictions could make them aware of this and this could lead to improvements in their selection procedure.

In cases in which some experts are more accurate than others, bootstrapping can be used to make the best expert=s forecasts available to others. For example, Dougherty, Ebert and Callender (1986) found that one personnel interviewer was much more accurate than the other two although all three were highly experienced. By developing bootstrapping models for the most accurate interviewer, one might learn how to improve the accuracy of other experts.

Bootstrapping should aid learning when a system is complex, involving such things as feedback loops, time delays, and nonlinearities. In a simulation of an inventory-control system, Diehl and Sterman (1995), by bootstrapping subjects= decisions, showed that they ignored important information about pending supply.

In this paper, I have concentrated on inferring rules for judgmental forecasting. One could use the same procedure to infer the rules used in any forecasting method. This might lead to a better understanding of what complex models are doing, and it might allow for a complex model to be replaced by a simple one. I worked on a project in which a company was using a highly complex model to make market-share forecasts. We conducted a series of interviews with people at various levels in the organization. Despite the fact that top management strongly supported the use of the models and the consultants who supplied the program had conducted expensive training sessions on the use of the model, no one in the organization understood how the model produced forecasts. We (Armstrong and Shapiro 1974) developed a bootstrapping model by using the model=s forecasts and its inputs. The result was a simple model that predicted market share (M) as a function of advertising (A). The model, $M = 20.7 + 0.6A$, explained 98 percent of the variation in the predictions made by the complex model.

Creating Experiments

In contrast to econometric models, bootstrapping is not restricted by the limitations of the actual data. For example, an econometric model used to predict how advertising expenditures affect sales for an item for which advertising expenditures have been constant, would be unable to estimate the effect of advertising. With bootstrapping, one can create situations in which the causal variables fluctuate. One can use such experimental situations to estimate relationships. While promising, this experimental approach has yet to be tested.

Because of its consistency, bootstrapping is superior to unaided judgment when one is trying to assess the impact of various policies. In other words, a bootstrapping model holds the procedures constant when forecasting the effects of different policies. Management can ask what-if questions and generate forecasts. This is analogous to the use of conjoint analysis.

Bootstrapping with artificial data can serve as an alternative to conjoint analysis. Whereas conjoint analysis requires data from hundreds of prospective customers, bootstrapping can use forecasts from only five experts on how consumers are likely to respond.

For important problems with much uncertainty, one might use both bootstrapping *and* conjoint analysis. Combining estimates of relationships from conjoint analysis and bootstrapping would be appropriate to such problems as forecasting sales for new products.

Acceptability

Despite the favorable evidence and the low costs of bootstrapping, its adoption has been slow. Dawes (1979) discusses this issue and offers explanations for the resistance. He suggests that some resistance is based on technical challenges to the quality of the studies on bootstrapping. Then there are psychological objections. People have

difficulty believing that a model could be superior to unaided judgment for *their* predictions. “After all, the evidence refers to other people on other problems at some time in the past, so why would it be relevant for me?” Ashton, Ashton and Davis (1994) and Grove and Meehl (1996) discuss similar problems in using models to replace unaided judgment.

Resistance persists even for areas that have been directly studied, such as graduate school admissions. Dawes (1979) reports that few schools have adopted the procedure. They resist using not just bootstrapping but econometric models as well. Instead, they cling to methods with low predictive ability. For example, Milstein et al. (1980, 1981) found that personal interviews were worthless for predicting which applicants would be successful at the Yale School of Medicine. DeVaul et al. (1987) reported on a study at the University of Texas Medical School where they admitted 50 students from the 100 applicants scoring *lowest* on the MCAT and grade point average. These students had initially been rejected by all the medical schools to which they applied. As it later turned out, the four-year performance records of these students were no different from those of the top 50 applicants. One would think that these findings would motivate university admissions officers to seek alternate procedures for selecting graduate students. An anonymous colleague of mine suggested the following explanations: Perhaps the performance of students is so far below their capabilities that anything above a modest level is irrelevant as a predictive factor. Alternatively, perhaps the system is designed so that the least capable students will be successful.

Arkes, Dawes and Christensen (1986) found that acceptance of a decision aid does not rest heavily on whether it outperforms unaided judgment. It depends more on the forecaster=s *perceived* level of expertise. Those who believe that they have a high level of expertise are less likely to adopt decision aids than those who are unsure about their expertise.

Bootstrapping might serve as the first step in introducing objective forecasting models. Managers may not take kindly to suggestions that they can be replaced by a quantitative model. They might offer less resistance to a model that mimics their rules. Once they adopt such a model, the question then becomes whether it is possible to improve it, so they might then incorporate estimates from econometric studies.

To overcome resistance to the use of a bootstrapping model, one could ask decision makers whether they would be interested in an experiment to examine its value. As mentioned earlier, Ed Snider used an experiment to persuade the Philadelphia Flyers’ management team to accept bootstrapping. Sometimes, however, decision makers cannot imagine any information that would change their minds. In the late 1970s, I offered to conduct an experiment for the Wharton School’s admissions committee. The members of the faculty committee said that they were unable to imagine any experimental outcome that would lead them to adopt a bootstrapping model. By asking about this before doing a study, I avoided working on a hopeless case. In the 1970s, I tried to convince the Philadelphia Eagles to consider bootstrapping for improving their selection of football players. I am still waiting for them to call, and they are still making poor draft picks.

Despite resistance, some organizations use bootstrapping models. Martorelli (1981) describes their use for draft selections in hockey and football. Christal (1968) reported that bootstrapping has been used for officer promotions in the U.S. Air Force.

Ethical concerns have been raised about bootstrapping. For example, why should a graduate school reject an applicant based on low numerical scores, they ask. Sometimes even the developers of the models do not think they should be used. DeDombal (1984) developed a model that was more accurate than senior physicians in recommending treatment of abdominal pain. But he did not recommend the system because ?human well-being is involved,” apparently believing that it is better to deal with a physician.

In some ways, bootstrapping is more ethical. Because a bootstrapping model=s rules are revealed, a model cannot be accused of concealing a prejudice against certain individuals. Should arguments arise, they can focus on what factors should be considered and how they should be weighted. Thus, bootstrapping can help to ensure that decisions are being made fairly and consistently.

IMPLICATIONS FOR RESEARCHERS

Studies on the use of bootstrapping in organizations are needed. For example, are managers more likely to accept models if the models use their rules?

Studies on the operational aspects of bootstrapping would be useful. Researchers might focus on how many cases one should present to experts, how to design cases so that the experts= task is easy, and how to scale variables. We also need studies on the conditions under which bootstrapping will be most effective.

Would accuracy improve if forecasts from bootstrapping models were combined with those from other methods? Unaided judgment is expected to be valid but unreliable, while bootstrapping improves reliability but at a possible loss of validity. Little work has been done on combinations of bootstrapping forecasts with those from other methods. Ashton, Ashton and Davis (1994) compared an equally weighted average of forecasts from bootstrapping models and from an expert. They found no improvement over the accuracy of the bootstrapping forecasts alone. Given that bootstrapping models are generally more accurate than an expert, it might have helped to have weighted them more heavily in this study.

It might be useful to combine bootstrapping estimates of a parameter with those from econometric analyses. One would expect that bootstrapping would play a vital role in assessing relationships that cannot be studied with econometric models because of collinearity, lack of variation, lack of data, or simply because a previously ignored factor becomes important. In other words, bootstrapping could be used to estimate relationships that cannot be estimated with actual data.

CONCLUSIONS

Bootstrapping, a type of expert system, is limited in that it is based only on data that experts use. Furthermore, it applies only to studies in which an expert's rules are inferred by regression analysis.

Bootstrapping is of particular interest because it is simple and inexpensive, and because of its demonstrated predictive validity. Its accuracy, to a large extent, derives from its being more reliable than experts; It applies the experts= rules more consistently than the experts can.

Here are some principles for bootstrapping:

- Judgmental bootstrapping provides more accurate forecasts than unaided judgment, especially when the
  - prediction problem is complex,
  - bootstrapping relationships can be reliably estimated,
  - experts have valid knowledge about relationships, and the
  - alternative is to obtain forecasts from individual unskilled experts.
- Judgmental bootstrapping provides an alternative to econometric models when
  - no data are available on the dependent variable (or there is little variation), and
  - actual data on the causal variables display little historical variation.
- Judgmental bootstrapping aids learning about judgmental prediction by identifying biases and the use of inappropriate cues.

One of the more promising uses of bootstrapping is to develop models for situations in which there are no data with variations in the causal variables. This can be done by creating sets of data. With an experimental design, the

analyst can ensure large variations in the causal variables and can avoid intercorrelations among them. The model can be used to forecast the outcomes of alternative policies in a systematic way. Surprisingly, this procedure has yet to be tested.

By revealing the current forecasting process, bootstrapping can facilitate learning. It can also reveal areas of high uncertainty and identify areas where judgmental forecasting seems deficient.

The use of judgmental bootstrapping poses few risks. In the eleven validation studies to date, it has been more accurate than experts in eight, less accurate in one, and equally accurate in the remaining two. The gains in accuracy have typically been large. Researchers obtained these results even though their bootstrapping procedures sometimes departed from ideal practice.

REFERENCES

Abdel-Khalik, A. Rashad & K. M. El-Sheshai (1980), "Information choice and utilization in an experiment on default prediction," *Journal of Accounting Research*, 18, 325-342.

Allen, G. & R. Fildes (2001), "Econometric forecasting," in J. S. Armstrong (ed.) *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Arkes, H. R., R. M. Dawes & C. Christensen (1986), "Factors influencing the use of a decision rule in a probabilistic task," *Organizational Behavior and Human Decision Processes*, 37, 93-110.

Armstrong, J. S. (1985), *Long-Range Forecasting: From Crystal Ball to Computer*. New York: John Wiley. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1997), "Peer review for journals: Evidence on quality control, fairness, and innovation," *Science and Engineering Ethics*, 3, 63-84. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S., F. Collopy & M. Adya (2001), "Rule-based forecasting: Using judgment in time-series extrapolation," in J. S. Armstrong (ed.) *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. & A. Shapiro (1974), "Analyzing quantitative models," *Journal of Marketing*, 38, 61-66. Full text at hops.wharton.upenn.edu/forecast.

Ashton, A. H. (1985), "Does consensus imply accuracy in accounting studies of decision making"" *Accounting Review*, 60, 173-185.

Ashton, A. H., R. H. Ashton & M. N. Davis (1994), "White-collar robotics: Levering managerial decision making," *California Management Review*, 37, 83-109.

Bowman, E. H. (1963), "Consistency and optimality in managerial decision making," *Management Science*, 9, 310-321.

Camerer, C. (1981), "General conditions for the success of bootstrapping models," *Organizational Behavior and Human Performance*, 27, 411-422.

Christal, R. E. (1968), "Selecting a harem and other applications of the policy-capturing model," *Journal of Experimental Education*, 36 (Summer), 35-41.

Collopy, F., M. Adya & J. S. Armstrong & (2001), "Expert systems for forecasting," in J. S. Armstrong (ed.) *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Cook, R. L. & T. R. Stewart (1975), "A comparison of seven methods for obtaining subjective descriptions of judgmental policy," *Organizational Behavior and Human Performance*, 13, 31-45.

Dawes, R. M. (1971), "A case study of graduate admissions: Application of three principles of human decision making," *American Psychologist*, 26, 180-188.

Dawes, R. M. (1979), "The robust beauty of improper linear models in decision making," *American Psychologist*, 34, 571-582.

Dawes, R. M. & B. Corrigan (1974), "Linear models in decision making," *Psychological Bulletin*, 81, 95-106.

DeDombal, F. T. (1984), "Clinical decision making and the computer: Consultant, expert, or just another test"" *British Journal of Health Care Computing*, 1, 7-12.

DeVaul, R. A. et al. (1987), "Medical school performance of initially rejected students," *Journal of the American Medical Association*, 257 (Jan 2), 47-51.

Diehl, E. & J. D. Sterman (1995), "Effects of feedback complexity on dynamic decision making," *Organizational Behavior and Human Decision Processes*, 62, 198-215.

Dougherty, T. W., R. J. Ebert & J. C. Callender (1986), "Policy capturing in the employment interview," *Journal of Applied Psychology*, 71, 9-15.

Ebert, R. J. & T. E. Kruse (1978), "Bootstrapping the security analyst," *Journal of Applied Psychology*, 63, 110-119.

Einhorn, H. J., D. N. Kleinmuntz & B. Kleinmuntz (1979), "Linear regression and process-tracing models of judgment," *Psychological Review*, 86, 465-485.

Ganzach, Y., A. N. Kluger & N. Klayman (2000), "Making decisions from an interview: Expert measurement and mechanical combination," *Personnel Psychology*, 53, 1-20.

Goldberg, L. R. (1976), "Man vs. model of man: Just how conflicting is that evidence?" *Organizational Behavior and Human Performance*, 16, 13-22.

Goldberg, L. R. (1971), "Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process," *Organizational Behavior and Human Performance*, 6, 458-479.

Goldberg, L. R. (1970), "Man vs. model of man: A rationale, plus some evidence, for a method of improving on clinical inferences," *Psychological Bulletin*, 73, 422-432.

Goldberg, L. R. (1968), "Simple models or simple processes? Some research on clinical judgments," *American Psychologist*, 23, 483-496.

Grove, W. M. & P. E. Meehl (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.

Hamm, R. H. (1991), "Accuracy of alternative methods for describing expert's knowledge of multiple influence domains," *Bulletin of the Psychonomic Society*, 29, 553-556.

Heeler, R. M., M. J. Kearney & B. J. Mehaffey (1973), "Modeling supermarket product selection," *Journal of Marketing Research*, 10, 34-37.

Hogarth, R. M. (1978), "A note on aggregating opinions," *Organizational Behavior and Human Performance*, 21, 40-46.

Hughes, H. D. (1917), "An interesting seed corn experiment," *The Iowa Agriculturist*, 17, 424-5,428

Johnson, E. (1988), "Expertise and decision under uncertainty: Performance and process," in M. Chi, R. Glaser & M. Farr (eds.). *The Nature of Expertise.* Lawrence Erlbaum Associates.

Kleinmuntz, B. (1990), "Why we still use our heads instead of formulas: Toward an integrative approach," *Psychological Bulletin*, 107, 296-310.

Kunreuther, H. (1969), "Extensions of Bowman's theory on managerial decision-making," *Management Science*, 15, 415-439.

Libby, R. (1976), "Man versus model of man: The need for a non-linear model," *Organizational Behavior and Human Performance*, 16 1-12.

Libby, R. & R. K. Blashfield (1978), "Performance of a composite as a function of the number of judges," *Organizational Behavior and Human Performance*, 21, 121-129.

Martorelli, W.P. (1981), "Cowboy DP scouting avoids personnel fumbles," *Information Systems News*, (November 16).

McClain, J. O. (1972), "Decision modeling in case selection for medical utilization review," *Management Science*, 18, B706-B717.

Milstein, R. M. et al. (1981), "Admissions decisions and performance during medical school," *Journal of Medical Education*, 56, 77-82

Milstein, R. M. et al. (1980), "Prediction of interview ratings in a medical school admission process, *Journal of Medical Education*, 55, 451-453.

Moskowitz, H. (1974), "Regression models of behavior for managerial decision making," *Omega*, 2, 677-690.

Moskowitz, H. & J. G. Miller (1972), "Man, models of man or mathematical models for managerial decision making" *Proceedings of the American Institute for Decision Sciences.* New Orleans, pp. 849-856.

Moskowitz, H., D. L. Weiss, K. K. Cheng & D. J. Reibstein (1982) "Robustness of linear models in dynamic multivariate predictions," *Omega*, 10, 647-661.

Roebber, P. J. & L. F. Bosart (1996), "The contributions of education and experience to forecast skill," *Weather and Forecasting*, 11, 21-40.

Roose, J. E. & M. E. Doherty (1976), "Judgment theory applied to the selection of life insurance salesmen," *Organizational Behavior and Human Performance*, 16, 231-249.

Schmitt, N. (1978), "Comparisons of subjective and objective weighting strategies in changing task situations," *Organizational Behavior and Human Performance*, 21, 171-188.

Schneidman, E. S. (1971), "Perturbation and lethality as precursors of suicide in a gifted group," *Life-threatening Behavior*, 1, 23-45.

Simester, D. & R. Brodie (1993), "Forecasting criminal sentencing decisions," *International Journal of Forecasting*, 9, 49-60.

Slovic, P., D. Fleissner & W. S. Bauman (1972), "Analyzing the use of information in investment decision making: A methodological proposal," *Journal of Business*, 45, 283-301.

Stewart, T. R. (2001), "Improving reliability of judgmental forecasts," in J. S. Armstrong (ed.) *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Taylor, F. W. (1911), *Principles of Scientific Management*. New York: Harper and Row.

Wallace, H. A. (1923), "What is in the corn judge's mind?" *Journal of the American Society of Agronomy*, 15 (7), 300-304.

Werner, P. D., T. L. Rose, J. A. Yesavage & K. Seeman (1984), "Psychiatrists' judgments of dangerousness in patients on an acute care unit," *American Journal of Psychiatry*, 141, No. 2, 263-266.

Wiggins, N. & E. Kohen (1971), "Man vs. model of man revisited: The forecasting of graduate school success," *Journal of Personality and Social Psychology*, 19, 100-106.

Wiggins, N. & P. J. Hoffman (1968), "Three models of clinical judgment," *Journal of Abnormal Psychology*, 73, 70-77.

Wittink, D. R. & T. Bergestuen (2001), "Forecasting with conjoint analysis," in J. S. Armstrong (ed.) *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Yntema, D. B. & W.S. Torgerson (1961), "Man-computer cooperation in decisions requiring common sense," *IRE Transactions of the Professional Group on Human Factors in Electronic*. Reprinted in W. Edwards & A. Tversky (eds.) (1967), *Decision Making*. Baltimore: Penguin Books, 300-314.

December 16, 2009

Acknowledgments: P. Geoffrey Allen, Fred Collopy, Ping Lin and Dick R. Wittink suggested extensive revisions. Monica Adya, Robin Dawes, Ronald J. Ebert, Lewis R. Goldberg, Stephen Hoch, Howard Kunreuther, John Mowen, Marcus O'Connor, Bill Remus and George Wright provided useful comments on early versions. Editorial assistance was provided by Raphael Austin, Natasha Miller, Ling Qiu and Mariam Rafi.