Forecasting Methods and Principles: Evidence-Based Checklists

J. Scott Armstrong[1]
Kesten C. Green[2]

Working Paper 128-clean
August 1, 2017

ABSTRACT

Problem: Most forecasting practitioners are unaware of discoveries from experimental research over the past half-century that can be used to reduce errors, often by more than half. The objective of this paper is to improve forecasting accuracy by providing evidence-based methods and principles to forecasters and decision-makers in a form that is easy for them to understand and use.

Methods: This paper reviews research on forecasting to identify methods that are useful for forecasting, and those that are not, and to develop checklists of evidence-based forecasting to help forecasters and their clients. The primary criterion for evaluating whether or not a method is useful is predictive validity as assessed by evidence on its contribution to *ex ante* predicve validity.

Findings: This paper identifies and describes 17 evidence-based forecasting methods and 8 that are not. Six checklists are provided to apply evidence-based findings to forecasting problems by guiding the selection of the most suitable methods and their implementation.

Originality: Four of the six checklists are new: They address (1) evidence-based methods, (2) regression analysis, (3) uncertainty, and (4) popular but not validated forecasting methods. Another checklist, the Golden Rule, was improved.

Usefulness: The checklists are low-cost tools that forecasters can use to obtain forecasts that are substantially more accurate than those provided by commonly used methods. The completed checklists provide assurance to clients and other interested parties that the resulting forecasts were derived using evidence-based procedures.

***Key words*:** big data, combining forecasts, decision-making, decomposition, equal weights, expectations, extrapolation, index method, intentions, Occam's razor, prediction intervals, regression analysis, scenarios, uncertainty

Authors' notes:

1. We received no funding for this paper and have no commercial interests in any forecasting method.
2. We estimate that most readers can read this paper in one hour.
3. We endeavored to conform with the Criteria for Science Checklist at GuidelinesforScience.com.

Acknowledgments: We thank Hal Arkes, Roy Batchelor, David Corkindale, Alfred G. Cuzán, Robert Fildes, Paul Goodwin, Andreas Graefe, Rob Hyndman, Randall Jones, Kostas Nikolopoulos, and Malcolm Wright for their reviews. We also thank those who made useful suggestions, including Frank Schmidt and Phil Stern. This is not to say that the reviewers and those who made suggestions agreed with all of the findings in this paper. Finally, we thank those who edited the paper: Maya Mudambi, Esther Park, Scheherbano Rafay, and Lynn Selhat.

[1] The Wharton School, University of Pennsylvania, 747 Huntsman, Philadelphia, PA 19104, U.S.A. and Ehrenberg-Bass Institute, University of South Australia Business School: +1 610 622 6480; armstrong@wharton.upenn.edu

[2] School of Commerce and Ehrenberg-Bass Institute, University of South Australia Business School, University of South Australia, City West Campus, North Terrace, Adelaide, SA 5000. kesten.green@unisa.edu.au

INTRODUCTION

This paper is concerned with improving the accuracy of forecasts by making scientific knowledge on forecasting available to forecasters. Accurate forecasts are important for businesses and other organizations, and for those concerned with developing useful government policies.

Thanks to findings from experiments testing multiple reasonable hypotheses, forecasting has advanced rapidly since the 1930s. In the mid-1990s, 39 leading forecasting researchers and 123 expert reviewers were involved in identifying and collating scientific knowledge on forecasting. These findings were summarized as principles (condition-action statements). One-hundred-and-thirty-nine principles were formulated (Armstrong 2001b, pp. 679-732). In 2015, two papers further summarized forecasting knowledge in the form of two overarching principles: simplicity and conservatism (Green and Armstrong 2015, and Armstrong, Green, and Graefe 2015, respectively). The guidelines for forecasting described in this paper draw upon those evidence-based principles.

This paper is concerned with methods that have been shown to improve forecast accuracy relative to methods that are commonly used in practice. Absent a political motive that a preferred plan be adopted, accuracy is the most important criterion for most parties concerned with forecasts (Fildes and Goodwin 2007). Other important criteria include forecast uncertainty, cost, and understandability (Yokum and Armstrong 1995).

CHECKLISTS TO IMPLEMENT AND ASSESS FORECASTING METHODS

We summarize knowledge on how best to forecast in the form of checklists. Structured checklists are an effective way to make complex tasks easier, to avoid the need for memorizing, to provide relevant guidance on a just-in-time basis, and to inform others about the procedures that were used.

In fields such as medicine, aeronautics and engineering, checklists are required, and a failure to follow them can be grounds for lawsuits. Much research supports the value of using checklists (see, e.g., Hales and Pronovost 2006). One experiment assessed the effects of using a 19-item checklist for a hospital procedure. The before-and-after design compared the outcomes experienced by thousands of patients in hospitals in eight cities around the world. The checklist led to a reduction in deaths from 1.5% to 0.8% in the month after the operations, and in complications, from 11% to 7% (Haynes et al. 2009).

While the advances in forecasting knowledge over the past century have provided the opportunity for substantial improvements in accuracy, most practitioners do not make use of that knowledge. There are a number of reasons why that happens: practitioners (1) prefer to stick with their current forecasting procedures; (2) wish to provide support for a preferred outcome; (3) are unaware of evidence-based methods; or (4) are aware of the evidence-based methods, but they have not followed any procedure to ensure that they use them, and (5) they have not been asked to do so. Practitioners who are not using evidence-based forecasting methods for reasons 3, 4, or 5 will benefit from reading this paper and using the checklists provided.

At the time that the original 139 forecasting principles were published in 2001, a review of 17 forecasting textbooks found that the typical textbook mentioned only 19% of the principles. At best, one textbook mentioned one-third of the principles (Cox and Loomis 2001). It is not surprising then, that few of the evidence- and document that they are following proper methods. When clients demand forecasts from based forecasting principles are used. The adoption of new discoveries in forecasting have typically lagged their discovery by decades.

The checklists presented in this paper can help funders in asking forecasters to provide evidence-based forecasts, policy makers and other decision makers to assess whether forecasts can be trusted, and forecasters to ensure the proper procedures specified in the checklists, and audit the forecasters' work, compliance is likely. Our experience is that when we have commissioned people to complete tasks that required them to use a checklist, *all who accepted the task did so effectively*. One such task involved assessing the persuasiveness of print advertisements by assessing them against 195 checklist items that was carried out by people recruited through Amazon's Mechanical Turk (Armstrong et al. 2016).

RESEARCH METHODS

We reviewed research findings and provided checklists to make this knowledge accessible to forecasters and researchers. The review involved searching for papers with evidence from experiments that compared the performance of alternative methods. We did this using the following procedures:

1) Searching the Internet, mostly using Google Scholar.
2) Contacting key researchers for assistance, which one study found to be substantially more comprehensive than computer searches (Armstrong and Pagell, 2003).
3) Using references from key papers.
4) Putting working paper versions of our paper online (e.g., ResearchGate) with requests for papers that might have been overlooked. In doing so, we emphasized the need for experimental evidence, especially evidence that would challenge the findings presented in this paper.
5) Asking reviewers to identify missing papers.
6) Posting on relevant websites such as ForecastingPrinciples.com and ResearchGate.

To ensure that we properly summarized previous findings, we attempted to contact authors of all papers from which we cited substantive findings. This was done given the substantial evidence that a high percentage of papers cited in scientific papers are incorrect (Wright and Armstrong, 2008). The References section for this paper includes coding of our efforts to contact authors and the results. The procedure reduced errors and improved our descriptions of their findings. In some cases, the cited authors also suggested additional relevant papers.

Given the enormous number of papers with promising titles, we screened papers by whether the "Abstracts" or "Conclusions" reported the findings and methods. If not, we stopped. If yes, we checked whether the paper provided full disclosure. If yes, we then checked whether the findings were important. Of the papers with promising titles, only a small percentage passed these criteria.

The primary criterion for evaluating whether or not a method is useful was predictive validity, as assessed by evidence on the accuracy of *ex ante* forecasts from the method relative to those from current practice or to existing evidence-based alternative methods. The papers that met the criteria were used to develop the checklist in Exhibit 1.

VALID FORECASTING METHODS: DESCRIPTIONS AND EVIDENCE

Exhibit 1 provides a listing of all 17 forecasting methods that have been shown to have predictive validity. For each of the methods, the exhibit identifies the knowledge that is needed—in addition to knowledge of the method—to use the method for a given problem. The forecaster should be aware of evidence from prior experimental research that is relevant to the forecasting problem. For the great majority of forecasting problems, several of the methods listed in Exhibit 1 will be usable.

Electronic versions of the Exhibit 1 checklist, and the other checklists in this paper, are available on the Software Page of the ForecastingPrinciples.com site.

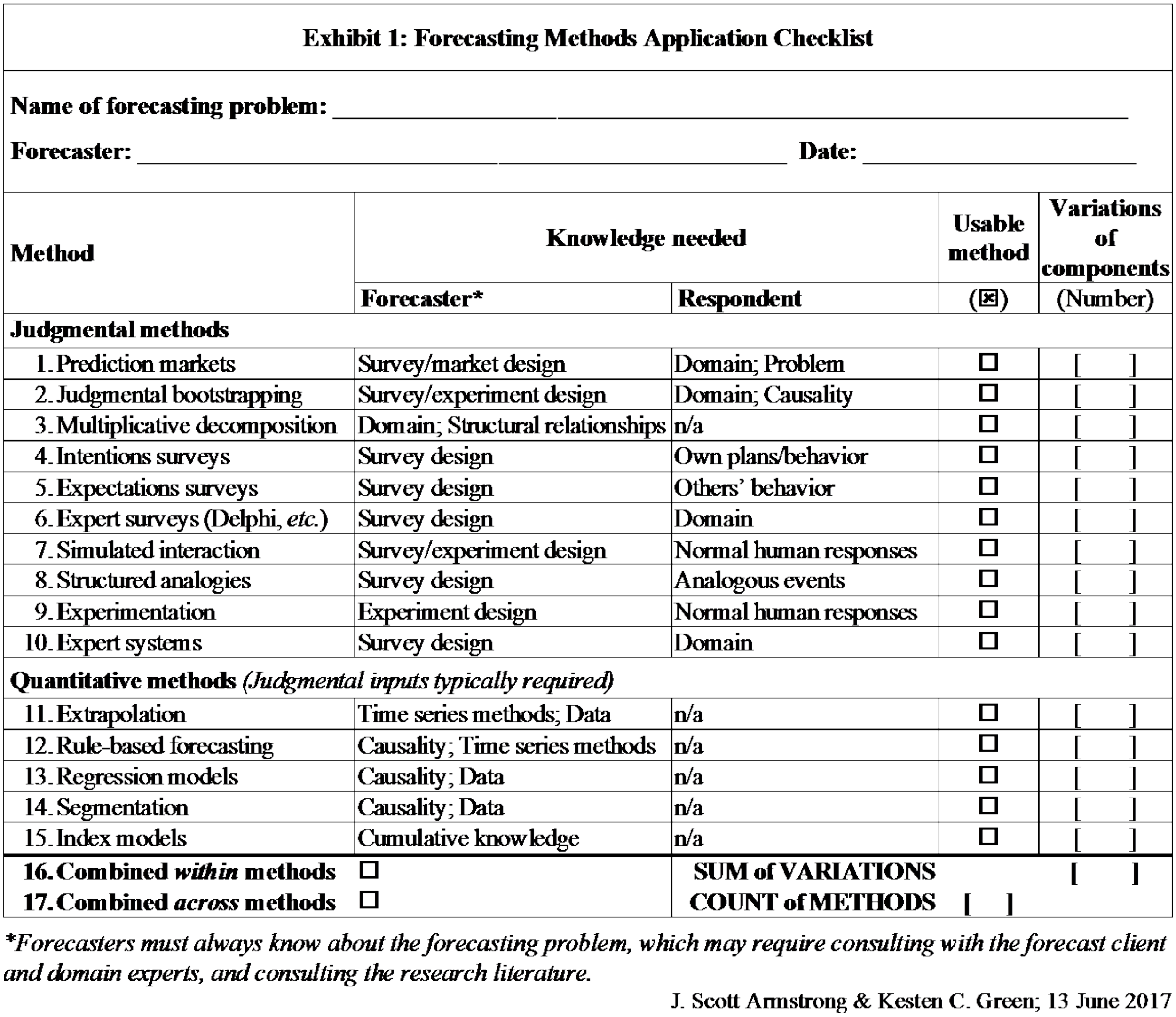

Exhibit 1: Forecasting Methods Application Checklist

Name of forecasting problem: ____

Forecaster: ______________________________________ **Date:** ____________________

Method	**Knowledge needed**		**Usable method**	**Variations of components**
	Forecaster*	**Respondent**	**(☒)**	**(Number)**
Judgmental methods				
1. Prediction markets	Survey/market design	Domain; Problem	□	[ ]
2. Judgmental bootstrapping	Survey/experiment design	Domain; Causality	□	[ ]
3. Multiplicative decomposition	Domain; Structural relationships	n/a	□	[ ]
4. Intentions surveys	Survey design	Own plans/behavior	□	[ ]
5. Expectations surveys	Survey design	Others' behavior	□	[ ]
6. Expert surveys (Delphi, *etc.*)	Survey design	Domain	□	[ ]
7. Simulated interaction	Survey/experiment design	Normal human responses	□	[ ]
8. Structured analogies	Survey design	Analogous events	□	[ ]
9. Experimentation	Experiment design	Normal human responses	□	[ ]
10. Expert systems	Survey design	Domain	□	[ ]
Quantitative methods *(Judgmental inputs typically required)*				
11. Extrapolation	Time series methods; Data	n/a	□	[ ]
12. Rule-based forecasting	Causality; Time series methods	n/a	□	[ ]
13. Regression models	Causality; Data	n/a	□	[ ]
14. Segmentation	Causality; Data	n/a	□	[ ]
15. Index models	Cumulative knowledge	n/a	□	[ ]
16. Combined *within* methods	□	**SUM of VARIATIONS**		**[ ]**
17. Combined *across* methods	□	**COUNT of METHODS**	**[ ]**	

Forecasters must always know about the forecasting problem, which may require consulting with the forecast client and domain experts, and consulting the research literature.

J. Scott Armstrong & Kesten C. Green; 13 June 2017

Practitioners typically use the method they are most familiar, or the method that they believe to be the best for the problem to hand. Both are mistakes. Instead, forecasters should arm themselves with all of the valid forecasting methods and seek to use all that are feasible for the problem. Further, forecasters should obtain forecasts from several implementations of each method, and combine the forecasts. When accuracy is important, we suggest forecasters obtain forecasts from at least two variations of each of three different methods, and to combine these combined forecasts over at least three different methods.

The predictive validity of a theory or a forecasting method is assessed by comparing the accuracy of forecasts from the method with forecasts from the currently used method, or from a simple plausible method, or from other evidence-based methods. Important, too, is to assess whether forecasts from a method can increase the accuracy of a combined forecast.

For qualitative forecasts—such as whether a, b, or c will happen, or which of x or y would be better—accuracy is typically measured as some variation of percent correct. For quantitative forecasts, accuracy is assessed by the size of the forecast errors. Forecast errors are measures of the absolute difference between *ex ante* forecasts and what actually transpired.

Evidence-based forecasting methods are described next. We start with judgmental methods, and follow with quantitative methods. The latter inevitably require some judgment.

Judgmental Methods

Judgmental methods should be fully disclosed and understandable. Contrary to common belief, expertise in a field or specific problem is, *on its own*, of no apparent value for making accurate forecasts in complex situations.

Prediction markets (1)

Prediction markets—also known as betting markets, information markets, and futures markets—have been used for forecasting since the 16th century (Rhode and Strumpf 2014). They attract experts who are motivated to use their knowledge to win money by making accurate predictions, thus being less likely to be biased. The forecasts are rapidly updated when news leads people to place bets. However, the markets can be manipulated.

Prediction markets are especially useful when knowledge is dispersed and many motivated participants are trading. In addition, they provide rapidly revised forecasts when new information becomes available. Forecasters using prediction markets will need to be familiar with designing online prediction markets, as well as with evidence-based survey design.

The accuracy of forecasts from prediction markets was tested in eight published comparisons in the field of business forecasting. Errors were 28% lower than those from no-change models, but 29% higher than those from combined judgmental forecasts (Graefe 2011). In another test, forecasts from the Iowa Electronic Market (IEM) prediction market across the three months before each U.S. presidential election from 2004 to 2016 were, on average, less accurate than forecasts from the RealClearPolitics poll average, a survey of experts, and citizen forecasts (Graefe 2017a). (We suspect the $500 limit on each bet reduces the number and motivation of participants and thus makes the market less efficient.) Comparative accuracy tests based on 44 elections in eight countries (not counting the U.S.) were more accurate than forecasts by experts, econometric models, and polls, although still less accurate than citizen expectations (Graefe 2017b).

In addition, prediction markets contributed substantially to improving the accuracy of combined forecasts of voting for political candidates.

Judgmental Bootstrapping (2)

Judgmental bootstrapping was discovered in the early 1900s, when it was used to make forecasts of agricultural crops. The method uses regression analysis with the variables that experts use to make judgmental forecasts. The dependent variable is not the actual outcome, but rather the experts' predictions of the outcome given the values of the causal variables.

The first step is to ask experts to identify causal variables based on their domain knowledge. Then ask them to make predictions for a set of hypothetical cases. For example, they could be asked to forecast which students were most likely to be successful doctoral candidates. By using hypothetical information on a variety of alternative potential candidates, the forecaster can ensure that the causal variables vary substantially and independently of one another. Regression analysis is then used to estimate the parameters of a model with which to make forecasts. In other words, judgmental bootstrapping is a method to develop a model of the experts' forecasting procedure.

In comparative studies to date, the bootstrap model's forecasts are more accurate than those of the experts. It is like picking oneself up by the bootstraps. This occurs because the model is more consistent than the expert in applying the expert's rules. It addition the model does not get distracted by irrelevant features, nor does it get tired or irritable. Finally, the forecaster can ensure that the model excludes irrelevant variables.

Judgmental bootstrapping models are especially useful for forecasting problems for which data on the dependent variable—such as sales for a proposed product—are not available. Once developed, the bootstrapping model can provide forecasts at a low cost and make forecasts for different situations—e.g., by changing the features of a product.

Judgmental bootstrapping can reduce bias in hiring and promoting employees by insisting that the variables are only included if they have been shown to be relevant to performance. Alternatively, making the model available to potential candidates would provide them with a low-cost way to assess their prospects and avoid the ethical problem of undeclared selection criteria.

Despite the discovery of the method and evidence on its usefulness, its early use seemed to have been confined to agricultural predictions. It was rediscovered by social scientists in the 1960s who examined its forecast accuracy. A review of those studies (Armstrong 2001a) found that judgmental bootstrapping forecasts were more accurate than those from unaided judgments in 8 of 11 comparisons, with two tests finding no difference and one finding a small loss in accuracy. The typical error reduction was about 6%. (The one failure occurred when the experts relied on an irrelevant variable that was not excluded from the bootstrap model.) A later study that compared financial analysts' recommendations with recommendations from bootstrap models of the analysts also found that trades based on the models' recommendations were more profitable than those by the analysts (Batchelor and Kwan 2007).

In the 1970s, in a chance meeting on an airplane, the first author sat next to Ed Snider, the owner of the Philadelphia Flyers hockey team. When Armstrong asked how Snider selected players, he said that he visited the Dallas Cowboys football team to find out why they were so successful. As it happened, were using a method similar to judgmental bootstrapping. Snider then asked his managers to use it, but they refused. He did, however, convince them to use both methods as a test. When they saw that the model made better draft picks, the managers changed to judgmental bootstrapping. Snider said that the other hockey team owners knew what the Flyers were doing, but they preferred to continue using their unaided judgments.

In 1979, when the first author was visiting a friend, Paul Westhead, then coach of the Los Angeles Lakers basketball team, he suggested the use of judgmental bootstrapping. Westhead was interested, but was unable to convince the owner. In the 1990s, a method apparently similar to judgmental bootstrapping—regression analysis with variables selected by experts—was adopted by the general manager of Oakland Athletics baseball team. It met with fierce resistance from baseball scouts, the experts who historically used a wide variety of data along with their judgment. The improved forecasts from the regression models were so profitable, however, that almost all professional sports teams now use some version of this method. Those that do not, pay the price in their won-loss tally.

Despite the evidence, judgmental bootstrapping appears to be ignored by businesses, where the "won-lost record" is not clear-cut. It is also ignored by universities for their hiring decisions despite the fact that one of the earliest validation tests showed that it provided a more accurate and less expensive way to decide who should be admitted to PhD programs (Dawes 1971).

Multiplicative decomposition (3*)*

Multiplicative 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 different methods are appropriate for forecasting each individual part, or when relevant data can be obtained for some parts of the problem, or when the directional effects of the causal factors differ among the components.

These conditions seem to be common and the decomposition principle had long been a key element in textbooks on decision-making. To assess the effect size of the method, an experiment was conducted to compare the accuracy of global estimates vs. estimates for elements of the decomposition. Five problems were drawn from an almanac, such as "How many packs (rolls) of Polaroid color films do you think were used in the United States in 1970?" Some subjects were asked to make global estimates while others were asked to estimate each of the decomposed elements. Decomposition did not harm accuracy for any of the five problems and the error reductions were enormous (Armstrong, Denniston and Gordon 1975) Additional testing by others added support to these findings (MacGregor, 2001).

Multiplicative decomposition is a general problem structuring method that can also be used in conjunction with other evidence-based methods listed in Exhibit 1 for forecasting the component parts.

Intentions surveys (4)

Intentions surveys ask people how they plan to behave in specified situations. Data from intentions surveys 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, and a meta-analysis of ten meta-analyses with data from over 83,000 subjects each found a strong relationship between people's intentions and their behavior (Kim and Hunter 1993; Sheeran 2002).

Intentions surveys are especially useful when historical data are not available, such as for new products. They are most likely to provide accurate forecasts when the forecast time-horizon is short, and the behavior is familiar and important to the respondent, such as with buying durable goods or voting for a presidential candidate. Plans are less likely to change when they are for the near future (Morwitz 2001; Morwitz, Steckel, and Gupta 2007).

To assess the intentions of others, prepare an accurate but brief description of the situation (Armstrong and Overton 1977). Intentions should be obtained by using probability scales such as 0 = 'No chance, or almost no chance (1 in 100)' to 10 = 'Certain, or practically certain (99 in 100)' (Morwitz 2001). 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 (2014).

Response error is often a large component of total error. The problem is acute when the situation is new to the people responding to the survey, as when forecasting demand for a new product category.

Expectations surveys (5*)*

Expectations surveys ask people how they *expect* themselves, or others, to behave. Expectations differ from intentions because people know that unintended events might interfere. For example, if you were asked whether you intend to purchase a vehicle over the next year, you might say that you have no intention to do so. However, you realize that your present car might be involved in an accident or that your financial position might change for the better, so you might expect a 15% chance that you will purchase a new car. As with intentions surveys, forecasters should use probability scales, follow evidence-based procedures for survey design, use representative samples, obtain high response rates, and correct for non-response bias by using extrapolation across waves (Armstrong and Overton 1977).

Following the U. S. government's prohibition of prediction markets for political elections, expectation surveys were introduced for the 1932 presidential election (Hayes, 1936). A representative sample of potential voters was asked how they expected *others* might vote. These "citizen expectations" surveys have predicted the popular vote winners of the U.S. Presidential elections from 1932 to 2012 on 89% of the 217 surveys (Graefe 2014) and won again in 2016.

Further evidence was obtained from the PollyVote project. Over the 100 days before the election, the error of citizens' expectations forecasts of the popular vote in seven U.S. Presidential elections from 1992 through 2016 averaged 1.2 percentage points compared to the error of *combined polls* of likely voters' intentions average of 2.6 (Graefe, Armstrong, Jones, and Cuzán 2017). Citizen forecasts are cheaper than the election polls because the respondents are answering for many other people, so the samples can be smaller. We expect that the costs of the few citizen surveys—typically only four per presidential election in the U.S.—would be a small fraction of one percent of the cost of the hundreds of intentions polls that are conducted.

Expert surveys (6*)*

Use written questions and instructions for the interviewers to ensure that each expert is questioned in the same way, thereby avoiding interviewers' biases. Word questions in more than one way in order to compensate for possible biases in wording. Pre-test each question to ensure that the experts understand what is being asked. Then average the answers from the alternative ways of asking each question. Additional advice on the design for expert surveys is provided in Armstrong (1985, pp.108-116).

Obtain forecasts from at least five experts. For important forecasts, use up to 20 experts (Hogarth 1978). That advice was followed in forecasting the popular vote in the four U. S. presidential elections from 2004 to 2016. Fifteen or so experts were asked for their expectations on the popular vote in several surveys over the last 96 days prior to each election. The average error of the expert survey forecasts was 1.6 percentage points versus 2.6 for the average error of combined polls (Graefe, Armstrong, Jones, and Cuzán 2017, and personal correspondence with Graefe).

Delphi is an extension of the above survey approach whereby the survey is given in two or more rounds with *anonymous* summaries of the forecasts and reasons provided as feedback after each round. Repeat the process until forecasts change little between rounds—two or three rounds are usually sufficient. Use the median or mode of the experts' final-round forecasts as the Delphi forecast. Software for the procedure is freely available at ForecastingPrinciples.com.

Delphi is attractive to managers because it is easy to understand and anonymous. It is relatively inexpensive because the experts do not need to meet. It has an advantage over prediction markets in that reasons are provided for the forecasts (Green, Armstrong, and Graefe 2007). Delphi is likely to be most useful when relevant information is distributed among the experts (Jones, Armstrong, and Cuzán 2007).

Delphi forecasts were more accurate than forecasts made in traditional meetings in five studies comparing the two approaches, about the same in two, and were less accurate in one. Delphi was more accurate than surveys of expert opinion 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 it in 12% (Rowe and Wright 2001.)

Simulated interaction (7)

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 union would respond to a contract offer by a company, or how a government would respond to demands by artists to be paid a pension by the government.

Simulated interactions can be conducted by using naïve subjects to play the roles. Describe the main protagonists' roles, prepare a brief description of the situation, and list possible decisions. Each participant in a simulation is given one of the roles and a description of the situation. The role-players are asked to engage in realistic interactions with the other role players, staying in their roles until they reach a decision. The simulations typically last less than an hour.

It is important to act out the roles. Another approach has been to "put yourself in the other person's shoes." U.S. Secretary of Defense Robert McNamara said that if he had done this during the Vietnam War, he would have made better decisions.[3] A test of that "role thinking" approach found no improvement in the accuracy of the forecasts relative to unaided judgment. Apparently, it is too difficult to think through the interactions of parties with divergent roles in a complex situation; active role-playing between parties is necessary to represent such situations with sufficient realism (Green and Armstrong 2011).

Relative to unaided expert judgment—the method usually used for such situations—simulated interaction reduced forecast errors on average by 57% for eight conflict situations (Green 2005). The conflicts used in the research included an attempt at the hostile takeover of a corporation, and a military standoff by two countries over access to water.

Structured analogies (8)

The structured analogies method involves asking ten or so experts in a given field to suggest situations that were similar to that for which a forecast is required (the target situation). The experts are given a description of the situation and are asked to describe analogous situations, rate their similarity to the target situation, and to match the outcomes of their analogies with possible outcomes of the target situation. An administrator takes the target situation outcome implied by each expert's top-rated analogy, and calculates the modal outcome as the forecast (Green and Armstrong, 2007). The method should not be confused with the common use of analogies to *justify an outcome* that is preferred by the forecaster or client.

Structured analogies were 41% more accurate than unaided judgment in forecasting decisions in eight real conflicts. These were the same situations as were used for research on the simulated interaction method described above, for which the error reduction was 57% (Green and Armstrong 2007).

A procedure akin to structured analogies was used to forecast box office revenue for 19 unreleased movies. 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 forecasts were less than half those of forecasts from simple and complex regression models (Lovallo, Clarke and Camerer 2012).

[3] From the 2003 documentary film, "Fog of War".

Responses to government incentives to promote laptop purchases among university students, and to a program offering certification on Internet safety to parents of high-school students were forecast by structured analogies. The error of the structured analogies forecasts was 8% lower than the error of forecasts from unaided judgment (Nikolopoulos, et al. 2015). Across the ten comparative tests from the three studies described above, the error reduction from using structured analogies averaged about 40%.

Experimentation (9)

Experimentation is widely used, and is the most realistic method to determine which variables have an important effect on the thing being forecast. Experiments can be used to examine how people respond to factors such as changes in the design of a government health care system. Different programs could be presented to subjects in laboratory experiment to see how buyers and sellers would respond. Alternatively, states could design their own plans, including sticking with the old plan, and each could be tracked to see which works best.

Laboratory experiments allow greater control than field experiments and they avoid revealing sensitive information. 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 using the different package in different geographical test markets. An analysis of experiments in the field of organizational behavior found that laboratory and field experiments yielded similar findings (Locke 1986).

Expert systems (10)

Expert systems involve asking experts to describe their step-by-step process for making forecasts. The procedure should be explicitly defined and unambiguous, such that it could be implemented using software. Use empirical estimates of relationships from econometric studies and experiments when available in order to help ensure that the rules are valid. The expert system should be simple, clear, and complete.

Expert system forecasts were more accurate than forecasts from unaided judgment in a review of 15 comparisons (Collopy, Adya and Armstrong 2001). Two of the studies—on gas, and on mail order catalogue sales—found that the expert systems errors were 10% and 5% smaller than those from unaided judgment. While little validation has been carried out, the method is promising.

Quantitative Methods

Quantitative methods require numerical data on or related to what is being forecast. However, these methods can also draw upon judgmental methods, such as decomposition.

Extrapolation (11)

Extrapolation methods use historical data only on the variable to be forecast. They are especially useful when little is known about the factors affecting the forecasted variable, or the causal variables are not expected to change much, or the causal factors cannot be forecast with much accuracy. Extrapolations are cost effective when many forecasts are needed, such as for production and inventory planning for thousands of products.

Exponential smoothing, which dates back to Brown (1959 & 1962), is a sensible approach to moving averages as it can use all historical data and it puts more weight on the most recent data.

Exponential smoothing is easy to understand and inexpensive to use. For a review of exponential smoothing, see Gardner (2006).

Damping a time-series trend toward the long-term historical trend or toward no trend often improves forecast accuracy. The greater the uncertainty about the situation, the greater is the need for damping. A review of ten experimental comparisons found that, on average, damping the trend toward zero reduced forecast error by almost 5% (Armstrong 2006). In addition, damping reduced the risk of large errors. Software for damping can be found at ForcastingPrinciples.com.

When there is a strong and consistent trend and the causal factors are expected to continue—as with the prices of resources (Simon 1996)—one can also damp toward the long-term trend.

When extrapolating daily, weekly, monthly, or quarterly data, remove the effects of seasonal influences first. Forecast the seasonally adjusted series, then multiply by the seasonal factors. In forecasts for 68 monthly economic series over 18-month horizons, seasonal adjustment reduced forecast errors by 23% (Makridakis et al. 1984, Table 14).

Given the inevitable uncertainty involved in estimating seasonal factors, they too should be damped. Miller and Williams (2003, 2004) provide procedures for damping seasonal factors. When they damped the seasonal adjustments that were made to the 1,428 monthly time series from the M3-Competition prior to forecasting, the accuracy of the forecasts improved for 59% to 65% of the series, depending on the horizon. These findings were successfully replicated by Boylan ey al. (2015). Software for the Miller-Williams procedures is freely available at ForPrin.com.

Damping by averaging seasonality factors across analogous situations also helps. In one study, combining seasonal factors from similar products such as snow blowers and snow shovels, reduced the average forecast error by about 20% (Bunn and Vassilopoulos 1999). In another study, pooling monthly seasonal factors for crime rates for six precincts in a city increased forecast accuracy by 7% compared to using seasonal factors that were estimated individually for each precinct (Gorr, Oligschlager, and Thompson 2003).

Multiplicative decomposition can be used to incorporate causal knowledge into extrapolation forecasts. For example, when forecasting a time-series, it often happens that the series is affected by causal forces–-growth, decay, opposing, regressing, supporting, or unknown—that affect trends in different ways. In such a case, decompose the time-series by causal forces that have different directional effects, extrapolate each component, and then recombine. Doing so is likely to improve accuracy under two conditions: (1) domain knowledge can be used to structure the problem so that causal forces differ for two or more of the component series, and (2) it is possible to obtain relatively accurate forecasts for each component. For example, to forecast motor vehicle deaths, forecast the number of miles driven—a series that would be expected to grow—and the death rate per million passenger miles—a series that would be expected to decrease due to better roads and safer cars—then multiply to get total deaths.

When tested on five time-series that clearly met the two conditions, decomposition by causal forces reduced *ex ante* forecast errors by two-thirds. For the four series that partially met the conditions, decomposition by causal forces reduced error by one-half (Armstrong, Collopy and Yokum 2005). There was no gain, or loss, when the conditions did not apply.

Additive decomposition of a time-series can be used to forecast the starting level and trend separately, and then add them—a procedure called “nowcasting.” Three comparative studies found that, on average, nowcasting reduced errors for *short-range* forecasts by 37% (Tessier and Armstrong, 2015). The percentage error reduction would be expected to decrease rapidly as the forecast horizon lengthened.

Rule-based forecasting (12)

Rule-based forecasting (RBF) allows the use of causal knowledge. To use RBF, first identify the features of the series. To date, 28 features have been tested and found useful—including the causal forces mentioned in the preceding section—and factors such as the length of the forecast horizon, the amount of data available, and the existence of outliers (Armstrong, Adya and Collopy 2001). These features are identified by inspection, statistical analysis, and domain knowledge. There are 99 rules for combining these different extrapolation forecasts.

For one-year-ahead *ex ante* forecasts of 90 annual series from the M-Competition (available from ForecastingPrinciples.com), the median absolute percentage error for RBF forecasts was 13% smaller than those from equally weighted combined forecasts. For six-year-ahead *ex ante* forecasts, the RBF forecast errors were 42% smaller, presumably due to the increasing importance of causal effects over longer horizons. RBF forecasts were also more accurate than equally weighted combinations of forecasts in situations involving strong trends, low uncertainty, stability, and good domain expertise. In cases where the conditions were not met, the RBF forecasts had little or no accuracy advantage over combined forecasts (Collopy and Armstrong 1992).

The contrary series rule is especially important. It states that when the expected direction of a time-series and the recent trend of the series are contrary to one another, set the forecasted trend to zero. The rule yielded substantial improvements in extrapolating time-series data from five data sets, especially for longer-term (six-year-ahead) forecasts for which the error reduction exceeded 40% (Armstrong and Collopy 1993). The error reduction was achieved even though the authors of that paper, who were not experts on the product categories, did the coding for “expected direction” of the trend; the domain knowledge of experts might provide more accurate forecasts, but that has not been tested.

Regression analysis (13)

As other sections of this paper describe, regression analysis plays an important and useful supporting role in forecasting as tool for quantifying relationships. In this section we discuss the use of regression analysis of nonexperimental data to develop causal models for forecasting, and present a checklist of guidelines.

Regression analysis can be useful for estimating the strength of relationships between the variable to be forecast and one or more *known* causal variables. Those estimates of the strengths of relationships can be useful in predicting the effects of policy changes and changes in the evironment, *if* the estimates are realistic representations of the important causal relationships. If not, the estimates are likely to result in misleading forecasts. Indeed, much literature published over the years has concluded that the validity of regression estimated model parameters has been poor. For example, Ioannidis, Stanley, and Doucouliagos (2015) reviewed research findings from 6,700 empirical economics studies that provided 64,076 estimates of effects across 159 topics, and concluded that the typical published effect size was likely to be exaggerated by a factor of two.

Moreover, regression analysis can be, and has been, abused for the purpose of advocating for preferred theories or policies, in violation of the core scientific principle of objectivity.

Regression analysis is in the sense that it reduces the size of coefficient estimates in response to random measurement error in the variables. Regression is blind, however, to other sources of error such as the omission of important causal variables, inclusion of irrelevant variables, intercorrelations and inaccurate forecasts of causal variables. Moreover, regression exaggerates the importance of outlier observations due to the use of the least-squares method, and correlations among causal variables confound the determination of their coefficients. Finally, changing a model specified on the basis of *a* *priori* analysis in order to improve statistical fit with historical data reduces predictive validity, and risks reports of humorous—or perhaps dangerously wrong—findings in the media.

Exhibit 2 provides a checklist of guidelines for using regression analysis for forecasting. The reasoning and evidence for the guidelines is described below.

Exhibit 2: Checklist for Developing Forecasting Models Using Regression Analysis

A priori analysis and model specification	
1. Use prior knowledge to identify relevant causal variables	□
2. Specify the direction and importance of causal variables' effects on the variable to be forecast	□
3. Discard a causal variable if it is unimportant, or cannot be forecast or controlled	□
4. Specify a model with a form and content that embodies prior knowledge	□
5. Specify a simple model that is understandable to those with a legitimate interest in the forecasts	□
Data analysis and estimation of the model	
6. Obtain all relevant data on the model variables	□
7. Ensure data are valid, reliable, and corrected for outliers, especially data on the variable to be forecast	□
8. Adjust coefficient estimates toward equal weights to compensate for intercorrelations	
9. Incorporate prior knowledge by averaging the regression coefficients with *a priori* coefficients	□
10. Combine forecasts from alternative multiple regression models or one-variable models	□
Validation of the forecasting model	
11. Estimate prediction intervals using *only* out-of-estimation-sample forecast errors	□
12. Provide access to all data, methods, revisions to hypotheses, and procedures to enable replication	□
13. Sign an oath that your analyses were ethical	□

1. Use prior knowledge to identify relevant causal variables— Follow the scientific method by using prior knowledge, including experimental findings, to identify causal variables before using regression analysis. Regression analysis is unsuited to identifying relationships from non-experimental data.

In some situations, causal factors are obvious from logical relationships. However, many causal relationships are uncertain. That is particularly the case with complex forecasting problems. If there are questions as to the validity of a proposed causal factor and its directional effect, one should consult published experimental research, especially meta-analyses of experimental findings. For example, opinions about gun control vary. While about two-thirds of the people in the U.S. believe that the right to carry guns reduces crime, other people believe the opposite. The opinions of voters and politicians have led states to change their laws over the years to variously restrict gun ownership and use, or to make ownership and use easier. These natural experiments provide a way to determine scientifically which opinion is correct, as was done by Lott (2010) and Lott (2016).

Do not go beyond obvious relationships and experimental evidence when searching for causal variables. Those cautions are especially important when forecasters have access to large databases with many variables.

Regression is not suited for identifying causal relationships from non-experimental data. Unfortunately, this improper use of regression analysis has been increasing. Ziliak and McCloskey's (2004) review of empirical papers published in the *American Economic Review* found that in the 1980s, 32% of the studies (N= 182) had relied on statistical significance for inclusion of variables in their models. The situation was worse in the 1990s, as 74% did so (N=137).

2. Specify the direction and importance of causal variables' effects on the variable to be forecast—The directional effects of some variables are obvious from logic or common knowledge on the domain. If the direction is not obvious, refer to the research literature. Failing that, survey say 3 to 5 domain experts.

Effect sizes are typically specified as elasticities so as to aid decision making. Elasticities are the percentage change in the variable to be forecast that would result from a one percent change in the causal variable. For example, a price elasticity of demand of -1.2 for beef would mean that a price increase of 10% would be expected to result in a decrease in the quantity demanded of 12%. To specify a model that has elasticity coefficients, convert the dependent and causal variables by taking logarithms of the values.

3. Discard a causal variable if it is unimportant, or cannot be forecast or controlled—This guideline should be obvious, and so requires no further explanation.

4. Specify a model with a form and content that embodies prior knowledge—Having done so, use your *a priori* estimates for all coefficients in the model and run a regression analysis to estimate constant term for this *a priori* model.

5. Specify a simple model that is understandable to those with a legitimate interest in the forecasts—This guideline is based on Occam's razor, which advises scientists to use simple models. Its effects on accuracy are substantial, as we show later in this paper.

6. Obtain all relevant data on the model variables—For example, when using time-series use data for all time periods unless a convincing case be made, prior to conducting the analyses, to do otherwise. Explain why this was done in the paper.

7. Ensure data are valid, reliable, corrected for outliers, especially data on the variable to be forecast—Be sure that the data provide valid measures of the dependent and causal variables. By reliable we mean that the measurement is free of large errors. One way to deal with the problem of data often not being a valid and reliable measure of the variable of interest is to find alternative measures of the same concept and average them. For example, to measure country A's exports to country B, average the reported figure with country B's reported imports from country A. As it happens, there are many situations where more than one data series can be used to measure the same conceptual variable, as Morganstern (1961) showed.

Outliers in data should be adjusted or removed in order to avoid their excessive influence on the least squares regression estimates (coefficients). There are a number of solutions; One is winsorizing, where outliers are reduced to the value of the most extreme observation in which you have confidence (Tukey 1962).

8. Adjust coefficient estimates toward equal weights to compensate for intercorrelations—Adjusting the causal variables coefficients toward equality (equalizing) is a conservative approach to dealing with the uncertainty introduced by intercorrelations among included variables. Equalizing coefficients requires that the variables are all measured on the same scale and positively correlated with the variable being forecast. To achieve that, standardize each variable by subtracting the mean of the variable from each observation and then divide by the variable's standard deviation. Reverse the signs on the observations of any causal variable that is negatively correlated with the dependent variable. Then use regression analysis to estimate the model's coefficients.

The first empirical demonstration on the power of equal weights was Schmidt's (1971). That was followed by Wainer's (1976) "Equal weights theorem" based on a statistical analysis leading to the conclusion that "Equally weighted linear models... are frequently superior" to least squares models (p. 214.) (Wainer was apparently unaware of Schmidt's empirical findings.) Additional empirical support was provided by Dana and Dawes (2004) in their analysis of large samples of real and synthetic data, and by Cuzán and Bundrick (2009) who found that Ray Fair's econometric forecasts for U.S. presidential elections were improved when equal (unit) weights were used. The equal weights approach was tested in election forecasting using an equal-weights model that included all of the 27 variables that had variously been used across ten independent econometric models. The model's *ex ante* forecast error was 29% lower than the error of the most accurate of the ten original regression models (Graefe 2015). Note that the econometric estimates in election studies typically have small sample sizes, usually less than 15.

While the findings have shown that "equal weights" often does better in out-of-sample tests of predictive validity, we suggest that *damping of regression coefficients toward equal weights* should be tailored to the forecasting problem and to prior knowledge on the relative importance of variables. We are not, however, aware of any direct research on this issue.

Equalizing was tested in election forecasting using eight independent econometric election forecasting models. Equalizing coefficients by between 10% and 60% reduced the Mean Absolute Errors of the forecasts relative to forecasts from the models with unadjusted coefficients for all of the models, and by between 70% and 100% reduced errors for seven of the eight models (Graefe, Armstrong, and Green 2014). On average, there was little gain or loss from equalizing by more than 50%: the typical error reduction from equalizing between 50% and 100% was about 5%.

9. Incorporate prior knowledge by averaging the regression coefficients with a priori coefficients—The procedure, sometimes called the "poor man's regression analysis" was described by Wold and Jureen (1955), who called it "conditional regression analysis." It involves averaging the modeller's *a priori* coefficients with the coefficients estimated from the data using regression analysis. Use zero for the estimated coefficient if it has the wrong sign (wrong direction). Use equal weights for averaging, unless there are strong a priori reasons for using differential weights.

10. Combine forecasts from alternative multiple regression models, or one-variable models—Consider the case of combining the forecasts of economists who ascribe to different economic theories. Combining the forecasts of the two economists whose theories were most different reduced the errors of Gross Domestic Product forecasts by 23%, whereas combining the forecasts of the two economists with the most similar theories reduced errors by 11%. Combining forecasts from economists who used different methods led to a 21% error reduction compared to a 2% error reduction for the forecasts from similar techniques (Table 2 in Batchelor and Dua 1995).

In another study, Graefe, Armstrong, and Green (2014) compared the accuracy of the forecasts from eight independent multiple regression models for forecasting the popular vote in U.S. presidential elections with the accuracy of an average of their forecasts. They found that the combined forecasts reduced error compared to the typical individual model's forecast by 36% across the 15 elections in the study.

Occam's razor suggests that the combination of forecasts from simple one-variable regressions might be more accurate than the forecasts from a multiple-variable regression model. Combining the forecasts of single-variable regressions overcomes some of the problems inherent in multiple regression. This was approach by making ten-year ahead ex *ante* population forecasts for 100 counties in North Carolina. They used six causal variables to develop six one-variable models, then calculated a combined forecast. They also developed a multiple regression with the same six variables. The Mean Absolute Percentage Error for the combined forecasts was 39% lower than that from the multiple regression. (Namboodiri and Lalu 1971.) The finding was surprising at the time, and perhaps even now for many statisticians. In general, of course, one comparative study is not enough to support a method.

That said, we are unaware of any empirical test that shows the superiority of multiple regression by testing against combined forecasts from one-variable regressions.

11. Estimate prediction intervals using only out-of-estimation-sample forecast errors—The prediction intervals are useful in deciding which methods are most accurate for forecasting. They also help in assessing uncertainty in forecasts.

12. Provide access to all data, methods, revisions to hypotheses, and procedures to enable replication—This is a basic scientific principle. It allows others to check your work and to make corrections.

13. Sign an oath that your analyses were ethical—Given the high level of cheating and questionable practices with regression, we recommend that analysts and authors voluntarily sign an oath that they will not engage in unethical or questionable behavior before starting a project, and then confirm that they have not done so after completion of the study. This procedure has been found to be effective (see Mazar, Amir and Ariely 2008).

The guidelines may be familiar to older readers who were acquainted with the econometric literature in an age when extensive *a priori* analysis was the norm. Researchers understood that any quantitative data they were able to obtain was likely to represent only a small part of the sum of knowledge about the situation they were concerned with For example, most of procedures described in the checklist of 14 regression guidelines were used in (Armstrong 1970b).

In those eary years, there was also a pragmatic reason for undertaking extensive a prior analysis in order to develop a model: the cost of regression analysis was very high. Analysts needed to compile the data manually from paper sources; keypunch the data; verify the keypunched data; write and keypunch computer code to describe the model; carry heavy boxes of punched cards to the computer center; fetch the cards and reams of paper detailing the results of the analysis; then repeat the whole process in the case of any mistake.

With the now low cost of regression analysis and easy access to much quantitative data in an electronic form, the pragmatic incentive for conducting laborious *a priori* analysis has gone. As a consequence—the improper use of regression analysis to identify statistical relationships rather than to measure known causal relationships using statistics—has been increasing. Ziliac and McCloskey's (2004) review of empirical papers published in the *American Economic Review* found that in the 1980s, 32% of the studies (N=182) had relied on statistical significance for inclusion of variables in their models. The situation was worse in the 1990s, as 74% did so (N=137). The proper number should be zero.The reasons for developing causal models by starting with a careful *a priori* analysis remains critical.

Segmentation (14)

Segmentation is based on additive decomposition. It involves breaking a problem into independent parts, using knowledge and data to make a forecast about each part, and then adding the forecasts of the parts. For a discussion of the process, see Armstrong (1985, pp. 249-270). For example, to forecast air travel demand in ten years' time, the Port of New York Authority in 1955 divided airline travelers into 130 business travelers and 160 personal traveler segments. The personal travelers were split by age, occupation, income, and education, and the business travelers by occupation, then industry, then income. Data on each segment was obtained from the census and from a survey on travel behavior. To derive the forecast, the official projected air travel population for 1965 was allocated among the segments, and the number of travelers and trip frequency were extrapolated using 1935 as the starting year with zero travelers. The forecast of 90 million trips was only 3% different from the

1965 actual figure (Armstrong 1985). Note that the Port Authority forecasting project is presented only as an example of an application, not as evidence on its value.

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 more cut-points should be used when there are the non-linearities in the relationships, and fewer cut points should be used when the samples sizes are smaller. Forecast the population of each segment and the behavior of the population within each segment by using the typical behavior or by using regression models within segment. Then combine the population and behavior forecasts for each segment, and then sum across segments.

Segmentation has advantages over regression analysis for situations where variables are intercorrelated, the effects of variables are non-linear, prior causal knowledge is good, and the sample sizes in each of the segments are large. These conditions occurred to a reasonable extent in a study where data from 2,717 gas stations were used to develop a regression model and a segmentation model. Data were available on nine binary variables and ten other variables including such variables as type of area, traffic volumes, length of street frontage, presence or not of a canopy, and whether or not the station was open 24 hours a day. The predictive validities of the two methods were then tested using a holdout sample of 3,000 stations. The error of the regression model was 58%, while that of the segmentation model was 41% (Armstrong and Andress 1970). Clearly this is a method that benefits from having much data.

Index models (15)

Some forecasting problems are characterized by having many important causal variables. Consider, for example, predicting which players will do well in sports, who would be an effective company executive, which countries will have the highest economic growth, and which applicants for immigration would pose a security risk. Regression analysis cannot take advantage of information about many variables.

Benjamin Franklin proposed a sensible solution. He described his method in a letter to his friend Joseph Priestley, who had written to Franklin about a “vexing decision” that he was struggling to make. The method is basically a pros and cons list with subjective weights on the variables; Franklin called it prudential algebra. A similar method, called experience tables, was used in the early 1900s for deciding which prisoners should be given parole (Burgess 1936). This method gave way to regression analysis. However, Gough (1962) tested the method against a regression model for making predictions about parolee success, and found that regression provided no improvements in accuracy.

We developed a version of Franklin’s prudential algebra that we refer to as the index method. It is ideally designed for studies where there are many important and well-known causal variables with known directional effects. Index models require good prior knowledge about the direction of the effects of the variables (Graefe and Armstrong, 2011). Use prior experimental evidence and domain knowledge to identify predictor variables and to assess each variable’s directional influence on the outcome. If prior knowledge on a variable’s effect is ambiguous or is not obvious, or is contrary to experimental findings, do not include the variable in the model.

The primary advantage of the index method is that one can include all important variables, no matter how many there are. The disadvantage is that a time-consuming search for relevant experimental evidence on the causal variables is often necessary.

Index scores are the sum of the values across the variables, which might be coded as 1 or 0 (favorable or unfavorable). The alternative with a higher index score is likely to turn out better, or is more likely to occur. Where sufficient historical data are available, one can then obtain quantitative forecasts by regressing index values against the variable of interest, such as the economic growth of nations and states. Unit weights should be used unless one has strong evidence that differential weighting will help.

Index models with unit weights have been found to be more accurate than regression models with "optimal weights" estimated from historical data. Graefe and Armstrong (2013) identified empirical studies that included comparisons of the two methods for forecasting problems in applied psychology, biology, economics, elections, health, and personnel selection. The index model forecasts were more accurate than regression model forecasts for ten of the thirteen studies.

One index model (Lichtman 2005) uses 13 variables selected by an expert to forecast the popular vote in U. S. Presidential elections. When tested on the 40 elections from 1860 through 2016, it was found to be correct on all elections except for 2016. See Armstrong and Cuzán (2006) for an analysis of this method. We are not aware of any other method that has matched this level of accuracy.

Another test assessed the predictions from an index model of the relative effectiveness of the advertising in 96 pairs of advertisements. There were 195 variables that were potentially relevant, so regression was not feasible. Guessing would result in 50% correct predictions. Judgment by novices was little better than guessing at 54%, expert judges made 55% correct predictions, and copy testing yielded 57%. By contrast, the index method was correct for 75% of the pairs of advertisements (Armstrong, Du, Green and Graefe 2016).

Combined Forecasts

The last two methods listed in Exhibit 1 deal with combining forecasts. We regard them as the most important methods to improve accuracy.

The basic rules for combining across methods are: (1) to use all valid evidence-based methods—and no other methods; (2) seek diverse data and procedures to use with the methods; (3) use equal weights for the components, unless strong evidence is available that some methods provide more accurate forecasts than others; (4) specify the weights and reasons for them *before* obtaining the forecasts (Graefe, Armstrong, Jones, and Cuzán 2014; Graefe 2015); and (5) use many components: and (6) averages are typically used but a trimmed mean might be appropriate when uncertainty is high. For important forecasts, we suggest using at least three components.

The above procedures *guarantee* that the resulting forecast will not be the worst forecast, and that it will perform at least as well as the typical component forecast. In addition, the absolute error of the combined forecast will be smaller than the average of the component forecast errors when the components bracket the true value. Finally, combining can be more accurate than the most accurate component—and that often occurs.

Combining is not intuitive. Most people believe that a combined forecast provide only average accuracy. For example, a paid panel of 203 U.S. adults was each asked to choose from five members of a campus film committee whose forecasts they would prefer to include in calculating average forecasts of attendance for proposed movie showings. The participants were given data on each of the committee members' recent forecast errors. Only 5% of the participants chose to ask for forecasts from all five members of the committee. The rest chose to include forecasts only from film committee members whose previous errors had been smallest (Mannes, Soll, and Larrick 2014). With the same intuition, when New York City officials received two different forecasts for an impending snowstorm in January 2015, they acted on the forecast that they believed would be the best—as it turned out, it was the worst. The counterintuitive conclusion on the effects of combing is the consequences of the effects of bracketing: a situation in which forecasts lay on opposite sides of the actual value. (You can create your own examples and you will see this.) Because bracketing is always possible, even knowing for sure which method would provide the most accurate forecast cannot not offer the guarantees that are offered by combining. Thus, when there are two or more forecasts *from evidence-based methods*, the method of combining forecasts should always be used.

Combining forecasts from a single method (16)

Combining forecasts from variations of a single method or from independent forecasters using the same method helps to compensate for errors in the data, mistakes, and small sample sizes in any of the component forecasts. In other words, combining within a single method is likely to be useful primarily in improving the reliability of forecasts. Given that a particular method might tend to produce forecasts that are biased in a consistent direction, however, forecasts from a single method are less likely to bracket the true value than are forecasts from different methods.

One review identified 19 studies that compared combinations of forecasts from a single method. The average error of the combined forecasts was 13% smaller than the average error of the typical component forecast, with a range from 3% to 24% (Armstrong 2001c).

Combining forecasts from several methods (17)

Different forecasting methods are likely to have different biases because they use different information and make different assumptions about relationships. As a consequence, forecasts from diverse methods are more likely than those from a single method to bracket the actual outcome. (They also use more information, which aids reliability.)

Armstrong, Morwitz and Kumar (2000) examined the effect of combining time-series extrapolations and intentions forecasts on accuracy. They found that combining forecast from the two different methods reduced errors by one-third compared to extrapolation forecasts alone.

The election-forecasting project (PollyVote.com) provided data for assessing the value of combining forecasts across up to six different methods. The combined forecast was more accurate than the *best* of the 4 to 6 component methods for each of the 100 days prior to the last six U.S. Presidential elections. Also, at 100 days prior to the election, the PollyVote combination of combined forecasts missed the actual outcome by an average of only one percentage point over the six elections. (see the PollyGraph at PollyVote.com).

GOLDEN RULE OF FORECASTING: BE CONSERVATIVE

The short form of the Golden Rule of Forecasting is to *be conservative*. The long form is to be conservative by adhering to cumulative knowledge about the situation and about forecasting methods. A conservative forecast is consistent with cumulative knowledge about the present and the past. To be conservative, forecasters must seek out and use all knowledge relevant to the problem, including knowledge of valid forecasting methods (Armstrong, Green, and Graefe 2015). The Golden Rule of Forecasting applies to all forecasting problems.

The Golden Rule of Forecasting is like the traditional Golden Rule, an ethical principle, that could be rephrased as “forecast unto others as you would have them forecast unto you.” The rule is especially useful when objectivity must be demonstrated, as in legal disputes or public policy disputes (Green, Armstrong, and Graefe 2015).

Exhibit 3 is a revised version of Armstrong, Green, and Graefe's Table 1 (2015, p.1718). It includes 28 guidelines logically deduced from the Golden Rule of Forecasting. Our literature search found evidence on the effects of 19 of the guidelines. On average the use of a typical guideline reduced forecast error by 28%. Stated another way, the *violation of a typical guideline increased forecast error by 39% on average.*

INSERT EXHIBIT 3 about here

We made several changes to the 2015 Golden Rule of Forecasting. First, we changed the Golden Rule Checklist item on adjusting forecasts (guideline 6 in Exhibit 3) to, "Avoid adjusting forecasts." Why? Adjusting forecasts risks introducing bias. Bias is particularly likely to occur when forecasters or clients might be subject to incentives to produce a given forecast.

Bias is common. For example, a survey of nine divisions within a British multinational firm found that 64% of the 45 respondents agreed that, "forecasts are frequently politically modified" (Fildes and Hastings 1994). In another study, 29 Israeli political surveys were classified according to the independence of the pollster from low to high as "in-house," "commissioned," or "self-supporting." The greater the pollsters' independence, the more accurate were their predictions. For example, 71% of the most independent polls had a relatively high level of accuracy, whereas 60% of the most dependent polls had a relatively low level of accuracy (Table 4, Shamir 1986).

Research in psychology has examined the effects of subjective adjustments to forecasts on accuracy. Meehl's (1954) conclusion from his review of that research was that forecasters should not make subjective adjustments to forecasts made by quantitative methods. He illustrated his conclusion with an allegory: "You're in the supermarket checkout lane, piling up your purchases. You don't say, 'This looks like $178.50 worth to me'; you do the calculations. And once the calculations are done, you don't say, 'In my opinion, the groceries were a bit cheaper, so let's reduce it by $8.00.' You use the calculated total." Research in psychology since then continues to support Meehl's findings (see Grove et al. 2000).

Another reason for the change in Guideline 6 is that if one follows the advice in this paper to use several valid methods, determining what is information is missing from the forecast and how to adjust for that would be problematic. Most importantly, we have been unable to find any evidence that adjustments would reduce forecast errors *relative to the errors of forecasts derived in ways that were consistent with the guidance in the checklists presented in Exhibits 1 and 2 above*. Research on adjusting forecasts from even single statistical models found that adjustments often increase errors (e.g., Belvedere and Goodwin 2017; Fildes et al. 2009) or have mixed results (e.g., Franses 2014; Lin, Goodwin, and Song 2014).

Take a problem that is often dealt with by judgmentally adjusting a statistical forecast: forecasting sales of a product that is subject to periodic promotions (see, e.g., Fildes and Goodwin 2007). The need for adjustment could be avoided by decomposing the problem into one of forecasting the level, trend, and effect of promotions separately. Trapero, Pedregal, Fildes, and Kourentzes (2013) provides support for that approach, finding an average reduction of mean absolute errors of about 20% compared to adjusted forecasts.

Exhibit 3: Golden Rule of Forecasting--Checklist II

			Comparisons*		
		Guideline	**N**	**Error reduction**	
1.		**Problem formulation**		**n**	**%**
1.1		Use all important knowledge and information by…			
1.1.1	□	selecting evidence-based methods validated for the situation	7	3	18
1.1.2	□	decomposing to best use knowledge, information, judgment	17	9	35
1.2		Avoid bias by…			
1.2.1	□	concealing the purpose of the forecast	–		
1.2.2	□	specifying multiple hypotheses and methods	–		
1.2.3	□	obtaining signed ethics statements before and after forecasting	–		
1.3	□	Provide full disclosure for independent audits, replications, extensions	1		
2.		**Judgmental methods**			
2.1	□	Avoid unaided judgment	2	1	45
2.2	□	Use alternative wording and pretest questions	–		
2.3	□	Ask judges to write reasons against the forecasts	2	1	8
2.4	□	Use judgmental bootstrapping	11	1	6
2.5	□	Use structured analogies	3	3	57
2.6	□	Combine independent forecasts from diverse judges	18	10	15
3.		**Extrapolation methods**			
3.1	□	Use the longest time-series of valid and relevant data	–		
3.2	□	Decompose by causal forces	1	1	64
3.3		Modify trends to incorporate more knowledge if the…			
3.3.1	□	series is variable or unstable	8	8	12
3.3.2	□	historical trend conflicts with causal forces	1	1	31
3.3.3	□	forecast horizon is longer than the historical series	1	1	43
3.3.4	□	short and long-term trend directions are inconsistent	–		
3.4		Modify seasonal factors to reflect uncertainty if…			
3.4.1	□	estimates vary substantially across years	2	2	4
3.4.2	□	few years of data are available	3	2	15
3.4.3	□	causal knowledge about seasonality is weak	–		
3.5	□	Combine forecasts from diverse alternative extrapolation methods	1	1	16
4.		**Causal methods**			
4.1	□	Use prior knowledge to specify variables, relationships, and effects	1	1	32
4.2	□	Modify effect estimates to reflect uncertainty	1	1	5
4.3	□	Use all important variables	5	4	45
4.4	□	Combine forecasts from diverse causal models	5	5	22
5.	□	**Combine combinations of forecasts from diverse methods**	16	15	18
6.	□	**Avoid adjusting forecasts**	–		
		Totals and Unweighted Average by Guideline 1 to 6	**106**	**70**	**28**

*** N**: Number of papers with findings on effect direction.
n: Number of papers with findings on effect size. **%**: Average effect size (geometric mean)

Expert and non-expert raters can complete the Golden Rule of Forecasting Checklist in less than an hour the first time they use it, and in less time after becoming familiar with it. Forecasters must fully disclose their methods and clearly explain them (Guideline 1.3). To help ensure the reliability of the checklist ratings, ask at least three people, each working independently, to do the rating.

SIMPLICITY IN FORECASTING: OCCAM'S RAZOR

Occam's razor is a principle of science attributed to 14th century scholar William of Ockham, but was earlier proposed by Aristotle. The principle is that the simplest explanation is best. The principle applies to forecasting: forecasters should use methods that are no more complex than needed.

Do forecasters believe in Occam's razor? In 1978, twenty-one of the world's leading experts in econometric forecasting were asked whether more complex econometric methods produced more accurate forecasts than simple methods, 72% replied that they did. In that survey, "complexity" was defined as an index reflecting the methods used to develop the forecasting model: (1) the use of coefficients other than 0 or 1; (2) the number of variables; (3) the functional relationship; (4) the number of equations; and (5) whether the equations involve simultaneity (Armstrong 1978).

A series of tests across different kinds of forecasting problems—such as forecasting of high school dropout rates—found that simple heuristics were typically at least as accurate as complex forecasting methods, and often more accurate (Gigerenzer, Todd et al. 1999).

In a recent paper, we (Green and Armstrong 2015) proposed a new operational definition of simplicity, one that could be used by any client. It consisted of a 4-item checklist to rate simplicity in forecasting as the *ease of understanding by a potential client*. The checklist was created before any analysis was done and it was not changed as a result of testing. Exhibit 4 provides an abridged version of the checklist provided on ForecastingPrinciples.com.

Exhibit 4: "Simple Forecasting" Checklist

Are the descriptions of the following aspects of the forecasting process sufficiently uncomplicated as to be easily understood by decision makers?	Simplicity rating (0…10)
1. method	[__]
2. representation of cumulative knowledge	[__]
3. relationships in models	[__]
4. relationships among models, forecasts, and decisions	[__]
Simple Forecasting Average (out of 10)	[___]

We found 32 published papers that allowed for a comparison of the accuracy of forecasts from simple methods with those from complex methods. Four of those papers tested judgmental methods, 17 tested extrapolative methods, 8 tested causal methods, and 3 tested forecast combining methods. The findings were consistent across the methods with a range from 24% to 28%. On average across each comparison, the more complex methods had *ex ante* forecast errors that were 27% higher than the simpler methods. This was surprising because the papers appeared to be proposing the more complex methods in the expectation that they would be more accurate.

ASSESSING FORECAST UNCERTAINTY

The uncertainty of a forecast affects its utility. For example, if demand for automobiles is forecast to increase by 20% next year, firms might consider hiring more employees and investing in more machinery. If the forecast had a high level of uncertainty such that a decline in demand is also likely, however, expanding operations might not be prudent.

Exhibit 5 presents a checklist for assessing uncertainty. We discuss these items below.

Exhibit 5: Assessing Forecast Uncertainty Checklist

❑ **1.** Avoid tests of statistical significance

❑ **2.** Avoid assessing uncertainty on the basis of statistical fit with historical data

❑ **3.** Estimate prediction intervals or likelihoods using *ex ante* forecasts for the situation

❑ **4.** Estimate prediction intervals or likelihoods using *ex ante* forecasts for analogous situations

❑ **5.** Obtain judgmental estimates of uncertainty from people familiar with the situation

Avoid tests of statistical significance for assessing uncertianty

It is not possible to accurately estimate forecast uncertainty using statistical significance alone. Any attempt to do so will likely lead to confusion, as well as poor decision-making. This is supported by extensive literature published over a range of over more than half a century. McShane and Gal (2015), for example, provide a short review, and describe an experiment that illustrates the fallibility of statistical significance testing. The experiment consisted of presenting leading researchers with a treatment difference between two drugs, as well as a p value for said difference, and asking which drug they would recommend to a potential patient. When the treatment difference was large, and reported to be $p > 0.05$, nearly half responded that they would advise that there was no difference between the two drugs. In contrast, when the difference between the treatment effects was *small*, but reported to be statistically significant ($p < 0.05$), 87% of the respondents replied that they would advise taking the drug that appeared to be only slightly more effective in the trial.

Another study found that even when the findings of two studies were presented only as confidence intervals, more than a quarter (26) of the 96 responses from psychology academics involved interpreting the findings in terms of statistical significance. Of that 26, all but 2 (8%) failed to recognize the two studies had "broadly consistent" results that "don't conflict". In contrast, 79% of the 61 responses that interpreted the findings in terms of the confidence intervals recognized that the studies both supported a positive finding (Coulson, Healey, Fidler and Cumming, 2010).

Avoid statistical fit with historical data

In a study using data consisting of 31 observations on 30 variables, stepwise regression was used with a rule that only variables with a *t* statistic greater than 2.0 would be included in the model. The final regression had eight variables and an R^2 (adjusted for degrees of freedom) of 0.85; in other words, the statistical fit was good. The data, however, were from Rand's book of random numbers (Armstrong 1970). A number of studies have used real world data to show that fit does not provide evidence on out-of-sample predictive validity (e.g., Pant and Starbuck 1990). Analysts should also ignore other statistical significance measures, such as *t*, *p*, and *F* (Soyer and Hogarth 2012).

Estimate prediction intervals using *ex ante* forecasts

Uncertainty is most accurately represented using empirical prediction intervals based on *ex ante* forecast errors from the same or analogous forecasting situations (Chatfield 2001). Simulate the actual forecasting procedure as closely as possible, and use the distribution of the errors of the resulting *ex ante* forecasts to assess uncertainty. For new situations, such as new product forecasts, assess uncertainty using the accuracy of forecasts for analogous situations.

Obtain judgmental estimates of uncertainty from people familiar with the situation

One common judgmental approach is to ask a diverse group of experts to express their confidence in their own judgmental forecasts in the form of 95% predictions intervals. Experts are typically overconfident about the accuracy of their forecasts. An analysis of judgmental confidence intervals for economic forecasts from 22 economists over 11 years found the actual values fell outside the range of their 95% individual confidence intervals about 43 percent of the time (McNees 1992). To reduce overconfidence, ask the experts to include all sources of uncertainty. Then ask them to list reasons why they might be wrong. These procedures have been shown to be effective in reducing overconfidence (Arkes 2001).

To improve the calibration of forecasters' estimates of uncertainty, ensure that they receive timely, accurate and well summarized information on what actually happened, and reasons why their forecasts were right or wrong. Weather forecasters use such procedures and their forecasts are well-calibrated for a few days ahead: When they say that there is a 40% chance of rain, on average over a large sample of forecasts, rain falls 40% of the time (Murphy and Winkler 1984).

Estimates of the standard error of survey research, such as those provided for election polls, are typically overconfident. That is to be expected, as such estimates are based only on the uncertainty due to sampling error; they ignore response error and non-response error. These error sources are typically at least as large as the sampling error. For example, in election forecasting, the empirically estimated confidence intervals tend to be about twice as large as those reported by election forecasters (Buchanan, 1986).

When uncertainty is high—such as with forecasting the effects of a change in a government regulation—response error is also likely to be high due to survey respondents' lack of self-knowledge about how they make decisions (see Nisbett and Wilson 1977). Response error is likely to be the primary source of error for new products.

Non-response error leads to large errors because the people who are most interested in the topic of the survey are much more likely to respond. While the error can be reduced to some extent by the "extrapolation-across-waves" method (Armstrong and Overton 1977), forecasters still need to consider this source of error in their assessment of uncertainty.

Quantitative estimates

Traditional statistical confidence intervals estimated from historical data are usually too narrow. One study showed that the percentage of actual values that fell outside the 95% confidence intervals for extrapolation forecasts was often greater than 50% (Makridakis, Hibon, Lusk, and Belhadjali 1987).

The confidence intervals estimated from the fit of a regression model are of no value to forecasters as a measure of the prediction interval estimated from out-of-sample accuracy (Pant and Starbuck 1990; Soyer and Hogarth 2012).

As with analyses of judgmental forecasts, regression models ignore key areas of uncertainty such as the omission of key variables, the difficulty in controlling or forecasting the causal variables, inability to make accurate forecasts of the causal variables, and the difficulty of assessing the relative importance of causal variables that are correlated with one another. These problems are magnified when analysts strive for a close fit, and even more so when data mining techniques strive for a close fit.

One way to address the problem of biases and multiple sources of error is to use a variety of different forecasting methods in the expectation that the sources of bias will vary across methods. We were unable to find any testing of this proposal in the literature. However, initial small sample tests in the PollyVote project suggest that prediction intervals based on using the standard deviation of forecasts provided by six different forecasting methods were well calibrated. This approach seems sensible, but must be considered as speculative at this time.

Forecast errors in time series are often asymmetric, which makes estimating prediction intervals difficult. Asymmetry of errors is likely to occur when the forecasting model uses an additive trend. This problem is effectively dealt with by transforming the forecast and actual values to logarithms, calculating the prediction intervals using differences in the logged values, and presenting the results in actual values. The log-log transformation specifically makes models much easier to understand, because the coefficients represent elasticities. (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. Regardless, asymmetric errors are a problem for the planner, not the forecaster; the planner must assess the damages due to forecasts where the supply is too high versus those where it is too low.

METHODS LACKING PREDICTIVE VALIDITY

Those involved with forecasting will notice that some commonly used methods were not discussed above**.** That is because we were unable to find evidence that they have improved accuracy relative to validated alternatives. Inaccurate forecasts can lead decision makers to threats or to miss opportunities. They can also lead to decisions that cause serious harm in the same way that a flawed medical diagnosis or the misspecification of a bridge structure can. For example, in Australia, expert opinions that dams would never fill again due to ongoing drought led governments to build desalination plants instead of more dams. Subsequent rains filled dams to overflowing, and the desalination plants have not been used.[4]

[4] Bolt, A., "An excuse from Flannery for his dud prediction." *Herald Sun*, March 1, 2012 and Bolt, A., "Floods sink climate change hysteria." *Herald Sun*, September 14, 2016.

Exhibit 6
Checklist of Methods to Avoid:
Popular methods that lack predictive validity* and ignore principles

		Principles Ignored	
		Occam's Razor	**Golden Rule**
Avoided	**Method**		
	Judgmental		
❑	Unaided judgment		✓
❑	Focus groups		✓
❑	Scenarios		✓
❑	Game theory	✓	✓
	Quantitative		
❑	Conjoint analysis		✓
❑	Box-Jenkins / ARIMA	✓	✓
❑	Neural networks	✓	✓
❑	Data mining / Big data analytics	✓	✓

*Method has failed tests of *ex ante* forecasting accuracy relative to naïve, commonly used, and previously validated methods, or evidence of success in fair tests is not available.

Popular methods that lack evidence on predictive validity are listed in Exhibit 6. Some of the methods have been tested against validated methods and have been found to be less accurate. That is the case, for example, with the Box-Jenkins method: studies of comparative accuracy have found that Box-Jenkins did poorly relative to simpler and more comprehensible alternatives (e.g., Makridakis et al. 1984). The Box-Jenkins method is also difficult for reasonably intelligent people to understand, thus violating Occam's razor.

Some methods might have promise, but have not been tested adequately. One such method “quantitative analogies.” For example, to forecast the effect of the launch of a generic drug on sales of an established drug, one could use the percentage change in the sales of similar established drug brands that had previously been exposed to such competition. The method is consistent with Occam's razor and the Golden Rule of Forecasting, and has intuitive appeal. One study used data on the five closest analogies to forecast sales of each of 20 target products; it reduced error by 8% on average compared to forecasts from regression analysis (last line of Table 2 in Goodwin, Dyussekeneva, and Meeran 2013).

Unaided judgment

Research examining the accuracy of experts' judgmental forecasts dates back to the early 1900s. Those findings led to the Seer-Sucker Theory (Armstrong 1980): “No matter how much evidence exists that seers do not exist, suckers will pay for the existence of seers.” The Seer-Sucker Theory has held up well over the years; in particular, a 20-year comparative evaluation study provided support (Tetlock 2005). Examples of unaided judgment abound (Cerf and Navasky 1998). Consider, for example, that many people invest in hedge funds despite the evidence that the returns from the expert stock pickers' portfolios—especially after commissions and management fees are deducted—are inferior to those from a portfolio that mimics the stock market (Malkiel 2016).

Given the evidence to date, efforts to find better experts would be an inefficient use of resources. It would also slow the progress of scientific forecasting by keeping alive the belief that experts' unaided judgments are useful. As we have shown above, experts are vital when their expertise is used in structured ways.

Focus groups

Focus groups are a popular method used to forecast customers' behavior, such as demand for a proposed TV series or suggestions from clients for the design of a new building. However, there is no evidence to support the use of this method for forecasting. Furthermore, it violates forecasting principles. First, the participants are seldom representative of the population of interest. Second, samples sizes are small, usually less than ten people per group. Third, in practice, questions for the participants are typically poorly structured and untested. Fourth, the responses of participants are influenced by the expressed opinions of others in the group, and by the way the moderator poses the questions. Fifth, subjectivity and bias are difficult to avoid when summarizing the responses of focus group participants.

Scenarios

Scenarios are stories about the future in which the outcomes are described in detail and written in the past tense. This might be expected to help people think through the likelihood of an event. However, the addition of details to the story leads people to *inflate* the likelihood of the event (Gregory and Duran 2001), which violates the logic that the occurrence of an event "A", cannot be more likely to occur if one states that events "B" and "C" occur at the same time as "A". In short, the expectation is that forecasts from scenarios would be biased.

Our Google Scholar search for "scenarios," "accuracy," and "forecast" or "predict" in May 2017 yielded 740,000 results. However, we have been unable to find any validation studies to support its value for improving accuracy. A review of research findings was unable to find any comparative studies to support the use of scenarios as a way to forecast what will happen. Scenarios also violate the Golden Rule, especially because they ignore prior knowledge.

Game theory

Game theory involves thinking about the goals of various actors in a situation where there are conflicts oven decisions. It involves thinking about the incentives that motivate parties and deducing the decisions they will make. The method sounds like a plausible way to forecast the decisions people will make, such as in negotiations among market participants and regulators, and in conflict situations. Authors of textbooks and research papers recommend game theory to make forecasts about outcomes of conflicts.

We have been unable to find evidence to support the claim—often implied rather that stated directly—that game theory provides useful forecasts. In the only tests of forecast validity to date, game theory experts' forecasts of the decisions that would be made in eight real (disguised) conflict situations were no more accurate than students' unaided judgment forecasts (Green 2002; 2005). Despite this, game theory is widely used as a way of predicting behavior. For example, our May 2017 Google Scholar search using "game theory," "accuracy," and "forecast" or "predict," yielded 45,000 "results." Research is needed that identifies conditions under which game theory is valid.

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. For example, various features of a laptop computer such as price, weight, dimensions, and screen size could be varied substantially while ensuring that the variations in features do not correlate with one another. The potential customers' choices can be analyzed by regressing them against the product features.

Conjoint analysis is based on sound principles, such as using experimental design and soliciting intentions independently from a representative sample of potential customers, so there is reason to expect the method to be useful. Our Google Scholar search for "conjoint analysis", "accuracy," and "forecast" or "predict" in May 2017 yielded over 10,000 results. However, we have been unable to find tests comparing the *ex-ante* accuracy of conjoint-analysis forecasts with those from other reasonable methods. Wittink and Bergestuen (2001) also failed to find such evidence. Despite the lack of sufficient testing for conjoint analysis, we believe that this method offers promise and urge that its predictive validity be tested against other methods such as experts' judgmental forecasts—the most common practice—and judgmental bootstrapping.

Neural networks

"Neural networks" is a method designed to identify patterns in time-series and to use them to make forecasts. Studies on neural nets have been popular with researchers, and we found nearly 400,000 results in our May 2017 Google Scholar search for "neural networks," "accuracy," and "predict or forecast." Little evidence on comparative forecast accuracy has been published. Perhaps the fairest and largest comparison, however, was 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). The poor results are not surprising, given that neural nets ignore prior knowledge and violate Occam's razor.

Data mining / big data analytics

Although data mining began in the 1960s with step-wise regression, we have been unable to find experimental evidence that looking for statistical patterns in data has improved ex *ante* forecast accuracy. In addition, data mining does not employ any of the 28 guidelines in the Golden Rule. An extensive review and reanalysis of 50 real-world data sets also found no evidence that data mining is useful (Keogh and Kasetty 2003). Finally, data mining techniques are expected to provide unrealistic overconfidence in the forecasts.

The "big data" movement has diverted researchers away from searching for experimental findings on causal factors. In effect, the approach ignores prior knowledge (see Einhorn 1972). In our judgment, "big data analytics" is presently a substantial barrier to the use of evidence–based forecasting methods.

CONCLUSIONS

Clients who are interested in accurate forecasts should require that forecasters adhere to the five evidence-based checklists provided in this paper. Checklists have successfully been used to obtain compliance in other domains. Clients, and other forecast users and commentators, should nevertheless assess compliance with the checklists.

Experimental research over the past century has led to the identification of 17 evidence-based forecasting methods. We described the methods, along with their estimated effects on *ex ante* forecast accuracy. The accuracy of forecasts can be substantially improved by using them instead of the methods that are commonly used.

All methods in the Forecasting Methods Application Checklist (Exhibit 1) that are suitable for a forecasting problem should be used when accuracy is important. Combining forecasts from diverse combined methods is likely to lead to large gains in forecast accuracy.

Forecasting can be abused to provide support for a preferred course of action at the expense of accuracy. Advocacy forecasting can be avoided, or detected, by using the 28 guidelines in the Golden Rule of Forecasting Checklist (Exhibit 3). On average across 70 relevant experimental comparisons, the violation of a single checklist item increased error by 39%.

Occam's razor—or simplicity—is a principle of science that improves accuracy when it is applied to forecasting methods and models. Methods than cannot be understood by potential clients provide forecasts that are less accurate than those from methods that can be understood. The simple forecasting checklist provided in this paper consists of four items. On average, methods that were more complex had *ex ante* forecast errors that were 27% larger than those from simple methods.

No forecasting method should be used without experimental evidence of predictive validity under specified and reproducible conditions. The "big data analytics" movement is a case in point. The method ignores findings from experimental research and key forecasting principles, and makes no use of prior knowledge relevant to the situation. Eight commonly used methods that should not be used for forecasting are listed in the last of the six checklists, Exhibit 6.

The good news is that enormous gains in accuracy are available by using relatively simple evidence-based methods and principles. The gains are especially large when long-term forecasts are involved. To obtain these benefits clients must ask analysts to follow the scientific checklists and then must audit their work to ensure that they complied with the checklists. We are not aware of any other way to benefit from the research on forecasting.

REFERENCES

Key

NS = not cited regarding substantive finding
AO = this paper's authors' own paper
NF = unable to find email address (including deceased)
NR = contact attempted (email sent) but no substantive reply received
FD = disagreement over interpretation of findings remains
FC = interpretation of findings confirmed in this or in a related paper

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

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting* 28, 689-694.AO

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

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

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

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

Armstrong, J. S. (2001b). Standards and practices for forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 679-732). Norwell, MA: Kluwer Academic Publishers. AO

Armstrong, J. S. (2001c). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 417-439). Norwell, MA: Kluwer Academic Publishers. AO

Armstrong, J. S. (1985). *Long-Range Forecasting*. New York: John Wiley and Sons. AO

Armstrong, J. S. (1980). The seer-sucker theory: The value of experts in forecasting. *Technology Review*, 83 (June/July 1980), 18-24. AO

Armstrong, J. S. (1978). Forecasting with econometric methods: Folklore versus fact. *The Journal of Business* 51, 549-64. AO

Armstrong, J. S. (1970a). How to avoid exploratory research. *Journal of Advertising Research,* 10, No. 4, 27-30. AO

Armstrong, J. S. (1970b). An application of econometric models to international marketing. *Journal of Marketing Research*, 7, 190-198. AO

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. AO

Armstrong, J. S., & Andress, J. G. (1970). Exploratory analysis of marketing data: Trees vs. regression. *Journal of Marketing Research*, 7, 487-492. AO

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

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

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

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

Armstrong, J. S., & Cuzán, F. (2006), Index methods for forecasting: An application to American presidential elections. *Foresight: The International Journal of Applied Forecasting,* 3, 10-13. AO

Armstrong, J. S., Denniston, W. B., & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance,* 14, 257-263. AO

Armstrong, J.S., Du, R., Green, K.C. & Graefe, A. (2016**),** Predictive validity of evidence-based persuasion principles, *European Journal of Marketing*, 50, 276-293. AO

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. AO

Armstrong, J.S., Green, K. C. (2017). Guidelines for science: Evidence and checklists. Working paper, ResearchGate. AO

Armstrong, J.S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, 68, 1717–1731. AO

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. AO

Armstrong, J. S., & Overton, T. S. (1977), Estimating nonresponse bias in mail surveys. *Journal of Marketing Research* 14, 396-402. AO

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

Armstrong, J.S. & Pagell, R. (2003), Reaping benefits from management research: Lessons from the forecasting principles project**,** *Interfaces*, 33, 89-111. ^{AO}

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

Batchelor, R. A., & Kwan, T. Y. (2007). Judgmental bootstrapping of technical traders in the bond market. *International Journal of Forecasting*, 23, 427-445. ^{FC}

Belvedere, V. and Goodwin, P. (2017). The influence of product involvement and emotion on short-term product demand forecasting. *International Journal of Forecasting,* 33, 652-661. ^{FC}

Boylan, J. E., Goodwin. P, Mohammadipour, M., & Syntetos, A.A. (2015). Reproducibility in forecasting research. *International Journal of Forecasting* 3, 79-90.

Brown, R. G. (1959), *Statistical Forecasting for Inventory Control*, New York: McGraw-Hill. ^{NF}

Brown, R.G. (1962). *Smoothing, Forecasting and Prediction of Discrete Time Series*. London: Prentice-Hall.^{NF}

Buchanan, W. (1986). Election predictions: An empirical assessment. *The Public Opinion Quarterly*, 50, 222-227. ^{NF}

Bunn, D.W. & Vassilopoulos, A.I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting*, 15, 431–443. ^{FC}

Burgess, E. W. (1936). Protecting the public by parole and by parole prediction, *Journal of Criminal Law and Criminology*, 27, pp. 491–502. ^{NF}

Cerf, C. & Navasky, V. (1998). *The Experts Speak*. New York: Villard. ^{NS}

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

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. ^{AO}

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. ^{AO}

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, 1(26), 1-9. doi: 10.3389/fpsyg.2010.00026 ^{FC}

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.

Cuzán, A. & Bundrick, C. M. (2009). Predicting presidential elections with equally weighted regressors in Fair's equation and the fiscal model. *Political Analysis*, 17, 333-340.

Dana, J. & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29 (3), 317-331. ^{FC}

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologis*t, 26, 180-188. ^{NF}

Dillman, D. A., Smyth J. D., & Christian, L. M. (2014). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. (4th ed.). Hoboken, NJ: John Wiley. ^{NS}

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37, 570-576. ^{FC}

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. ^{FC}

Fildes, R. & R. Hastings (1994), The organization and improvement of market forecasting, *Journal of the Operational Research Society*, 45, 1-16. FC

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

Franses, P. H. (2014). *Expert adjustments of model forecasts: theory, practice and strategies for improvement*. Cambridge, U.K.: Cambridge University Press.

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

Gigerenzer, G. Todd, P.M.& the ABC Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press. FC

Goodwin, P., Dyussekeneva, K. and Meeran, S. (2013). The use of analogies in forecasting the annual sales of new electronics products. *IMA Journal of Management Mathematics*, 24, 407-422. FC

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

Gough, H. G. (1962). Clinical versus statistical prediction in psychology," in L. Postman (ed.), *Psychology in the Making*. New York: Knopf, pp. 526-584.

Graefe, A. (2017a). Political markets. In Arzheimer, K., and Lewis-Beck, M. S. (eds.). *The Sage Handbook of Electoral Behavior*. Los Angeles??: Sage

Graefe, A. (2017b). Prediction market performance in the 2016 U. S. presidential election, *Foresight*, – *The International Journal of Applied Forecasting*, 45, 38-42. FC

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, 68, 1792–1799. FC

Graefe, A. (2014). Accuracy of vote expectation surveys in forecasting elections. *Public Opinion Quarterly*, 78 (S1): 204–232. FC

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

Graefe, A., & Armstrong, J.S. (2013). Forecasting elections from voters' perceptions of candidates' ability to handle issues. *Journal of Behavioral Decision Making*, 26, 295-303. DOI: 10.1002/bdm.1764. AO

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. AO

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. AO

Graefe, A., Armstrong, J. S., & Green, K. C. (2014). Improving causal models for election forecasting: Further evidence on the Golden Rule of Forecasting. *APSA 2014 Annual Meeting Paper*. Available at SSRN: https://ssrn.com/abstract=2451666

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2017). Assessing the 2016 U.S. Presidential Election Popular Vote Forecasts," in *The 2016 Presidential Election: The causes and consequences of an Electoral Earthquake*. Lexington Books, Lanham, MD. AO

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2014). Combining forecasts: An application to political elections. *International Journal of Forecasting*, 30, 43-54. AO

Graefe, A., Küchenhoff, H., Stierle, V. & Riedl, B. (2015). Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting*, 31(3), 943-951. FC

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. AO

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

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research* 68 (8), 1678-1685. AO

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. AO

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

Green, K. C., Armstrong, J.S., Du, R. & Graefe, R. (2015). Persuasion principles index: Ready for pretesting advertisements, *European Journal of Marketing*, 50, 317–326. AO

Green, K. C., Armstrong, J. S., & Graefe, A. (2015). Golden rule of forecasting rearticulated: forecast unto others as you would have them forecast unto you. *Journal of Business Research*, 68, 1768-1771. AO

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

Gregory, W. L., & Duran, A. (2001), Scenarios and acceptance of forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting* (pp. 519–540). Norwell, MA: Kluwer Academic Publishers. FC

Grove, W. M. et. al. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 19-30.

Hales, B. M., & Pronovost, P. J. (2006). The checklist—a tool for error management and performance improvement. *Journal of Critical Care*, *21*, 231-235. NS

Hayes, S. P. Jr. (1936).The inter-relations of political attitudes: IV. Political attitudes and party regularity. *The Journal of Social Psychology*, *10*, 503-552. NF

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.

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

Hubbard, R. (2016). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. New York: Sage.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2015). The power of bias in economics research. Deakin University Economics Series Working Paper SWP 2016/1. Available from https://www.deakin.edu.au/__data/assets/pdf_file/0007/477763/2016_1.pdf

Jones, R. J., Armstrong, J. S., & Cuzán, A. G. (2007). Forecasting elections using expert surveys: An application to U. S. presidential elections. AO Working paper available at http://repository.upenn.edu/marketing_papers/137

Kabat, G. C. *Hyping Health Risks* (2008). N.Y., N.Y.: Columbia University Press. FC

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371.

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. FC

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. FC

Lichtman, A. J. (2005). The keys to the White House: Forecast for 2008. *Foresight: The International Journal of Applied Forecasting*, 3, 5-9.

Lin, S., Goodwin, P. and Song, H. (2014). Accuracy and bias of experts' adjusted forecasts. *Annals of Tourism Research,* 48, 156-174. FC

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

Lott, J. R., Jr. (2016). *The War on Guns*. Regnery Publishing: Washington, D.C.

Lott, J. R., Jr. (2010). *More Guns, Less Crime*. Third Edition. University of Chicago Press.

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. FC

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.

Malkiel, B. G. (2016). *A random walk down Wall Street: The time-tested strategy for successful investing*. New York, NY: Norton. NS

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology,* 107, 276-299. FC

Mazar, N., Amir, O, & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept management. *Journal of Marketing Research,* 45, 633-644. FC

McNees, S. K. (1992), The uses and abuses of 'consensus' forecasts. *Journal of Forecasting,* 11, 703-710. NF

McShane, B. B. & Gal, D. (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science, 62,* 1707-1718. FC

Meehl, P.E. (1954). *Clinical vs. Statistical Prediction*, Minneapolis: University of Minnesota Press. NF

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. FC

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

Morganstern, O. (1963). *On the Accuracy of Economic Observations*, Princeton: Princeton University Press,1963.NS

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. FC

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

Murphy, A.H. & Winkler, R.L. (1984). Probability forecasting in meteorology, *Journal of the American Statistical Association*. 79, 489-500.

Namboodiri, N.K. & Lalu, N.M. (1971). The average of several simple regression estimates as an alternative to the multiple regression estimate in postcensal and intercensal population estimates: A case study, *Rural Sociology*, 36, 187-194.

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukos, V. & Khammash, M. (2015), Relative performance of methods for forecasting special events, *Journal of Business Research*, 68, 1785-1791. FC

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, 84, 231-259. FC

Pant, P. N. & Starbuck, W. H. (1990). Innocents in the forest: Forecasting and research methods. *Journal of Management*, 16, 433-446.

Perry, M. J. (2017). 18 spectacularly wrong predictions made around the time of first Earth Day in 1970, expect more this year. *AEIdeas*, April 20. FC
Available from http://www.aei.org/publication/18-spectacularly-wrong-predictions-made-around-the-time-of-first-earth-day-in-1970-expect-more-this-year/

Rhode, P.W. & Stunpf, K.S. (2004), Historical presidential betting markets. *Journal of Economic Perspectives*, 18, 127-141.

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. FC

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31, 699-714.

Shamir, J. (1986). Pre-election polls in Israel: Structural constraints on accuracy, *Public Opinion Quarterly*, 50, 62-75.

Sheeran, P. (2002). Intention-behavior relations: A conceptual and empirical review. in W. Stroebe and M. Hewstone, *European Review of Social Psychology*, 12, 1-36. FC

Simon, J. L. (1996). *The ultimate resource II: people, materials, environment*. Princeton, NJ: Princeton University Press. NF Available from http://www.juliansimon.com/writings/Ultimate_Resource/

Soyer, E., & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28, 695-711.

Tessier, T.H. & Armstrong, J.S. (2015). Decomposition of time-series by level and change. *Journal of Business Research*, 68, 1755–1758. AO

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

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, 29, 234-243. FC

Tukey, J.W. (1962), The future of data analysis. *Annals of Mathematical Statistics*,33,1-67.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83, 213-217.

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.

Wold, H. & Jureen, L. (1953). *Demand Analysis*. New York: John Wiley.NS

Wright, M. & Armstrong, J.S. (2008), Verification of citations: Fawlty towers of knowledge. *Interfaces*, 38, 125-139. AO

Yokum, J.T. & Armstrong, J.S (1995) Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, 11, 591-597. AO

Ziliak, S. T. & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: The University of Michigan Press.

Ziliak, S. T., & McCloskey D. N. (2004). Size matters: The standard error of regressions in the *American Economic Review*. *The Journal of Socio-Economics, 33,* 527–546. FC

Total Words 18,400
Text only 14,600