≠

DEMAND FORECASTING: EVIDENCE-BASED METHODS

Kesten C. Green[1]

J. Scott Armstrong[2]

October 2012
Version 165

[1] International Graduate School of Business, University of South Australia, City West Campus, North Terrace, Adelaide, SA 5000, Australia, T: +61 8 8302 9097 F: +61 8 8302 0709 kesten.green@unisa.edu.au

[2] The Wharton School, University of Pennsylvania, 747 Huntsman, Philadelphia, PA 19104, U.S.A. T: +1 610 622 6480 F: +1 215 898 2534 armstrong@wharton.upenn.edu

ABSTRACT

In recent decades, much comparative testing has been conducted to determine which forecasting methods are more effective under given conditions. This evidence-based approach leads to conclusions that differ substantially from current practice, . This paper summarizes the primary findings on what to do – and what not to do. When quantitative data are scarce, impose structure by using expert surveys, intentions surveys, judgmental bootstrapping, prediction markets, structured analogies, and simulated interaction. When quantitative data are abundant, use extrapolation, quantitative analogies, rule-based forecasting, and causal methods. Among causal methods, use econometrics when prior knowledge is strong, data are reliable, and few variables are important. When there are many important variables and extensive knowledge, use index models. Use structured methods to incorporate prior knowledge from experiments and experts' domain knowledge as inputs to causal forecasts. Combine forecasts from different forecasters and methods. *Avoid* methods that are complex, that have not been validated, and that ignore domain knowledge; these include intuition, unstructured meetings, game theory, focus groups, neural networks, stepwise regression, and data mining.

Keywords: checklist, competitor behavior, forecast accuracy, market share, market size, sales forecasting.

Demand forecasting asks how much can be sold given the situation? The situation includes the broader economy, social and legal issues, and the nature of sellers, buyers, and the market. The situation also includes actions by the firm, its competitors, and interest groups.

Demand forecasting knowledge has advanced in the way that science always advances: through accumulation of evidence from experiments that test multiple reasonable hypotheses (Armstrong 2003). Chamberlin was perhaps the first to describe this method, by which he hoped that "the dangers of parental affection for a favorite theory can be circumvented" (1890; p. 754, 1965). The evidence-based approach led to the agricultural and industrial revolutions that are responsible for our current prosperity (Kealey 1996), and to the more recent enormous progress in medicine (Gratzer 2006). From the evidence of progress in those fields, Chamberlin's optimistic 1890 conclusion that "…one of the greatest moral reforms that lies immediately before us consists in the general introduction into social and civic life of… the method of multiple working hypotheses" (p. 759) was partly born out.

Despite the impressive results in other fields, however, management researchers have largely ignored this evidence-based approach. Few conduct experiments to test multiple reasonable hypotheses. For example, fewer than 3% of the 1,100 empirical articles in a study on marketing publications involved such tests and many of those few paid little attention to conditions (Armstrong, Brodie, and Parsons 2001).

In medicine, a failure to follow evidence-based procedures can be the basis of expensive lawsuits. The idea that practitioners should follow evidence-based procedures is less developed in business and government. Consider, for example, the long obsession with statistical significance testing despite the evidence that it confuses people and harms their decision-making (Ziliak and McCloskey 2008).

The *Journal of Forecasting* was founded in 1981 on a belief that an evidence-based approach would lead to a more rapid development of the field. The approach met with immediate success. Almost 58% of the empirical papers published in the *Journal of Forecasting* (1982 to 1985) and the *International Journal of Forecasting* (1985-1987) used the method of multiple reasonable hypotheses. These findings compare favorably with the only 22% of empirical papers in *Management Science* that used the method of multiple hypotheses (Armstrong 1979) and the 25% from leading marketing journals (Armstrong, Brodie, and Parsons 2001). By 1983, the *Journal of Forecasting* had the second highest journal impact factor of all management journals.

In the mid-1990s, the forecasting principles project began by summarizing findings from experimental studies from all areas of forecasting. The project involved the collaborative efforts of 39 leading forecasting researchers from various disciplines, and was supported by 123 expert reviewers. The findings were summarized as principles (condition-action steps). That is, under what conditions is a method effective? One-hundred-and-thirty-nine principles were formulated, They were published in Armstrong (2001, pp 679-732).

This article summarizes the substantial progress in demand forecasting by first describing evidence-based methods and then describing principles for selecting the best methods for demand forecasting problems and conditions. It summarizes procedures to improve forecasts by combining, adjusting, and communicating uncertainty. Finally, it describes procedures to ease the implementation of new methods.

Forecasting Methods

Demand forecasters can draw upon many methods. These methods can be grouped into 17 categories. Twelve rely on judgment, namely unaided judgment, decomposition, expert surveys,

structured analogies, game theory, judgmental bootstrapping, intentions and expectations surveys, simulated interaction, conjoint analysis, experimentation, prediction markets, and expert systems. The remaining five methods require quantitative data. They are extrapolation, quantitative analogies, causal models, neural nets, and rule-based forecasting. Additional information on the methods is available in *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Armstrong 2001).

Methods that rely primarily on judgment

Unaided judgment

Expert's judgments are convenient for many demand forecasting tasks such as forecasting sales of new products, effects of changes in design, pricing, or advertising, and competitor behavior. Experts' unaided judgments can provide useful forecasts if the experts make many forecasts about similar situations that are well understood and they receive good feedback that allows them to learn. Most demand forecasting tasks are not of this kind, however.

When, as is often the case, the situations that are faced by demand forecasters are uncertain and complex, experts' judgments are of little value (Armstrong 1980). Few people are aware of this. When told about it most people are sure that the findings do not apply to them. Indeed, companies often pay handsomely for such expert forecasts. Thus it has been labeled the Seer-sucker Theory: "No matter how much evidence exists that seers do not exist, suckers will pay for the existence of seers". In a recent test of this theory, subjects were willing to pay for sealed-envelop predictions of the outcome of the next toss of a sequence of fair coin tosses. Their willingness to pay and the size of their bets increased with the number of correct predictions (Powdthavee and Riyanto 2012).

In a 20-year experiment on the value of judgmental forecasts, 284 experts made more than 82,000 forecasts about complex and uncertain situations over short and long time horizons. Forecasts related to, for example, GDP growth and health and education spending for different nations. Their forecasts turned out to be little more accurate than those made by non-experts, and they were less accurate than forecasts from simple models (Tetlock 2005).

Experts are also inconsistent in their judgmental forecasts about complex and uncertain situations. For example, when seven software professionals estimated the development effort required for six software development projects a month or more after having first been asked to do so, their estimates had a median difference of 50% (Grimstad and Jørgensen 2007). SEEMS OUT OF PLACE HERE>

Judgmental Decomposition

Judgmental decomposition involves dividing a forecasting problem into multiplicative parts. For example, to forecast sales for a brand, a firm might separately forecast total market sales and market share, and then multiply those components. Decomposition makes sense when deriving forecasts for the parts is easier than for the whole problem and when different methods are appropriate for forecasting each part.

Forecasts from decomposition are generally more accurate than those obtained using a global approach. In particular, decomposition is more accurate when the aggregate forecast is highly uncertain and when large numbers (over one million) are involved. In three studies involving 15 tests, judgmental decomposition led to a 42% error reduction when uncertainty about the situation was high (MacGregor 2001).

Expert surveys

Experts often have knowledge about how others might behave. To gather this knowledge,, use written questions in order to ensure that each question is asked in the same way of all experts. This also helps to avoid interviewers' biases. Avoid revealing expectations that might anchor the experts' forecasts. For example, knowledge of customers' expectations of 14 projects' costs had very large effects on eight experts' forecasts—they were eight times higher when customer expectation were high than when they were low—even when the experts were warned to ignore the expectations due to their lack of validity (Jørgensen and Sjøberg 2004). Word the questions in different ways to compensate for possible biases in wording and pre-test all questions. Dillman, Smyth, and Christian (2009) provide advice on questionnaire design.

The *Delphi technique* provides a useful way to obtain expert forecasts from diverse experts while avoiding the disadvantages of traditional group meetings. Delphi is likely to be most effective in situations where relevant knowledge is distributed among experts. For example, decisions regarding where to locate a retail outlet would benefit from forecasts obtained from experts on real estate, traffic, retailing, consumers, and on the area to be serviced.

To forecast with Delphi, select between five and twenty experts diverse in their knowledge of the situation. Ask the experts to provide forecasts and reasons for their forecasts, then provide them with anonymous summary statistics on the panels' forecasts and reasons. Repeat the process until forecasts change little between rounds—two or three rounds are usually sufficient. The median or mode of the experts' final-round forecasts is the Delphi forecast. Software to help administer the procedure is available at forecastingprinciples.com.

Delphi forecasts were more accurate than those from traditional meetings in five studies, less accurate in one, and equivocal in two (Rowe and Wright 2001). Delphi was more accurate than expert surveys for 12 of 16 studies, with two ties and two cases in which Delphi was less accurate. Among these 24 comparisons, Delphi improved accuracy in 71% and harmed accuracy in 12%.

Delphi is attractive to managers because it is easy to understand and the record of the experts' reasoning is informative and it provides credibility. Delphi is relatively cheap because the experts do not meet. Delphi's advantages over prediction markets include (1) broader applicability, (2) ability to address complex questions, (3) ability to maintain confidentiality, (4) avoidance of manipulation, (5) revelation of new knowledge, and (6) avoidance of cascades. Points 5 and 6 refer to the fact that whereas the Delphi process requires participants to share their knowledge and reasoning and to respond to that of others, prediction markets' participants do not exchange qualitative information (Green, Armstrong, and Graefe 2007). In addition, one study found that Delphi was more accurate than prediction markets. Participants were more favorably disposed toward Delphi (Graefe and Armstrong, 2011).

Structured analogies

The structured analogies method is a formal, unbiased process for gathering information about similar situations and processing that information to make forecasts. The method should not be confused with the informal use of analogies to justify forecasts obtained by other means.

To use structured analogies, prepare a description of the situation for which forecasts are required (the target situation) and select experts who are likely to be familiar with analogous situations, preferably from direct experience. Instruct the experts to identify and describe analogous situations, rate

their similarity to the target situation, and match the outcomes of their analogies with potential outcomes of the target situation. Take the outcome of each expert's top-rated analogy, and use a median or mode of these as the structured analogies forecast.

The research to date on structured analogies is limited but promising. Structured analogies were 41% more accurate than unaided judgment in forecasting decisions in eight real conflicts. Conflicts used in the research that are relevant to the wider problem of demand forecasting include union-management disputes, a hostile takeover attempt, and a supply channel negotiation (Green and Armstrong 2007). A procedure akin to structured analogies was used to forecast box office revenue for 19 unreleased movies (Lovallo, Clarke, and Camerer 2012). Raters identified analogous movies from a database and rated them for similarity. The revenue forecasts from the analogies were adjusted for advertising expenditure, and if the movie was a sequel. Errors from the structured analogies based forecasts were less than half those of forecasts from a simple regression model, and those from a complex one. Structured analogies is easily implemented and understood, and can be adapted for diverse forecasting problems.

Game theory

Game theory involves identifying the incentives that motivate parties and deducing the decisions they will make. This sounds plausible, and the authors of textbooks and research papers recommend game theory to make forecasts about conflicts such as those that occur in oligopoly markets. However, there is no evidence to support this viewpoint. In the only test of forecast validity to date, game theory experts' forecasts of the decisions that would be made in eight real conflict situations were no more accurate than students' unaided judgment forecasts (Green 2002 and 2005). Based on the evidence to date, then, we recommend against the use of game theory for demand forecasting.

Judgmental bootstrapping

Judgmental bootstrapping estimates a forecasting model from experts' judgments. The first step is to ask experts what information they use to make predictions about a class of situations. Then ask them to make predictions for a set of real or hypothetical cases. Hypothetical situations are preferable, because the analyst can design the situations so that the independent variables vary substantially and do so independently of one another. For example, experts, working independently, might forecast first year sales for proposed new stores using information about proximity of competing stores, size of the local population, and traffic flows. These variables are used in a regression model that is estimated from the data used by the experts, and where the dependent variable is the *expert's forecast.*

Judgmental bootstrapping models are most useful for repetitive, complex forecasting problems for which data on the dependent variable are not available (e.g. demand for a new product) or where the available data on the causal variable do not vary sufficiently to allow the estimation of regression coefficients. For example, it was used to estimate demand for advertising space in *Time* magazine. Once developed, judgmental bootstrapping models can provide forecasts that are less expensive than those provided by experts.

A meta-analysis found that the judgmental bootstrapping forecasts were more accurate than those from unaided judgment in 8 of the 11 comparisons, with two tests showing no difference and one showing a small loss (Armstrong 2006) [Any more recent studies?? The typical error reduction was about 6%. The one failure occurred when the experts relied heavily on an erroneous variable. In other

words, when judges use a variable that lacks predictive validity—such as the country of origin—consistency is likely to harm accuracy.

Intentions and expectations surveys

Intentions surveys ask people how they *intend* to behave in specified situations. The data collected can be used, for example, to predict how people would respond to major changes in the design of a product. A meta-analysis covering 47 comparisons with over 10,000 subjects finds a strong relationship between people's intentions and their behavior (Kim and Hunter 1993). Sheeran (2002) reaches the same conclusion with his meta-analysis of ten meta-analyses with data from over 83,000 subjects.

Surveys can also be used to ask people how they *expect* they would behave. Expectations differ from intentions because people know that unintended things happen. For example, if you were asked whether you intended to visit the dentist in the next six months you might say no. However, you realize that a problem might arise that would necessitate a visit, so your expectation would be that visiting the dentist in the next six months had a probability greater than zero.

To forecast demand using a survey of potential consumers, prepare an accurate and comprehensive description of the product and conditions of sale. Expectations and intentions can be obtained using probability scales such as 0 = 'No chance, or almost no chance (1 in 100)' to 10 = 'Certain, or practically certain (99 in 100)'. Evidence-based procedures for selecting samples, obtaining high response rates, compensating for non-response bias, and reducing response error are described in Dillman, Smyth, and Christian (2009). Response error is often a large component of error. This problem is especially acute when the situation is new to the people responding to the survey, as would be the case for questions about a new product. Intentions data provide unbiased forecasts of demand, so no adjustment is needed for response bias (Wright and MacRae 2007).

Intentions and expectations surveys are useful when historical demand data are not available, such as for new product forecasts or for a new market. They are most likely to be useful in cases where survey respondents have had relevant experience. Other conditions favoring the use of surveys of potential customers include: (1) the behavior is important to the respondent, (2) the behavior is planned, (3) the respondent is able to fulfill the plan, and (4) the plan is unlikely to change (Morwitz 2001).

Focus groups have been proposed to forecasts customers' behavior. However, there is no evidence to support this approach for demand forecasting, Furthermore, the approach violates important forecasting principles. First, the participants are seldom representative of the population of interest. Second, they use small samples. Third, in practice, questions for the participants are often not well structured or well tested. Fourth, in summarizing the responses of focus group participants, subjectivity and bias are difficult to avoid. Fifth, and most important, the responses of participants are influenced by the presence and expressed opinions of others in the group.

Simulated interaction

Simulated interaction is a form of role-playing that can be used to forecast decisions by people who are interacting. For example, a manager might want to know how best to secure an exclusive distribution arrangement with a major supplier, how a competitor would respond to a proposed sale, or how important customers would respond to possible changes in the design of a product.

Simulated interactions can be conducted inexpensively by using students to play the roles. Describe the main protagonists' roles, prepare a brief description of the situation, and list possible

decisions. Participants adopt a role, then read the situation description. They engage in realistic interactions with the other role players, staying in their roles until they reach a decision. Simulations typically last between 30 and 60 minutes.

Relative to the usual forecasting method of unaided expert judgment, simulated interaction reduced forecast errors by 57% for eight conflict situations (Green 2005). These were the same situations as for structured analogies (described above), where the error reduction was 41%

If the simulated interaction method seems onerous, you might think that following the common advice to put yourself in the other person's shoes would help a clever person such as yourself to predict decisions. For example, Secretary of Defense Robert McNamara said that if he had done this during the Vietnam War, he would have made better decisions.[3] He was wrong: A test of "role thinking" by the authors found no improvement in the accuracy of the forecasts (Green and Armstrong 2011). Apparently, thinking through the interactions of parties with divergent roles in a complex situation is too difficult; active role-playing between parties is necessary to represent such situations with sufficient realism to derive useful forecasts

Conjoint Analysis

Conjoint analysis can be used to examine how demand varies as important features of a product are varied. Potential customers are asked to make selections from a set of offers such as 20 different designs of a product. For example, various features of a tablet computer such as price, weight, dimensions, software features, communications options, battery life, and screen clarity could be varied substantially while ensuring that the variations in features do not correlate with one another. The potential customer chooses from among various offerings. The resulting data can be analyzed by regressing respondents' choices against the product features.

Conjoint analysis is based on sound principles, such as using experimental design and soliciting independent intentions from a representative sample of potential customers. So it should be useful. However, despite a large academic literature and widespread use by industry, experimental comparisons of conjoint-analysis with other reasonable methods are scarce (Wittink and Bergestuen 2001). In an experiment involving 518 subjects making purchase decisions about chocolate bars, conjoint analysis led to forecasts of willingness to pay that were between 70% and 180% higher than those that were obtained using a lottery that was designed to elicit true willingness to pay figures (Sichtmann, Wilken, Diamantopoulos 2011). In this context, users of conjoint analysis should consider conducting their own experiments to compare the accuracy of the conjoint analysis forecasts with those from methods.

Experimentation

Experimentation is widely used and is the most realistic method for forecasting the effects of alternative courses of action. Experiments can be used to examine how people respond to such things as a change in the design of a product or to changes in the marketing of a product. For example, how would people respond to changes in the automatic answering systems used for telephone inquiries? Trials could be conducted in some regions but not others. Alternatively, different subjects might be exposed to different telephone systems in a laboratory experiment.

[3] From the documentary film, "Fog of War."

Laboratory experiments allow greater control, testing of conditions is easier, costs are usually lower, and they avoid revealing sensitive information to competitors. A lab experiment might involve testing consumers' relative preferences by presenting a product in different packaging, and recording their purchases in a mock retail environment. A field experiment might involve, for example, charging different prices in different geographical markets to estimate the effects on total revenue. Researchers sometimes argue over the relative merits of laboratory and field experiments. An analysis of experiments in organizational behavior found that the two approaches yielded similar findings (Locke 1986).

Prediction markets

Prediction markets—which are also known as betting markets, information markets, and futures markets—have been used to make forecasts since the 1800s. Prediction markets can be created to predict such things as the proportion of U.S. households with three or more vehicles by the end of 2015. Confidential markets can be established within firms to motivate employees to reveal their knowledge, as forecasts, by buying and selling contracts that reward accuracy. Forecasting first year sales of a new product is one possible application. Prediction markets are likely to be superior to unstructured meetings because they efficiently aggregate the dispersed information of anonymous self-selected experts. However, this applies to the use of any structured approach. For example the second author was invited to a meeting at a consumer products company in Thailand in which a new advertising campaign was being proposed. The company's official forecast was for a substantial increase in sales. The author asked the 20 managers in the meeting for their anonymous forecasts along with 95% confidence intervals. None of the mangers forecast an appreciable increase in sales. The official forecast was greater than the 95% confidence intervals of all of the mangers.

Some unpublished studies suggest that prediction markets can produce accurate sales forecasts. Despite the promise, the average improvement in accuracy across eight published comparisons in the field of business forecasting—relative to forecasts from, variously, naïve models, econometric models, individual judgment, and statistical groups—is mixed. While the error reductions range from +28% (relative to naïve models) to -29% (relative to average judgmental forecasts), the comparisons were insufficient to provide guidance on the conditions that favor prediction markets (Graefe 2011). Nevertheless, without strong findings to the contrary and with good reasons to expect some improvement, when knowledge is dispersed and a sufficient number of motivated participants are trading, assume that prediction markets will improve accuracy relative to unaided group forecasts.

Expert systems

Expert systems are codifications of the rules experts use to make forecasts for a specific product or situation. An expert system should be simple, clear, and complete. To identify the rules, record experts' descriptions of their thinking as they make forecasts. Use empirical estimates of relationships from econometric studies and experiments when available in order to ensure that the rules are sound. Conjoint analysis, and bootstrapping can also provide useful information.

Expert system forecasts were more accurate than those from unaided judgment in a review of 15 comparisons (Collopy, Adya and Armstrong 2001). Two of the studies, on gas and mail order catalogue sales, involved forecasting demand. The expert systems error reductions were 10% and 5% respectively in comparison with unaided judgment. Given the small effects, limited evidence, and the

complexity of experts systems, it would be premature to recommend expert systems for demand forecasting.

Methods requiring quantitative data

Extrapolation

Extrapolation methods require historical data only on the variable to be forecast. They are appropriate when little is known about the factors affecting a variable to be forecast. Statistical extrapolations are cost effective when many forecasts are needed. For example, some firms need frequent forecasts of demand for each of hundreds of inventory items.

Perhaps the most widely used extrapolation method, with the possible exception of using last year's value, is exponential smoothing. Exponential smoothing is sensible in that recent data are weighted more heavily and, as a type of moving average, the procedure smoothes out short-term fluctuations. Exponential smoothing is understandable, inexpensive, and relatively accurate. Gardner (2006) provides a review of the state-of-the-art on exponential smoothing.

When extrapolation procedures do not use information about causal factors, uncertainty can be high, especially about the long-term. The proper way to deal with uncertainty is to be conservative. For time series, conservatism requires that estimates of trend be damped toward no change: The greater the uncertainty about the situation, the greater the damping that is needed. Procedures are available to damp the trend and some software packages allow for damping. A review of ten comparisons found that, on average, damping reduced forecast error by almost 5% when used with exponential smoothing (Armstrong 2006). In addition, damping reduces the risk of large errors and can moderate the effects of recessions. Avoid software that does not provide proper procedures for damping.

When extrapolating data of greater than annual frequency, remove the effects of seasonal influences first. Seasonality adjustments lead to substantial gains in accuracy, as was shown in a large-scale study of time-series forecasting: In forecasts over an 18-month horizon for 68 monthly economic series, they reduced forecast errors by 23 percent (Makridakis, *et al*. 1984, Table 14).

Because seasonal factors are estimated, rather than known, they should be damped. Miller and Williams (2003, 2004) provide procedures for damping seasonal factors. Their software for calculating damped seasonal adjustment factors is available at forecastingprinciples.com. When they applied the procedures to the 1,428 monthly time series from the M3-Competition, forecast accuracy improved for 68% of the series. In another study, damped seasonal estimates were obtained by averaging estimates for a given series with seasonal factors estimated for related products. This damping reduced forecast error by about 20% (Bunn and Vassilopoulos 1999).

One promising extrapolation approach is to decompose time series by causal forces. This is expected to improve accuracy when a time series can be effectively decomposed under two conditions: (1) if domain knowledge can be used to structure the problem so that causal forces differ for two or more component series, and (2) when it is possible to obtain relatively accurate forecasts for each component. For example, to forecast the number of people that will die on the highways each year, forecast the number of passenger miles driven (a series that is expected to grow), and the death rate per million passenger miles (a series expected to decrease), then multiply these forecasts. When tested on five time series that clearly met the conditions, decomposition by causal forces reduced forecast errors by two-thirds. For the four series that partially met the conditions, decomposition by causal forces reduced errors by one-half. Although the gains in accuracy were large, to date there is only the one study on decomposition by causal forces (Armstrong, Collopy and Yokum 2005).

For many years Box-Jenkins was the favored extrapolation procedure among statisticians and it was admired for its rigor. Unfortunately, there are two problems: First, it is difficult for reasonably intelligent human beings to understand. And, second, studies of comparative accuracy found that Box-Jenkins does not improve accuracy (e.g. Makridakis, *et al.*, 1984).

Quantitative analogies

When few data are available on the item being forecast, data from analogous situations can be used to extrapolate what will happen. For example, in order to assess the annual percentage loss in sales when the patent protection for a drug is removed, one might examine the historical pattern of sales when patents were removed for similar drugs in similar markets.

To forecast using quantitative analogies, ask experts to identify situations that are analogous to the target situation and for which data are available. Analogous data may be as directly relevant as, for example, previous per capita ticket sales for a play that is touring from city to city.

It often helps to combine data across analogous situations. Pooling monthly seasonal factors for crime rates for six precincts of a city increased forecast accuracy by 7% compared to when seasonal factors were estimated individually for each precinct (Gorr, Oligschlager, and Thompson 2003). Forecasts of 35 software development project costs from four automated analogy selection procedures were 2% more accurate than forecasts from four atheoretical statistical models (Li, Xie, and Goh 2009). The analogies-based forecasts were 11% more accurate than those from the neural networks models alone.

Causal Models

Causal models include models derived using segmentation, regression analysis, and the index method. These methods are useful if knowledge and data are available for variables that might affect the situation of interest. For situations in which large changes are expected, forecasts from causal models are more accurate than forecasts derived from extrapolating the dependent variable (Armstrong 1985, pp. 408-9; Allen and Fildes 2001). Theory, prior research, and expert domain knowledge provide information about relationships between explanatory variables and the variable to be forecast. The models can be used to forecast the effects of different policies.

Causal models are most useful when (1) strong causal relationships exist, (2) the directions of the relationships are known, (3) large changes in the causal variables are expected over the forecast horizon, and (4) the causal variables can be accurately forecast or controlled, especially with respect to their direction.

Segmentation involves breaking a problem down into independent parts of the same kind, using knowledge and data to make a forecast about each part, and combining the forecasts of the parts. For example, a hardware company could forecast industry sales for each type of product and then add the forecasts.

To forecast using segmentation, identify important causal variables that can be used to define the segments, and their priorities. Determine cut-points for each variable such that the stronger the relationship with the dependent variable, the greater the non-linearity in the relationship, and the more data that are available the more cut-points that should be used. Forecast the population of each segment and the behavior of the population within each segment then combine the population and behavior forecasts for each segment, and sum the segments.

Segmentation has advantages over regression analysis where variables interact, the effects of variables on demand are non-linear, and clear causal priorities exist. Segmentation is especially useful when errors in segment forecasts are likely to be in different directions. This situation is likely to occur where the segments are independent and of roughly equal importance, and when information on each segment is good. For example, one might improve accuracy by forecasting demand for the products of each division of a company separately, then adding the forecasts. But if segments have only small samples and erratic data, the segment forecasts might include large errors (Armstrong 1985, pp. 412-420).

Segmentation based on *a priori* selection of variables offers the possibility of improved accuracy at a low risk. Experts prefer segmentation's bottom-up approach as the approach allows them to use their knowledge about the problem effectively (Jørgensen 2004). Bottom-up forecasting produced forecasts that were more accurate than those from top-down forecasting for 74% of 192 monthly time-series (Dangerfield and Morris 1992). In a study involving seven teams making estimates of the time required to complete two software projects, the typical error from the bottom-up forecast was half of that for the top-down approach (Jørgensen 2004). Segments can be too small. For example, 40 students each predicted completion times for one composite and three small individual office tasks, and were then discretely timed completing the tasks. The individual tasks were completed in between 3 and 7 minutes on average. The forecast errors were biased towards overestimation and the absolute errors were twice the size of the errors from estimating the composite task (Forsyth and Burt 2008). The problem of overestimation did not arise when another group of 40 students made forecasts of the time to complete when the individual tasks were of longer durations; roughly 30 minutes. The bottom-up absolute forecast errors were 13% smaller than the top-down forecast errors.

Regression analysis is used to estimate the relationship between a dependent variable and one or more causal variables. Regression is typically used to estimate relationships from historical (non-experimental) data. Regression is likely to be useful in situations in which three or fewer causal variables are important, effect sizes are important, and effect sizes can be estimated from many reliable observations that include data in which the causal variables varied independently of one another (Armstrong 2012).

Important principles for developing regression models are to (1) use prior knowledge and theory, not statistical fit, for selecting variables and for specifying the directions of their effects, (2) discard variables if the estimated relationship conflicts with prior evidence on the nature of the relationship, and (3) keep the model simple in terms of the number of equations, number of variables, and the functional form (Armstrong 2012). Choose between theoretically sound models on the basis of out-of-sample accuracy, not on the basis of R^2. Unfortunately, the improper use of regression analysis seems to be increasing, thus producing misleading demand forecasts.

Because regression analysis tends to over-fit data, the coefficients used in the forecasting model should be damped toward no effect. This adjustment tends to improve out-of-sample forecast accuracy, particularly when one has small samples and many variables. As this situation is common for many prediction problems, unit (or equal weight) models—the most extreme case of damping—often yield more accurate forecasts than models with statistically fitted (un-damped) regression coefficients .

The *index method* is suitable for situations with little data on the variable to be forecast, where many causal variables are important, and where prior knowledge about the effects of the variables is good (Graefe and Armstrong, 2011). Use prior empirical evidence to identify predictor variables and to assess each variable's directional influence on the outcome. Experimental findings are especially valuable. Better yet, draw on findings from meta-analyses of experimental studies. If prior studies are not available, independent expert judgments can be used to choose the variables and determine the

directions of their effects. If prior knowledge on a variable's effect is ambiguous or contradictory, do not include the variable in the model.

Index scores are the sum of the values across the variables, which might be coded as 1 or 0 (favorable or unfavorable), depending on the state of knowledge. An alternative with a higher index score is more likely. Where sufficient historical data are available, by regressing index values against the variable of interest, such as sales, one can obtain quantitative forecasts

The index method is especially useful for selection problems, such as for assessing which geographical location offers the highest demand for a product. The method has been successfully tested for forecasting the outcomes of U.S. presidential elections based on information about candidates' biographies (Armstrong and Graefe 2011) and voters' perceptions of candidates' ability to handle the issues (Graefe and Armstrong 2012).

In general, avoid causal methods that lack theory or do not use prior knowledge. Data mining, step-wise regression, and neural networks are such methods. For example, data mining uses sophisticated statistical analyses to identify variables and relationships. Although data mining is popular, no evidence exists that the technique provides useful forecasts. An extensive review and reanalysis of 50 real-world data sets also finds little evidence that data mining is useful (Keogh and Kasetty 2002).

Neural nets

Neural networks are designed to pick up nonlinear patterns in long time-series. Studies on neural nets have been popular with researchers with more than 7,000 articles identified in an August 2012 Social Science Citation Index (Web of Knowledge) search for the topic of neural networks and forecasting. Early reviews on the accuracy of forecasts from neural nets were not favorable. However, Adya and Collopy (1998) found only eleven studies that met the criteria for a comparative evaluation, and in eight of these, neural net forecasts were more accurate than alternative methods. Tests of *ex ante* accuracy in forecasting 111 time series, however, found that neural network forecasts were about as accurate as forecasts from established extrapolation methods (Crone, Hibon, and Nikolopoulos 2011). Perhaps the fairest comparison has been the M3-Competition with 3,003 varied time series. In that study, neural net forecasts were 3.4% less accurate than damped trend-forecasts and 4.2% less accurate than combined extrapolations (Makridakis and Hibon 2000).

Given that neural nets ignore prior knowledge, the results are difficult to understand, and the evidence on accuracy is weak, demand forecasters are unlikely to benefit from using the method. Furthermore, with many studies published on neural nets, the published research might not properly reflect the true value of the method due to journals preference for statistically significant results. The situation is much like that for Box-Jenkins methods.

Rule-based forecasting

Rule-based forecasting, or RBF, allows an analyst to integrate evidence-based forecasting principles and managers' knowledge about the situation with time-series forecasts in a structured and inexpensive way. RBF is an evidence-based general-purpose expert system for forecasting time-series data.

To implement RBF, first identify the features of the series. There are 28 series features, including the causal forces (growth, opposing, regressing, supporting, or unknown) and such things as the length of the forecast horizon, the amount of data available, and the existence of outliers (Armstrong, Adya and

Collopy 2001). The features can be identified by inspection, statistical analysis, or domain knowledge. There are presently 99 rules for adjusting the data and estimating the starting value and the short- and long-range trends. RBF forecasts are a blend of the short- and long-range extrapolations For one-year ahead ex *ante* forecasts of 90 annual series, the median absolute percentage error for RBF forecasts were 13% smaller than those from equally weighted combined forecasts. For six-year ahead *ex ante* forecasts, the RBF forecast errors were 42% smaller. RBF forecasts were more accurate than equal-weights combined forecasts in situations involving significant trends, low uncertainty, stability, and good domain expertise. In cases where the conditions were not met, the RBF forecasts were no more accurate (Collopy and Armstrong 1992).

If implementing RBF is too big a step, consider the contrary series rule. The rule states that when the expected direction of a time-series and the historical trend of the series are contrary to one another, set the forecasted trend to zero. The rule yielded substantial improvements, especially for longer-term (6-year-ahead) forecasts where the error reduction exceeded 40% (Armstrong and Collopy 1993).

Matching methods with problems and conditions

Managers need forecasts of the total size of the relevant market. They also need forecasts of the actions and reactions of key decision makers such as competitors, suppliers, distributors, competitors, or government officials. Forecasts of these actions help to forecast market share. The resulting forecasts of market size and market share allow the calculation of a demand forecast. Selection of methods to match the conditions the demand forecaster is faced with—principally the type and quantity of data that is available and knowledge about the situation. Finally conditions that prevail when forecasting demand for new products are treated as a special case.

Forecasting market size

Market size is influenced by environmental factors. For example, the demand for alcoholic beverages will be influenced by such things as local climate, size and age distribution of the population, disposable income, laws, and culture.

Market forecasts for relatively new or rapidly changing markets in particular are often based on judgment. Given the risk of bias from unaided judgment, use structured methods. For example, the Delphi technique could be used to answer questions about market size such as: "By what percentage will the wine market grow over the next 10 years?" or "What proportion of households will watch movies via the Internet five years from now?"

When sufficient data are available, such as when the market is well established or when data on analogous markets or products are available, use time-series extrapolation methods or causal methods. Simple time-series extrapolation is inexpensive. Rule-based forecasting is more expensive, but less likely to produce large errors. Use causal methods, such as econometrics and segmentation, when the causal variables are known, large changes are expected in the causal variables, the direction of the change can be predicted accurately, and good knowledge exists about the effects of such changes.

Forecasting decision makers' actions

The development of a successful business strategy sometimes depends upon having good forecasts of the actions and reactions of competitors whose actions might have an influence on market shares. For example, if you lower your price, what will your competitors do? A variety of judgmental methods can be used to forecast competitors' actions. These include:

- expert opinion (ask experts who know about the relevant markets);
- intentions (ask competitors how they would respond in a given situation);
- structured analogies (analyze similar situations and the decisions that were made);
- simulated interaction (act out the interactions among decision makers for the firm and competitors); and
- experimentation (try the strategy on a small scale and monitor the results).

In some situations, forecasting the actions of interest groups is important. For example, how would organizations that lobby for environmental causes react to the introduction of packaging changes by a large fast-food restaurant chain? Use structured analogies and simulated interaction for such problems.

The need to forecast behavior in one's own organization is sometimes overlooked. Company plans typically require the cooperation of many people. Managers may decide to implement a given strategy, but will the organization be able to carry out the plan? Sometimes an organization fails to implement a plan because of a lack of resources, misunderstanding, or opposing groups. Intentions surveys of key decision makers in an organization may help to assess whether a given strategy can be implemented successfully. Simulated interaction can also provide useful forecasts in such situations.

Predict the *effects* of strategies intended to influence demand. One can make such forecasts by using expert judgment, judgmental bootstrapping, or econometric methods.

Forecasting market share

If one expects the same causal forces and the same types of behavior to persist, a naïve extrapolation of market share, such as from a no-change model, or in the case of a consistent trend in market share that is expected to continue, use a damped trend.

Draw upon methods that incorporate causal reasoning when large changes are expected. If small changes in the factors that affect market share are anticipated, use judgmental methods such as expert surveys or Delphi. If the changes in the factors are expected to be large, the causes are well understood, and data are scarce, use judgmental bootstrapping.

Use econometric methods when (1) the marketing activities differ substantially from previous activity; (2) data are sufficient and sufficiently variable; (3) models can allow for different responses by different brands; (4) models can be estimated at brand level; and (5) competitors' actions can be forecast (Brodie, Danaher, Kumar, and Leeflang 2001).

Knowledge about relationships can sometimes be can be obtained from prior research. For example, a meta-analysis of price elasticities of demand for 367 branded products, estimated using econometric models, reported a mean value of -2.5 (Tellis 2009). Estimates can also be made about other measures of market activity, such as advertising elasticity.

Choosing methods to suit the conditions

Evidence-based forecasting identifies the conditions that favor each method. Selecting the best forecasting method for a given situation is not a simple task. Often more than one method will provide useful forecasts.

The first question a forecaster confronts is whether the data are sufficient to develop a quantitative model. If not, you will need to use judgmental procedures. The two are not mutually exclusive: In many situations, both quantitative and judgmental methods are possible and useful.

For situations involving small changes, where no policy analysis is needed, and where forecasters get good feedback—such as with the number of diners that will come to a restaurant at a given time—unaided judgment can work well. If, however, the feedback is poor or uncertainty is high, using experts in a structured manner such as with a questionnaire or, if the relevant information is distributed among experts, with a Delphi panel, will help. Where policy analysis is needed, judgmental bootstrapping or decomposition will help to use experts' knowledge effectively.

For situations involving large changes, but which do *not* involve conflicts among a few decision makers, ask whether policy analysis is required. If policy analysis *is* required, as with situations involving small changes, use judgmental bootstrapping or decomposition to elicit forecasts from experts.

Experimentation is the most relevant and valid way to assess how customers would respond to changes in products or in the way of marketing products.

If policy analysis is not required, intentions or expectations surveys of potential customers may be useful. Consider also expert surveys, perhaps using the Delphi technique.

To make forecasts about situations that involve conflict among a few decision makers, ask whether similar cases exist. If they do, use structured analogies. If similar cases are hard to identify or the value of an accurate forecast is high, such as where a competitor reaction might have major consequences, use simulated interaction.

Turning now to situations where sufficient quantitative data are available to consider the estimation of quantitative models, ask whether knowledge about the relationships between causes and effects is also available. If knowledge about such relationships is good, use the knowledge to specify regression models so as to assess effect size. For example, to forecast the extent of an increase on the employment of unskilled people due to a decrease in the minimum wage rate, estimate a regression model using data from different jurisdictions and over time.

If the data are cross-sectional (e.g. for stores in different locations or product launches in different countries) use the method of quantitative analogies. For example, the introduction of new products in U.S. markets can provide analogies for the outcomes of the subsequent release of similar products in other countries.

If time-series data are available and domain knowledge is not good, use extrapolation methods to forecast. Where good domain knowledge exists (such as when a manager knows that sales will increase due to the advertising of a price reduction), consider using rule-based forecasting.

Much of the benefit of rule-based forecasting can be obtained by using the contrary-series rule. The rule is easy to implement: ignore the historical trend when managers expect causal forces to act against the trend. For example, where sales of new cars have been increasing over recent times, forecast flat sales when signs of economic recession are emerging.

For situations where knowledge of relationships is good and large changes are unlikely, as is common in the short-term, use extrapolation. If large changes are likely, causal methods provide forecasts that are more accurate. Models estimated using regression analysis, or econometrics, may

provide useful forecasts when important variables are few, much good quantitative data are available, relationships are linear, correlations among causal variables are low, and interactions are absent.

If the relationships are complicated, consider segmentation. Forecast the segments independently using appropriate methods.

Often the conditions are not favorable for regression analysis. In such situations, consider using the index method.

Forecasting demand for a new product

New product forecasting is important given that large investments are commonly involved and uncertainty is high. The choice of a forecasting method depends on what life-cycle stage the product has reached.

Surprisingly, surveys of what consumers want and of how they make decisions are of little value. For example, such an approach was said to have led to the conclusion that customers would not be interested in 3M's proposed Post-its. As shown in a meta-analysis of many studies from diverse areas of decision-making, customers are largely unaware of how they make decisions to purchase products (Nisbett and Wilson 1977).

Rather than asking consumers what they want, it is better to provide them with product choices and ask about their intentions and expectations. A product description may involve prototypes, visual aids, product clinics, or brochures. A relatively simple description of the key features of the proposed product is the best place to start, given the findings that decision makers cannot handle substantial amounts of information, as shown in a study of a proposed car-share system for Philadelphia (Armstrong and Overton 1971) Consumer intentions (or expectations) can improve forecasts even when some sales data are available (Armstrong, Morwitz and Kumar 2000).

It is sometimes difficult to identify potential customers for a new product. An inexpensive way around this is to create a role for subjects and asked them about their intentions to adopt the product when in that role.

Expert opinions are useful in the concept phase. Obtaining forecasts from the sales force is common. The Delphi method provides an effective way to conduct such surveys. In doing so, avoid biased experts, adjust for biases, or recruit a diverse panel.

Improve expert forecasts by decomposing the problem into parts that are better known than the whole. Thus, to forecast the sales of very expensive cars, rather than making a direct forecast ask "How many households will exist in the U.S. in the forecast year?' "Of these households, what percentage will make more than $500,000 per year?" and so on. The forecast is obtained by multiplying the components.

Experts can make predictions about a set of situations (20 or so) involving alternative product designs and alternative marketing plans. These predictions would then be related to the situations by regression analysis. Expert judgments have advantages over conjoint analysis in that few experts—between five and twenty—are needed. In addition, expert judgments can incorporate policy variables, such as advertising, that are difficult for consumers to assess.

Information about analogous products can be used to forecast demand for new products. Collect historical data on the analogous products and examine their growth patterns. Use the typical pattern as a forecast for the new product.

Once a new product is on the market, extrapolation is possible. Much attention has been given to selecting the proper functional form. The diffusion literature recommends an S-shaped curve to predict new product sales. That is, growth builds up slowly at first and then becomes rapid (if word-of-

mouth is good, and if people see the product being used by others). Then growth slows as sales approach a saturation level. Evidence on what is the best way to model the process is limited and the benefits of choosing the best functional form are modest (Meade and Islam 2001). In the absence of evidence to the contrary, use simple and understandable growth curves.

Improving forecasts

Even when forecasts have been derived from evidence-based methods that were selected to suit the conditions, it may still be possible to improve the accuracy of the forecasts by combining and adjusting forecasts. But first, it is necessary to consider how to measure accuracy so as to know when it has improved. There are many error measures that might be used for assessing forecast accuracy, and the choice of measures is important. A key lesson from evidence-based forecasting is, do not use mean square error (MSE). While MSE has characteristics that statisticians find attractive, the measure is not reliable (Armstrong and Collopy 1992). Though still commonly used, MSE use by firms has dropped substantially in recent years (McCarthy, Davis, Golicic, and Mentzer 2006). The median absolute percentage error (MdAPE), on the other hand, is appropriate for many situations because the measure is not affected by scale or by outliers. The cumulative relative absolute error (CumRAE) is another measure that is easy to interpret, and it useful for comparing the accuracy of forecasts from the method of interest with those from a benchmark.

Combining forecasts

Combining forecasts is one of the most powerful procedures in forecasting and is applicable to a wide variety of problems. Combining is most useful in situations where the true value might fall between forecasts.

In order to increase the likelihood that two forecasts bracket the true value, use methods and data that differ substantially. The extent and probability of error reduction through combining is higher when differences among the methods and data that produced the component forecasts are greater

Use trimmed averages or medians for combining forecasts. Avoid differential weights unless there is strong empirical evidence that the relative accuracy of forecasts from the different methods differs.

Gains in accuracy from combining are higher when forecasts are made for an uncertain situation, and many forecasts are available from several reasonable methods especially when using different data sources. Under such favorable conditions, combining can cut errors by half (Graefe, Armstrong, Jones, and Cuzán 2012). Combining forecasts helps to avoid large errors, and often improves accuracy even when the best method if known beforehand.

Adjusting Forecasts

If judgmental forecasts are likely to be biased, adjust the forecasts based on evidence of bias from similar forecasting situations. When forecasts are likely to be too optimistic consider instructing the forecasters to assume the first forecast reflect ideal conditions and ask them to now provide forecasts based on realistic conditions (Jørgensen 2011). For new situations, consider obtaining a second forecast assuming the first one was wrong, and average the two (Herzog and Hertwig 2009). When judgmental forecasts are made repeatedly, regress errors against variables forecasters should

have used, then combine statistical forecasts of error from the resulting model with new judgmental forecasts to improve accuracy (Fildes, Goodwin, Lawrence, and Nikolopoulos 2009).

When making judgmental adjustments of statistical forecasts: (1) Adjust only for important information about future events; (2) Record reasons for adjustments; (3) Decompose the adjustment task if feasible; (4) Mechanically combine judgmental and statistical forecasts; and (5) Consider using a Delphi panel for determining adjustments (Goodwin 2005). Future events might include new government regulations coming into force, a planned promotion, the loss of an important client, or a competitor's actions. Consider estimating a regression model to correct judgmental forecasts for biases (Goodwin, Önkal, and Lawrence 2011).

When statistical forecasts are derived using causal methods, judgmental adjustments can help accuracy if important variables are missing from the causal model, data are poor, relationships are miss-specified, relationships are believed to have changed, or the environment has changed (Goodwin *et al*. 2011).

Assessing and communicating uncertainty

In addition to improving accuracy, the discipline of forecasting is concerned with assessing uncertainty about accuracy and measuring error. Improved assessments of forecast uncertainty or risk help with decision making and planning, such as in determining safety stocks for inventories

Present estimates of uncertainty about as prediction intervals, such as "we estimate an 80% chance that demand for new passenger vehicles in Australia in 2020 will be between 400,000 and 700,000." Do not use the fit of a model to historical data to estimate prediction intervals. Do consider (1) experts' assessments, (2) the distribution of forecasts from different methods and forecasters, and (3) the distribution of *ex ante* forecast errors.

Traditional confidence intervals, which are estimated from historical data for quantitative forecasts, tend to be too narrow. Empirical studies show that the percentage of actual values that fall outside the 95% confidence intervals is often greater than 50% (Makridakis, Hibon, Lusk, and Belhadjali 1987). The problem occurs because confidence interval estimates ignore important sources of uncertainty.

Forecast errors in time series are often asymmetric, and this asymmetry makes estimating confidence intervals difficult. Asymmetry of errors is likely to occur when the forecasting model uses an additive trend. The most sensible procedure is to transform the forecast and actual values to logs, calculate the prediction intervals using logged differences, and present the results in actual values (Armstrong and Collopy 2001).

Loss functions can also be asymmetric. For example, the losses due to a forecast that is too low by 50 units may differ from the losses if a forecast is too high by 50 units. But asymmetric loss functions are a problem for the planner, not the forecaster.

Overconfidence arising from historical fit is compounded when analysts use the traditional statistics provided with regression programs (Soyer and Hogarth 2012). Tests of statistical significance are of no value to forecasters *even when properly used and properly interpreted* and the tests often mislead decision makers (Armstrong 2007).

Experts also are typically overconfident and hence underestimate uncertainty (Arkes 2001). For example, in an examination of economic forecasts from 22 economists over 11 years, the actual values fell outside the range of their prediction intervals about 43% of the time. This problem occurs even when the economists were warned in advance against overconfidence. Group interaction and providing explanations both increase overconfidence. A series of four studies provide support for explanations for

overconfidence that include poor feedback, belief in uniqueness, misunderstanding of confidence levels, desire to appear skilled, and rewards for overconfidence (Jørgensen, Teigen, and Moløkken 2004).

To improve the calibration of judges, ensure they receive timely and accurate information on what actually happened, along with reasons why their forecasts were right or wrong. Receiving this kind of feedback is part of the reason why weather forecasters are well calibrated for day-ahead forecasts of, for example, the chance of rain. In cases where good feedback is not possible, ask experts to write all the reasons why their forecasts might be wrong; doing so will tend to moderate overconfidence (Arkes, 2001).

Still another way to assess uncertainty is to examine the agreement among forecasts. For example, agreement, or lack of agreement, among judgmental forecasts of annual advertising sales for *Time* magazine was a good proxy for uncertainty (Ashton 1985). The differences between the forecasts of the individual experts participating in a Delphi panel can be used in this way.

Finally, uncertainty is most faithfully represented using empirical prediction intervals estimated from *ex ante* forecast errors from the same or similar forecasting situations (Chatfield 2001). Simulating the actual forecasting procedure as closely as possible, and using the distribution of the resulting *ex ante* forecasts to assess uncertainty is best. For example, if you need to make forecasts for two years ahead, withhold enough data to be able to estimate the forecast errors for two-year-ahead *ex ante* forecasts. When organizations make many similar forecasts, use evidence on errors from previous forecasts to develop heuristics for estimating prediction intervals for new forecasts. For example, NASA's Software Engineering Laboratory guidelines for estimating prediction intervals were simply factors between 1.05 and 2.00 to apply to the forecasts, such that the PI is from the forecast divided by the factor to the forecast multiplied by the factor (Jørgensen, Teigen, and Moløkken 2004).

New product forecasts are particularly prone to uncertainty, and there are no previous forecasts for the product to use for estimating empirical prediction intervals. Looking at the record of forecasting new products can help, especially if it is possible to obtain accuracy data for forecasting situations that are somewhat similar to the one being forecast. Published benchmark accuracy data for new product forecasting is a good place to start (see Armstrong 2002).

Implementation of evidence-based methods

The forecastingprinciples.com is a free website dedicated to helping people on business and government to improve their forecasting procedures.[4] It provides the forecasting principle as a checklist. Most of the principles are relevant for demand forecasting.

Structured checklists are an effective way to make complex tasks routine, to avoid the need for memorizing, and to provide relevant guidance on a just-in-time basis. This is useful for applying principles that are already agreed upon such as in flying an airplane or in doing a medical operation. Consider the following experiment: In 2008, an experiment was used to assess the effects of using a 19-item checklist for hospital procedures. This before/after experimental design was used for thousands of

[4] Forecasting is especially important for the not-for-profit sector as there is no guidance from market prices, and also, because there is no self-correcting mechanism. Publicpolicyforecasting.com was created to enable governments and disinterested parties to show that their proposed projects follow proper forecasting procedure. To date, the three public project audits on this site showed virtually no awareness of proper forecasting procedures.

patients in eight hospitals in eight cities around the world. In the month after the operations, the checklist led to a reduction in deaths from 1.5% to 0.8%, and in complications, from 11% to 7% (Haynes et al. 2009).

Checklists serve an additional role in forecasting as they introduce analysts to principles they were unaware of. At the time that the original 139 forecasting principles were published, a review of 18 forecasting textbooks found that, the typical textbook mentioned only 19% of the principles. At best, one textbook mentioned 34% of the principles (Cox and Loomis 2001).

The Forecasting Audit Software is essentially a checklist to guide the choice and implementation of a demand forecasting process that is evidence-based and suited for the situation. The Software can also be used to assess the extent to which a forecasting process is consistent with evidence-based forecasting principles and to make suggestions for how the process might be improved.

Full disclosure of the methods and data provide the primary requirement for the audit. Unfortunately many forecasting efforts fail to provide sufficient information. For example, in 2012, we attempted to conduct a forecasting audit of the proposed California high-speed line. We were able to obtain many reports that were said to support the decision to move ahead with this controversial project, but they did not provide sufficient information on the data and methods to allow for a meaningful audit. Proper forecasts are critical in this case given that the private market is unwilling to develop such a transportation system.

The most effective way to introduce new principles would be to do so via forecasting software. Unfortunately, this does not seem to happen. An early attempt to review demand-forecasting software failed because none of the commercial providers would cooperate (Tashman and Hoover 2001). Some providers mention the use of principles (e.g., damped trend and the use of better measures of forecast errors than the Root Mean Square Error), but in general few of the principles seem to have been implemented, Other than Forecast Pro and SAS, software provides have shown little interest in the forecasting principles project. The general opinion is that the providers will respond to clients' requests. Clients might want to use the checklists to see whether their providers use – or will use—the evidence-based principles.[5]

Conclusions

Evidence-based forecasting involves experimental testing of multiple reasonable hypotheses. Although only a few researchers have adopted the approach, their contributions have led to remarkable progress over the past four decades. The gains from evidence based research have been to reveal which methods do not seem to help under any conditions, (e.g., game theory), which help under given conditions (e.g., index methods, for causal models in complex and uncertain situations,) and what is the most effective way to use each method (e.g., proposes analogies prior to making forecasts). We also know which methods offer little promise despite enormous efforts devoted to them. These include focus groups, conjoint analysis, and complex models.

Advances touch on many aspects of demand forecasting. Some relate to the use of judgment, such as with Delphi, simulated interactions, intentions surveys, expert surveys, judgmental bootstrapping, and combining. Others relate to quantitative methods such as extrapolation, rule-based forecasting, and the index method. Many of these methods are relatively simple to use and easy to

[5] We speculate that the problem with software providers is that the methods are designed by statisticians who apparently are unaware of the evidence-based research on forecasting. For example, statistician who are interested in forecasting seldom refer to the evidence-based literature. (Fildes and Makridakis 1995.)

understand. Most recently, gains have come from the integration of statistical and judgmental forecasts. Much has been learned about how to implement these forecasting methods.

Over the past few years, despite much effort to help practitioners by providing understandable evidence-based forecasting principles and techniques, and by making them freely available at forecastingprinciples.com, most firms, consultants, and software developers seem to unaware of the evidence-based research on forecasting. As a consequence, there are great opportunities to improve the accuracy and cost effectiveness of demand forecasts .

~9,940 words excluding abstract and footnotes.

Acknowledgments: Sven Crone, Paul Goodwin, and Andreas Graefe made suggestions on early versions…

REFERENCES

Adya, M., & Collopy, F. (1998). How effective are neural nets at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, 17, 451–461.

Allen, P. G., & Fildes, R. (2001). Econometric forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 303–362). Norwell, MA: Kluwer Academic Publishers.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 495–515). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting* [Forthcoming].

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, 23, 321–327.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, 22, 583–598.

Armstrong, J. S. (2003). Discovery and communication of important marketing findings, evidence, and proposals, *Journal of Business Research*, 56 (2003), 69-84.

Armstrong, J. S. (2002). Benchmarks for new product forecast errors. Published on the Internet at http://forecastingprinciples.com/files/Benchmark%20New%20Product%20Errors.pdf

Armstrong, J. S. (Ed.). (2001). *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 171–192). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (1988) "Research Needs in Forecasting, *International Journal of Forecasting*, 4 (1988), 449-465.

Armstrong, J. S. (1980). The Seer-Sucker Theory: The Value of Experts in Forecasting, *Technology Review*, 83 (June/July 1980), 18-24.

Armstrong, J. S., Adya, M., & Collopy, F. (2001). Rule-based forecasting: Using judgment in time-series extrapolation. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 259–282). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S., Brodie, R. & Parsons, A. (2001). Hypotheses in Marketing Science: Literature Review and Publication Audit. *Marketing Letters*, 12, 171–187

Armstrong, J. S., & Collopy, F. (2001). Identification of asymmetric prediction intervals through causal forces. *Journal of Forecasting*, 20, 273–283.

Armstrong, J. S., & Collopy, F. (1993). Causal forces: Structuring knowledge for time series extrapolation. *Journal of Forecasting,* 12, 103–115.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting,* 8, 69–80.

Armstrong, J. S., F. Collopy, F. & Yokum, T. (2005). Decomposition by causal forces: A procedure for forecasting complex time series. *International Journal of Forecasting,* 21, 25-36.

Armstrong, J. S., & Graefe, A. (2011). Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research,* 64, 699–706.

Armstrong, J. S., Morwitz, V., & Kumar, V. (2000). Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy? *International Journal of Forecasting,* 16, 383–397.

Armstrong, J. S., & Overton, T. S. (1971). Brief vs. comprehensive descriptions in measuring intentions to purchase. *Journal of Marketing Research,* 8, 114–117.

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

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science,* 41, 68–75.

Brodie, R. J., Danaher, P., Kumar, V., & Leeflang, P. (2001). Econometric models for forecasting market share. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 597–611). Norwell, MA: Kluwer Academic Publishers.

Chamberlin, T. C. (1965). The method of multiple working hypotheses. *Science,* 148, 754–759.

Chatfield, C. (2001). Prediction intervals for time series. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 475–494). Norwell, MA: Kluwer Academic Publishers.

Collopy, F., Adya, M. & Armstrong, J. S. (2001). Expert systems for forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 285–300). Norwell, MA: Kluwer Academic Publishers.

Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting: Development and validation of an expert systems approach to combining time-series extrapolations. *Management Science,* 38, 1394–1414.

Cox, J. E. & Loomis, D. G. (2001), Diffusion of forecasting principles through books. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 633-–649). Norwell, MA: Kluwer Academic Publishers.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting,* 27, 635–660.

Dangerfield, B. J., & Morris, J. S. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting,* 8, 233–241.

Dillman, D. A., Smyth J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. (3rd ed.). Hoboken, NJ: John Wiley.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting,* 25, 3–23.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. International Statistical Review, 65, 289-308.

Forsyth, D. K., & Burt, C. D. B. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory & Cognition*, 36, 791–798.

Gardner, E. S., Jr. (2006). Exponential smoothing: The state of the art – Part II (with commentary). *International Journal of Forecasting*, 22, 637–677.

Goodwin, P. (2005). How to integrate management judgment with statistical forecasts. *Foresight*, 1, 8–12.

Goodwin, P., Önkal, D., & Lawrence, M. (2011). Improving the role of judgment in economic forecasting. In M. P. Clements, & D. F. Hendry (Eds.), *The Oxford Handbook of Economic Forecasting* (pp. 163–189). Oxford, UK: OUP.

Gorr, W., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19, 579–594.

Graefe, A. (2011). Prediction market accuracy for business forecasting. In L. Vaughan-Williams (Ed.), *Prediction Markets* (pp. 87–95). New York: Routledge.

Graefe, A. and Armstrong, J.S. (2012), Forecasting Elections from Voters' Perceptions of Candidates' Ability to Handle Issues**,** *Journal of Behavioral Decision Making,* DOI: 10.1002/bdm.1764.

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2012). Combining forecasts: An application to political elections. *Working paper*. [Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1902850]

Graefe, A., & Armstrong, J. S. (2011). Conditions under which index models are useful: Reply to Bio-index Commentaries. *Journal of Business Research*, 64, 693–695.

Graefe, A., & Armstrong, J. S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task, *International Journal of Forecasting*, 27, 183-195.

Gratzer, D. (2006) *The Cure.* New York: Encounter Books.

Grimstad, S., & Jørgensen, M. (2007). Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, 80, 1770-1777.

Gregory, W. L., & Duran, A. (2001). Scenarios and acceptance of forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 519–541). Norwell, MA: Kluwer Academic Publishers.

Green, K. C. (2005). Game theory, simulated interaction, and unaided judgment for forecasting decisions in conflicts: Further evidence. *International Journal of Forecasting*, 21, 463–472.

Green, K. C. (2002). Forecasting decisions in conflict situations: a comparison of game theory, role-playing, and unaided judgement. *International Journal of Forecasting*, 18, 321–344.

Green, K. C., & Armstrong, J. S. (2011). Role thinking: Standing in other people's shoes to forecast decisions in conflicts, *International Journal of Forecasting*, 27, 69–80.

Green, K. C., & Armstrong, J. S. (2007). Structured analogies for forecasting. *International Journal of Forecasting*, 23, 365–376

Green, K. C., Armstrong, J. S., & Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared. *Foresight*, 8, 17–20. Available from http://kestencgreen.com/green-armstrong-graefe-2007x.pdf

Haynes, Alex B., et al. (2009), "A surgical checklist to reduce morbidity and mortality in a global population," *New England Journal of Medicine*, 360 (January 29), 491-499.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual *Psychological Science*, 20 (2), 231-237.

Jørgensen, M. (2011). Contrasting ideal and realistic conditions as a means to improve judgment-based software development effort estimation. *Information and Software Technology*, 53, 1382–1390.

Jørgensen, M. (2004). Top-down and bottom-up expert estimation of software development effort. *Journal of Information and Software Technology*, 46 (1), 3–16.

Jørgensen, M. & Sjøberg, D. I. K. (2004). The impact of customer expectation on software development effort estimates. *International Journal of Project Management*, 22, 316–325.

Jørgensen, M. & Sjøberg, D. I. K. (2003). An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. *Information and Software Technology*, 45, 123-136.

Kealey, T. (1996) *The Economic Laws of Scientific Research*. London: Macmillan.

Keogh, E. J., & Kasetty, S. (2002). On the need for time series data mining benchmarks: A survey and empirical demonstration. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 102–111.

Kim, M. S. & Hunter, J. E. (1993). Relationships among attitudes, behavioral intentions, and behavior: A meta-analysis of past research. *Communication Research,* 20, 331–364.

Locke, E. A. (1986). *Generalizing from Laboratory to Field Settings*. Lexington, MA: Lexington Books.

Li, Y. F., Xie, M. Goh, T. N. (2009). A study of project selection and feature weighting for analogy based software cost estimation. *The Journal of Systems and Software*, 82, 241–252.

Lovallo, D., Clarke, C., Camerer, C. (2012). Robust analogizing and the outside view: Two empirical tests of case-based decision making. *Strategic Management Journal*, 33, 496–512,

MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 107–123). Norwell, MA: Kluwer Academic Publishers.

Makridakis, S. G., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J. Parzen, E., & Winkler, R. (1984). *The Forecasting Accuracy of Major Times Series Methods.* Chichester: John Wiley.

Makridakis, S. G., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of time series in the M-competition. *International Journal of Forecasting*, 3, 489–508.

Makridakis, S. G. & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, 451–476.

McCarthy, T. M., Davis, D. F., Golicic, S. L. & Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting*, 25, 303–324.

Meade, N., & Islam, T. (2001). Forecasting the diffusion of innovations: Implications for time series extrapolation. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 577–595). Norwell, MA: Kluwer Academic Publishers.

Miller, D. M., & Williams, D. (2003). Miller, D. M. and Williams, D. (2003). Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy, *International Journal of Forecasting*, 19, 669-684.

Miller, D. M., & Williams, D. (2004). Shrinkage estimators for damping X12-ARIMA seasonals. *International Journal of Forecasting*, 20, 529–549.

Morwitz, V. G. (2001). Methods for forecasting from intentions data. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 33–56). Norwell, MA: Kluwer Academic Publishers.

Morwitz, V. G., Steckel, J. H., & Gupta, A. (2007). When do purchase intentions predict sales?" *International, Journal of Forecasting*, 23, 347-364.

Nisbett, R. E. and T. D.Wilson (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes, *Psychological Review*, 84, 231-259

Powdthavee, N., & Riyanto, Y. (2012). Why do people pay for useless advice? Implications of gambler's and hot-hand fallacies in false-expert setting. IZA Discussion Paper No. 6557.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 125–144). Norwell, MA: Kluwer Academic Publishers.

Sheeran, Paschal (2002). "Intention-behavior relations: A conceptual and empirical review," in Wolfgang Stroebe and Miles Hewstone, *European Review of Social Psychology*, volume 12, pp. 1-36.

Sichtmann, C., Wilken, R., & Diamantopoulos, A. (2011). Estimating willingness-to-pay with choice-based conjoint analysis – Can consumer characteristics explain variations in accuracy? *British Journal of Management*, 22, 628–645.

Soyer, E., & Hogarth, R. (2012). Illusion of predictability: How regression statistics mislead experts.

Tashman, L. J. & Hoover, J. (2001). Diffusion of forecastingprinciples through software. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 651-676). Norwell, MA: Kluwer Academic Publishers.

Tellis, G. J. (2009). Generalizations about Advertising Effectiveness in Markets *Journal of Advertising Research*, 49 (2), 240-245.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* New Jersey: Princeton University Press.

Wittink, D. R., & Bergestuen, T. (2001). Forecasting with conjoint analysis. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 147–167). Norwell, MA: Kluwer Academic Publishers.

Wright, M., & MacRae, M. (2007). Bias and variability in purchase intention scales. *Journal of the Academy of Marketing Science*, 35, 617–624.

Ziliak, S. T., & McCloskey, D, N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.

Authors

Kesten C. Green (Ph.D., VUW, 2003) teaches managerial economics at the International Graduate School of Business of the University of South Australia and is a Senior Research Associate of the Ehrenberg-Bass Institute for Marketing Science. He is also a Director of the International Institute of Forecasters and co-director of the Forecasting Principles public service Internet site devoted to the advancement of evidence-based forecasting. His research has led to improvements in forecasting the decisions people make in conflicts such as occur in business competition, supply chains, mergers and acquisitions, and between customers and businesses. His other interests include forecasting for public policy, forecasting demand, forecasting for recessions and recoveries, and the effect of business objectives on performance. His research has been covered in the *Australian Financial Review*, the *London Financial Times*, the *New Yorker*, and the *Wall Street Journal*. He has advised the Alaska Department of Natural Resources, the U.S. Department of Defense, the Defense Threat Reduction Agency, the National Security Agency (NSA) and more that 50 other business and government clients. Kesten can be contacted at kesten@me.com.

J. Scott Armstrong (Ph.D., MIT, 1968), Professor of Marketing at the Wharton School, University of Pennsylvania, is a founder of the *Journal of Forecasting*, the *International Journal of Forecasting*, and the International Symposium on Forecasting. He is the creator of forecastingprinciples.com and editor of *Principles of Forecasting* (Kluwer 2001), an evidence-based summary of knowledge on forecasting. In 1996, he was selected as one of the first "Honorary Fellows" by the International Institute of Forecasters. In 2004 and 2008, his PollyVote.com team showed how

scientific forecasting principles can produce highly accurate forecasts of US presidential elections. He was named by the Society of Marketing Advances as "Distinguished Marketing Scholar of 2000." One of Wharton's most prolific scholars, he is the most highly cited professor in the Marketing Department at Wharton. His current projects involve the application of scientific forecasting methods to climate change, the effectiveness of learning at universities, and the use of the index method to make predictions for situations with many variables and much knowledge. His book, *Persuasive Advertising*, was published by Palgrave Macmillan in 2010. The book summarizes evidence-based knowledge on persuasion and is supported by advertisingprinciples.com. He can be contacted at armstrong@wharton.upenn.edu.