Forecasting Methods and Principles: Evidence-Based Checklists

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

Forecasting Methods 171-KCG~~0 clean~~ (November 6, 2017)

Please suggest evidence we have overlooked and tell us about mistakes we have made.

ABSTRACT

Problem: Few practitioners or academics use findings from nearly a century of experimental research that would allow them to substantially reduce forecast errors. In order to improve forecasting practice, this paper develops evidence-based guidance in a form that is easy for forecasters and decision-makers to access, understand, and use: checklists.

Methods: Meta-analyses of experimental research on forecasting were used to identify the principles and methods that lead to accurate out-of-sample forecasts. Cited authors were contacted to check that summaries of their research were correct. Checklists to help forecasters and their clients practice evidence-based forecasting were then developed from the research findings. Finally, appeals to identify errors of omission or commission in the analyses and summaries of research findings were sent to leading researchers.

Findings: Seventeen simple forecasting methods can between them be used to provide accurate forecasts for diverse problems. Knowledge on forecasting is summarized in the form of five checklists with guidance on the selection of the most suitable methods for the problem, and their implementation.

Originality: Three of the five checklists—addressing (1) evidence-based methods, (2) regression analysis, and (3) assessing uncertainty—are new. A fourth—the Golden Rule checklist—has been improved. The fifth—the Simple Forecasting checklist (Occam's Razor)—remains the same.

Usefulness: Forecasters can use the checklists as tools to reduce forecast errors—often by more than one-half—compared to those of forecasts from commonly used methods. Scientists can use the checklists to devise valid tests of the predictive validity of their hypotheses. Finally, clients and other interested parties can use the checklists to determine whether forecasts were derived using evidence-based procedures and can, therefore, be trusted.

***Key words*:** combining forecasts, decision-making, decomposition, equal weights, equalizing coefficients, expectations, extrapolation, knowledge models, intentions, Occam's razor, prediction intervals, predictive validity, regression analysis, 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, but suggest reading more slowly.
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, Magne Jorgensen, Kostas Nikolopoulos, Don Peters, and Malcolm Wright for their reviews. We also thank those who made useful suggestions, including Raymond Hubbard, Frank Schmidt, Phil Stern, and Firoozeh Zarkesh. 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: Harrison Beard, Amy Dai, Simone Liao, Brian Moore, 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

Forecasts are important for decision-making in businesses, governments, and other organizations. Researchers since the 1930s have responded to the need for forecasts by conducting experiments testing multiple reasonable methods. The findings from those experiments have led to great improvements in knowledge about forecasting.

In the mid-1990s, 39 leading forecasting researchers and 123 expert reviewers were involved in identifying and collating scientific knowledge on forecasting. The findings were summarized in the form of 139 principles (condition-action statements), in Armstrong (2001b). 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).

While the advances in forecasting knowledge have provided the opportunity for substantial improvements in forecast accuracy, most practitioners and academics do not make use of that knowledge. Possible reasons are that they: (1) prefer to stick with their current forecasting procedures; (2) wish to provide support for a belief or preferred decision; (3) are unaware of evidence-based methods; (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 use evidence-based procedures. In regard to the third reason, 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 (Cox and Loomis, 2001). Practitioners who are not using evidence-based forecasting methods for reason numbers 3, 4, or 5 will benefit from reading this paper.

This paper develops guidelines for forecasting that draw heavily on the evidence-based principles mentioned above, and on more recent research. To help forecasters and decision-makers, the guidelines are provided as checklists. The guidelines are intended primarily for the purpose of improving the accuracy of out-of-sample forecasts for diverse situations. Accuracy is the most important criterion for most parties concerned with forecasts (Fildes and Goodwin 2007). We also discuss other important criteria, such as uncertainty, cost, and understandability of the methods (Yokum and Armstrong 1995).

CHECKLISTS TO IMPLEMENT AND ASSESS FORECASTING METHODS

Evidence-based checklists completed and verified by the person responsible for decisions avoid the need for memorizing, make complex tasks easier, and provide relevant guidance on a timely basis. In fields such as medicine, aeronautics, and engineering, a failure to follow an appropriate checklist can be grounds for a lawsuit. Much research supports the value of using checklists (e.g., Hales and Pronovost 2006). One experiment assessed the effects of using a 19-item checklist for a hospital procedure. The study compared the outcomes experienced by thousands of patients in hospitals in eight cities around the world before and after the checklist was used. Use of the checklist led to a reduction in deaths from 1.5% to 0.8% in the month after the operations (Haynes et al. 2009).

When we commissioned people to complete tasks that required them to use a checklist, the vast majority of those who accepted the task did so effectively. For example, to assess the persuasiveness of print advertisements, raters hired through Amazon’s Mechanical Turk used 195 checklist items to code advertisements on their conformance to persuasion principles. Their ratings had high inter-rater reliability (Armstrong, Du, Green, and Graefe 2016).

We present checklists of the forecasting guidelines. The checklists are intended for use by forecasters, and by all who have stake in accurate forecasts and predictive validity.

RESEARCH METHODS

We reviewed research findings to develop forecasting guidelines. To do so, we first identified relevant research by:

1) searching the Internet, mostly using Google Scholar;
2) contacting leading researchers for suggestions on key studies;
3) checking papers referred to in key studies;
4) putting our working paper online with requests for evidence that we might have overlooked.

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

We developed our guidelines using findings from papers that conformed to the *Checklist of* C*riteria for Useful Scientific Research* at GuidelinesforScience.com. In particular, we discarded papers that did not test multiple reasonable hypotheses or that relied on non-experimental data or that did not test out-of-sample forecast accuracy.

Formatted: Default Paragraph Font, Not Expanded by / Condensed by

To ensure that we properly summarized findings from prior research, we attempted to contact the authors of all papers that we cited regarding substantive findings. We did so on the basis of evidence that a high percentage of findings cited in papers in leading scientific journals are described incorrectly (Wright and Armstrong 2008); largely because researchers seldom read the papers that they cite (Simkin and Roychowdhury 2005). We asked the authors we contacted to suggest relevant papers that we had overlooked—especially papers describing experiments with findings that conflicted with ours. Many of the authors helped. The practice of contacting leading researchers was shown to produce reviews that are substantially more comprehensive than those done with computer searches (Armstrong and Pagell 2003). We have coded efforts to contact authors, and the results, in the references section of this paper.

Finally, we developed checklists of our evidence-based guidelines in order to make forecasting knowledge accessible to all. Draft versions of the checklists were modified as our review of research findings progressed.

VALID FORECASTING METHODS: DESCRIPTIONS AND EVIDENCE

The predictive validity of a forecasting method is assessed by comparing the accuracy of forecasts from the method with forecasts from the currently used method, or from other evidence-based methods. That is the scientific method of testing multiple reasonable hypotheses (Chamberlin 1890).

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 differences between *ex ante* forecasts and data on what actually transpired. The benchmark error measure for evaluating forecasting methods is the easily understood and decision-relevant Relative Absolute Error, abbreviated as "RAE" (Armstrong and Collopy 1992).

Tests of a new method—a development of the RAE—called the Unscaled Mean Bounded Relative Absolute Error (UMBRAE)—suggest that it is superior to the RAE and other alternatives that have been proposed (Chen, Twycross, and Garibaldi 2017). Given that the evidence on UMBRAE is based on only this one study, however, we suggest using both the RAE and UMBRAE until such time as the evidence on its usefulness allows a definitive conclusion on which is the better measure.

Exhibit 1 lists 17 valid forecasting methods: methods that are consistent with forecasting principles and have been shown to provide out-of-sample forecasts with superior accuracy. For each, the Exhibit identifies the knowledge needed to use the method. For most forecasting problems, several of the methods will be usable. An electronic version of the Exhibit 1 checklist will be provided at ForecastingPrinciples.com in the top menu bar. This paper provides a description of each method and a brief review of the evidence.

Formatted: Default Paragraph Font, Not Expanded by / Condensed by

Exhibit 1: Forecasting Methods Application Checklist

Name of forecasting problem: ____

Forecaster: ____ Date: _____________________

Method	**Knowledge needed**		**Usable method**	**Variations of components**
	Forecaster*	**Respondents/Experts**	(☒)	(Number)
Judgmental methods				
1. Prediction markets	Survey/market design	Domain; Problem	□	[ ]
2. Multiplicative decomposition	Domain; Structural relationships	Domain	□	[ ]
3. Intentions surveys	Survey design	Own plans/behavior	□	[ ]
4. Expectations surveys	Survey design	Others' behavior	□	[ ]
5. Expert surveys (Delphi, *etc.*)	Survey design	Domain	□	[ ]
6. Simulated interaction	Survey/experiment design		□	[ ]
7. Structured analogies	Survey design	Analogous events	□	[ ]
8. Experimentation	Experimental design	Normal human responses	□	[ ]
9. Expert systems	Survey design	Domain	□	[ ]
Quantitative methods *(Judgmental inputs typically required)*				
10. Extrapolation	Time series methods; Data	n/a	□	[ ]
11. Rule-based forecasting	Causality; Time series methods	Domain	□	[ ]
12. Regression analysis	Causality; Data	Domain	□	[ ]
13. Judgmental bootstrapping	Survey/experiment design	Domain; Causality	□	[ ]
14. Segmentation	Causality; Data	Domain	□	[ ]
15. Knowledge models	Cumulative knowledge	Domain	□	[ ]
16. Combining forecasts from a single method… □		**SUM of VARIATIONS**		**[ ]**
17. Combining forecasts from several 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; October 15, 2017

Because we are concerned with methods that have been shown to improve forecast accuracy relative to methods that are commonly used in practice, we do not discuss all methods that have been used for forecasting. Forecast users should ask forecasters what methods they will use, and the reasons why. If they do not provide good reasons, find another forecaster. If they mention a method that is not listed in Exhibit 1, ask them to produce evidence that their method provides forecasts smaller errors than the relevant methods listed in the Exhibit 1 checklist.

We start our descriptions of evidence-based forecasting methods with judgmental methods, and follow with descriptions of quantitative methods. The latter also require judgment.

Judgmental Methods

Expertise based on experience in similar situations can be useful for forecasting, by the use of relative frequencies, for example. Experience can also lead to the development of simple "rules of thumb," or heuristics, that provide quick forecasts that are usually sufficiently accurate for making good decisions, such as choosing between options. A dramatic demonstration was provided the emergency landing of US Airways Flight 1549—the "Miracle on the Hudson." The landing was a success because the pilot used the gaze heuristic to forecast that landing on the Hudson was the only viable option (Hafenbrädl, Waeger, Marewski, and Gigerenzer 2016). The superiority of simple heuristics for many recurrent practical problems has been shown by extensive research conducted by Gerd Gigerenzer and the ABC group of the Max Planck Institute for Human Development in Berlin. Goodwin (2017) also describes situations where expertise, translated into rules-of-thumb helps to make accurate forecasts.

Importantly, however, expertise and experience in a field or specific problem area is, *on its own*, of no apparent value for making accurate forecasts in complex situations with poorly understood or uncertain cause and effect relationships, where experts and managers do not receive frequent well-summarized feedback on the accuracy of their predictions, and where there are three or more important causal factors. Such situations are common in business and government decision making.

Research on the accuracy of experts' unaided judgmental forecasts about complex and nonrecurring situations dates from the early 1900s. An early review of the research 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 study comparing the accuracy of many forecasts from experts with that of forecasts from novices and from naïve rules provided support (Tetlock 2005). Consider also that many people invest in hedge funds despite the evidence that the returns from the expert stock pickers' portfolios are inferior to those from a portfolio that mimics the stock market (Malkiel 2016).

As a general rule, unaided expert judgment should be avoided for complex nonrecurring situations for which simple heuristics have not been shown to be valid. For such situations, structured judgmental methods are needed. This section describes nine evidence-based structured methods for forecasting using judgment.

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 2004). They attract experts who are motivated to use their knowledge to win money by making accurate predictions. Because market participants are anonymous, there is no penalty for bets that are inconsistent with the experts' personal beliefs or public statements, and so their bets are more likely to be unbiased ~~forecasts~~.

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 prediction markets, as well as with survey design.

The accuracy of forecasts from prediction markets was tested in eight published comparisons in the field of business forecasting. Out-of-sample forecast errors were 28% lower than those from no-change models (Graefe 2011). In another test, forecasts from the Iowa Electronic Market (IEM) prediction market across the 100 days before each U.S. presidential election from 2004 though 2016 were, on average, less accurate than forecasts from the RealClearPolitics poll average, a survey of experts, and citizen forecasts (Graefe 2017a). This prediction market limits the bets to no more than $500, which is likely to reduce the number and motivation of participants. However, comparative accuracy tests based on 44 elections in eight countries other than the U.S. found that forecasts from betting markets were more accurate than forecasts by experts, econometric models, or polls (Graefe 2017b).

Multiplicative decomposition (2)

Multiplicative decomposition involves dividing a forecasting problem into parts, the forecasts for which are multiplied together to forecast the whole. 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 different parts, when relevant data can be obtained for some parts of the problem, and when the directional effects of the causal factors differ among the components.

Those conditions seem to be common, and the decomposition principle has long been a key guideline for decision-making (e.g., a Google search for “management decision making” and “decomposition” found almost 100,000 results in October 2017). To assess the size of the effect of using decomposition for forecasting, an experiment was conducted to compare the accuracy of global estimates with the combined estimates of elements of the decomposed whole. 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 reduce accuracy for any of the five problems (Armstrong, Denniston, and Gordon 1975). MacGregor (2001) summarized three studies (including the above study) and found that judgmental decomposition led to a 42% reduction in error.

Intentions surveys (3)

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 vital when historical data are not available. They are most likely to provide useful forecasts when the forecast time-horizon is short, the behavior is familiar, and the behavior is sufficiently important that respondents tend to plan it (Morwitz 2001; Morwitz, Steckel, and Gupta 2007).

To assess ~~the~~ intentions of others, prepare accurate but brief descriptions of the situation and present them without bias (Armstrong and Overton 1971). Intentions are most usefully 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)’ so that the survey responses can be used to calculate the proportion of the population that is likely to behave in a given way (Morwitz 2001).

The way that a question is asked can have an enormous effect on responses. Even when researchers strive for objectivity, there are often differences in responses due to seemingly minor changes in wording. Over the past century, much research has been done on this issue. Here are two recommendations for reducing response error: (1) pretest the questions to ensure that the respondents understand them in the way the forecaster intends; (2) use alternative ways to state a question and then average responses across questions. For more advice, see Bradburn, Sudman, and Wansink (2004).

To reduce non-response error, provide monetary inducements when you send out the questionnaire (Armstrong and Yokum 1994). In addition, resend the survey to non-responders in follow-up waves after allowing ample time for the respondent to reply. Doing so allows the forecaster to estimate the effect of non-response by extrapolating across waves (Armstrong and Overton 1977). Additional evidence-based procedures for selecting samples and obtaining high response rates, are described in Dillman, Smyth, and Christian (2014).

Expectations surveys (4)

Expectations surveys ask people how they *expect* they or others will behave. Expectations differ from intentions because people know that things can change. 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 need expensive repairs, so you might expect a 10% 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 (Haynes, Weiser, Berry, et al. 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. In addition, 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 percentage points (Graefe, Armstrong, Jones, and Cuzán, 2017).

Expert surveys (5)

Use written questions and instructions for interviewers or self-completion surveys to ensure that each expert is questioned in the same way. Apply the same procedures for developing questions as those described for expectations surveys, above. Additional advice on the design of expert surveys is provided in Armstrong (1985, pp.108-116).

Obtain forecasts from at least five experts; up to 20 for important forecasts (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 percentage points for the average error of combined polls (Graefe, Armstrong, Jones, and Cuzán 2017, and personal correspondence with Graefe).

Delphi is an extension of the expert survey approach whereby the survey is given in two or more rounds with *anonymous* summaries of the experts' 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. Depending on whether the forecasts is scaler or categorical, use the median or the 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 anonymous and easy to understand. Moreover, 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 the different experts each have different information relevant to the problem or different interpretations of the nature of the situation (Jones, Armstrong, and Cuzán 2007).

Forecasts made with Delphi were more accurate than forecasts made in traditional meetings in five studies, 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 (6)

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, or how government would respond to artists demanding that the government buy paintings that they were unable to sell.

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

Relative to unaided expert judgment—the method usually used for such situations—simulated interaction reduced forecast errors by 57% on average 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 between two countries over access to water.

The alternative approach of "putting oneself in the other person's shoes" has been proposed. 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, however, 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).

Structured analogies (7)

The structured analogies method involves asking ten or so experts to suggest situations that were similar to the one for which a forecast is required. The experts are given a description of the situation and are asked to identify analogous situations, rate their similarity to the target situation, and 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 among the independent experts as the forecast (Green and Armstrong 2007). The method should not be confused with the common use of analogies to justify a decision 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).

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

A procedure akin to structured analogies was used to forecast box office revenue for 19 unreleased movies, in which raters identified analogous movies from a database and rated them for similarity. The revenue forecasts from the analogies were adjusted for advertising expenditure and whether 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).

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, Petropoulos, Bougioukos, and Khammash 2015).

Across the ten comparative tests from the three studies described above, the error reductions from using structured analogies averaged about 40%.

Experimentation (8)

Experimentation is widely used and is the most valid and reliable method to determine the effects that changes in a causal variable will have. Estimates of effects can then be used to make forecasts. Experiments can be conducted in a laboratory or other artificial environment, or in the field. An analysis of experiments in the field of organizational behavior found that laboratory and field experiments yielded similar findings (Locke 1986).

Alternatively, forecasters can analyze natural experiments to identify causal relationships and use that information to make forecasts. For example, regulation and deregulation of industries provide natural experiments on the effect of regulation on consumer welfare. Winston (1993) found that regulation harmed customers in eight of the nine markets for which experimental data were available, and was of no net benefit in the case of the ninth.

Expert systems (9)

Expert systems are developed by asking experts to describe their step-by-step process for making forecasts. The process must be explicitly defined and unambiguous, such that it could be implemented using software. The resulting expert system should be simple, clear, and complete.

Expert system forecasts were more accurate than forecasts from unaided judgments based 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 forecast errors were 10% and 5% smaller than those from unaided judgment. While little validation evidence is available, the method appears promising.

Quantitative Methods

Quantitative methods require at least some, and typically much, numerical data on or related to what is being forecast. Quantitative methods can also draw upon judgmental methods, such as decomposition, in order to structure the forecasting problem in ways that make the best use of knowledge and data. This section describes six evidence-based quantitative forecasting methods.

Extrapolation (10)

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 forecast variable, 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 and 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).

It is, however, risky to assume that a trend will continue at the same rate, even in the short term. It could increase or decrease. Which is more likely? That depends on the causal forces that drive the trend. In many situations, you do not know. The greater the uncertainty about the situation, the greater is the need for damping the trend toward zero—the no change forecast. A review of ten experimental comparisons found that, on average, damping the trend toward zero reduced forecast errors by almost 5% , as compared to using a constant trend (Armstrong 2006). In addition, damping reduced the risk of large errors. (Gardner's software for damped-trend extrapolation can be found at ForcastingPrinciples.com.) When there is a consistent long-term trend and the causal factors are expected to continue—as with the prices of resources (Simon 1996)—damping toward the long-term trend is appropriate.

Formatted: Default Paragraph Font, Not Expanded by / Condensed by

When extrapolating data more frequent than annual, 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, Andersen, Carbone, 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 for the 1,428 monthly time series from the M3-Competition prior to forecasting, the accuracy of the forecasts improved for between 59% to 65% of the series, depending on the horizon. These findings were successfully replicated with respect to the direction and broad magnitude of improvement (Boylan, Goodwin, Mohammadipour, et al. 2015). Software for the Miller-Williams procedures is freely available at ForcastingPrinciples.com.

Formatted: Default Paragraph Font, Not Expanded by / Condensed by

Damping by averaging seasonal factors across analogous series also improves forecast accuracy. 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—which can be characterized as growth, decay, opposing, regressing, supporting, or unknown—that affect trends in different ways. In such a case, one can 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, one study forecasted 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. The two forecast series were then multiplied to get total deaths. When tested on five time series that clearly met the two conditions, decomposition by causal forces reduced out-of-sample forecast errors by two-thirds. For the four series that partially met the conditions, decomposition by causal forces reduced error by one-half. There was no gain, or loss, in forecast accuracy when the conditions did not apply (Armstrong, Collopy and Yokum 2005).

Additive decomposition can also be considered for extrapolation problems. One useful approach is 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).

Rule-based forecasting (11)

Rule-based forecasting (RBF) allows the use of causal knowledge for time-series extrapolation. 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 forecast horizon, the amount of data available, and the existence of outliers (Armstrong, Adya and Collopy 2001). Features are identified by inspection, statistical analysis, and domain knowledge. Alternative extrapolation models are used, and 99 rules are used for combining the forecasts from the models.

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 unweighted combinations of forecasts (Collopy and Armstrong 1992). Testing by Vokurka, Flores, and Pearce~~. et al.~~ (~~1999~~1996) provided supporting evidence.

The contrary series rule is especially important—and simple. It states that if the expected direction of a time series and the recent trend of a time-series are contrary to one another, one should not forecast a trend. 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 coding for "expected direction" of the trend was done by the authors of that paper, who were not experts on the product categories.

Regression analysis (12)

As other sections of this paper describe, regression analysis is an important tool for quantifying relationships. In this section, we discuss the use—and misuse— of regression analysis to develop data models for forecasting. We use the term "data models" to refer to models that use data to estimate causal relationships.

Regression analysis can be useful for estimating the strength of relationships between the variable to be forecast and one or more *known* causal, or predictor, variables. Such estimates can be useful in predicting the effects of policy changes or of changes in the environment. One of the benefits of regression analysis is that it is conservative, in that it reduces the size of coefficient estimates to adjust for random measurement error in the variables. Much literature published over the years, however, has concluded that the validity of regression estimated model parameters has been poor. For example, Ioannidis, Stanley, and Doucouliagos (~~2015~~2017) 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.

Regression model estimation is harmed by the omission of important causal variables, inclusion of irrelevant variables, interactions among causal variables 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, when a model that was properly specified on the basis of *a priori* analysis is then changed in order to improve statistical fit with historical data, predictive validity is reduced, and the risk of reports of humorous—or perhaps dangerously wrong—findings in the media is increased.

Exhibit 2 provides a checklist of guidelines for using regression analysis for forecasting. The reasoning and evidence for the guidelines are 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.	Develop alternative multiple and one-variable regression models and combine their forecasts	□
Validation of the forecasting model		
11.	Validate models 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—Regression analysis is unsuited for identifying relationships by application to non-experimental data. Instead, follow the scientific method by using prior knowledge. This involves using obvious unchallenged relationships from domain knowledge or logic from known relationships. In cases where the relationships are not obvious, one needs to summarize experimental evidence and domain knowledge to identify causal variables.

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

Unfortunately, the improper use of regression analysis to detect relationships in non-experimental data is increasingly common. 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).

The problems with non-experimental data were shown in a study of 56 persuasion principles from Armstrong (2010). Evidence on the direction of the effects of each principle was available from regression analyses of non-experimental data, and from experimental data. The findings from the experimental analyses were in in the same direction for each of these principles. However, for non-experimental data, agreement was found for only two-thirds of the principles (Armstrong and Patnik 2009).

2. Specify the direction and importance of causal variables' effects on the variable to be forecast—List all the variables considered and rank them according to their expected (*a priori*) importance when forecasting the dependent variable. The directional effects of some variables are obvious from logic or common knowledge about the domain. If the direction is not obvious, refer to the research literature. Failing that, survey three to five domain experts. Record your *a priori* estimates of the causal variable coefficients.

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 means that one would expect a price increase of 10% to result in a 12% decrease in the quantity demanded, all else being equal. 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 is based on logic. That is, if you cannot accurately forecast or control a causal variable, how could inclusion of it in a model improve the accuracy of forecasts of the dependent variable?

4. Specify a model with a form and content that embodies prior knowledge—This guideline, too, should be obvious if one's purpose is to obtain accurate *ex ante* forecasts.

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 prefer the simplest model that is consistent with the evidence. The effect of model simplicity on forecast accuracy is substantial, as we show later in this paper. When you are satisfied that you have specified a simple model that embodies prior knowledge (regression guidelines 4 and 5), apply your *a priori* estimates of the coefficients (guideline 2) to data on the variables in the model—transformed as appropriate—and then use regression analysis to estimate the constant term for this *a priori* model.

6. Obtain all relevant data on the model variables—For example, when using time-series include data for all time periods unless a convincing case can be made—prior to conducting regression analysis—for not doing so. Record why you excluded data from your analysis if you did so.

7. Ensure data are valid, reliable, and 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; that is, that they truly represent the variable of interest. For example, does the Gross Domestic Product provide a good measure of "economic well-being" if it excludes production of goods and services where no money changes hands? By reliable, we mean that the method that is used to measure the variable will repeatedly provide similar result. Problems with validity and reliability are common, as Morgenstern (1963) showed.

One way to improve validity and reliability is to identify 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. There are many situations where more than one data series can be used to measure the same conceptual variable.

Outliers in data should be adjusted or removed in order to avoid their excessive influence on the least-squares regression estimates (coefficients). One solution is to reduce the outlier to the value of the most extreme observation in which you have confidence, a procedure called "winsorizing" (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 variable to be forecast. 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 Einhorn and Hogarth (1975) and Dana and Dawes (2004) who showed some of the conditions under which regression is and is not effective, relative to equal weights.

The equal weights approach was tested in election forecasting using a regression model that included all of the 27 variables that had been used across ten independent econometric (regression) models. The equal weights model's *ex ante* average forecast error was 29% lower than the average error of the most accurate of the ten original regression models (Graefe 2015). Note that econometric model coefficients in election forecasting are typically based on small samples of data, often fewer than 15 observations.

While the findings have shown that the equal weights approach is often the best approach in out-of-sample tests of the predictive validity of causal models, we speculate that equalizing multiple regression model coefficients should be tailored to the forecasting problem and conditions, such as sample size—more equalizing when samples are small—and prior knowledge on the relative importance of variables—more equalizing when prior knowledge is weak. Equalizing was tested in election forecasting using eight independent econometric election forecasting models estimated using data that was standardized and positively correlated with the dependent variable. Where equalizing coefficients by 100% amounts to using equal weights, equalizing by between 10% and 60% reduced the absolute errors of the forecasts from all of the models. Equalizing by between 70% and 100% reduced the errors of forecasts from seven of the eight models (Graefe, Armstrong, and Green 2014).

9. Incorporate prior knowledge by averaging the regression coefficients with a priori *coefficients*—The procedure was described by Wold and Juréen (1953), who called it "conditional regression analysis." It involves averaging coefficients the modeler has derived from prior knowledge with the coefficients estimated from the data using regression analysis. Drop the variable from the model if the regression-estimated coefficient has the opposite sign to the *a priori* sign. Use equal weights for averaging each pair of coefficients, unless there are strong *a priori* reasons for using differential weights.

10. Develop alternative multiple and one-variable regression models and combine their forecasts—Consider the case of combining the forecasts of economists who ascribe to different economic theories. In one study, combinations of forecasts from two economists with similar theories reduced the mean square errors of 12-month ahead real GNP growth forecasts by 11% on average, whereas combinations of forecasts from two economists whose theories were dissimilar reduced errors by 23%. The equivalent analysis comparing combinations from pairs of economists who used similar forecasting techniques with those from pairs who used dissimilar techniques found a 2% error reduction for similar technique combinations, and a 21% error reduction for dissimilar technique combinations (Table 2 in Batchelor and Dua 1995). The error-reduction advantage for diversity in combinations was *much* larger for five of the six other comparisons in the study, in which economists with similar/dissimilar theories/techniques forecast the GNP deflator, corporate profit growth, and the unemployment rate.

Another study 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. The combined forecasts reduced error compared to the typical individual model's forecast by 36% across the 15 elections in the study (Graefe, Armstrong, and Green 2014).

Occam's razor suggests that a combination of forecasts from simple one-variable regressions might be more accurate than forecasts from a multiple-variable regression model using the same variables. Combining the forecasts of single-variable regressions overcomes some of the problems inherent in multiple regression. The approach was tested by making ten-year ahead ex *ante* population forecasts for 100 counties in North Carolina. The researchers used six causal variables to develop six one-variable models, then calculated a combined forecast. They also developed a multiple regression model with the same six variables. The Mean Absolute Percentage Error for the combined single-regression forecasts was 39% lower than that for forecasts from the multiple regression models (Namboodiri and Lalu 1971). The finding was surprising at the time, and perhaps even now for many statisticians. One comparative study can provide only weak evidence, but we are not aware of any empirical test that found multiple regression forecasts were more accurate than combined one-variable regression forecasts

11. Validate models using only out-of-estimation-sample forecast errors—For cross-sectional data, use hold-out samples. If there are few observations, use the jack-knife method; that is, use all but one data point to estimate the model and make a forecast for the excluded observation. Then replace that observation and pick another observation to exclude, and so on until forecasts have been made for all data points.

For time series, use successive updating. For example, to test the predictive validity of alternative models for forecasting the next 100 years of global mean temperatures, annual forecasts were made for horizons one to 100 starting in 1851 by Green, Armstrong, and Soon (2009). They repeated the process, starting in 1852, then 1853, and so on until they had obtained forecast errors for 157 one-year-ahead forecasts, 156 two-years-ahead,… and 58 one-hundred-years-ahead.

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. Mistakes are common in forecasting.

13. Sign an oath that your analyses were ethical—Cheating often occurs in forecasting as described in Goodwin (2017, Chapter 7). Also consider the high level of cheating and questionable practices associated with the use of regression analysis in particular (see Armstrong and Green 2017). We recommend that analysts and authors voluntarily sign an oath that they *will not* engage in unethical or questionable behavior, and later to confirm that they did not engage in such behavior. That procedure has been found to be effective in reducing unethical behavior (see Mazar, Amir and Ariely 2008).

These guidelines may be familiar to much older readers who were acquainted with best practice in an age when extensive *a priori* analysis was the norm. Ten of the 13 guidelines described in Exhibit 2 were used in (Armstrong 1970b). Guidelines 8 and 10 were the only ones we could not trace prior to 1970. (Guideline 13 went unstated because unethical behavior in science was previously rare and was taken for granted.)

In those early years, there was also a pragmatic reason for undertaking extensive *a priori* 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 the next day; then repeat the whole process to remedy mistakes.

Judgmental bootstrapping (13)

In the early 1900s, a method was developed to provide forecasts of the size of the upcoming corn harvest in the U.S.. In the 1940s, the method was used successfully for personnel selection (Meehl 1951), and this has been supported by a continuing stream of research in universities since then (e.g., Dawes and Corrigan 1974; Grove, Zaid, Lebow, Snitz, and Nelson, C. ~~Grove, et al.~~ 2000). The method uses regression analysis to estimate coefficients for 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. Among researchers in forecasting, this method has, in recent decades, been called "judgmental bootstrapping." In effect, it uses a quantitative model of the experts' use of causal information for forecasting to improve upon the expert's forecast accuracy.

In comparative studies to date, the bootstrap model's forecasts are more accurate than those of the experts. The gain in accuracy occurs because the quantitative model is more consistent than the expert is in applying the expert's mental model. In addition, the model does not get distracted by irrelevant information, nor does it get tired or irritable. The developer of the bootstrap model can also ensure that the model excludes irrelevant variables, such as the beauty or height of an applicant for a position as a senior executive.

The first step for developing a judgmental bootstrap model is to ask experts to identify causal variables based on their domain knowledge. Then ask them to make predictions using data on the variables. For example, they could be asked to forecast which students are most likely to be successful doctoral candidates.

Judgmental bootstrap models can be estimated from experts' predictions made on the basis of hypothetical data on the causal variables. Doing so allows the forecaster to ensure that the causal variables vary substantially and independently of one another. That use of experimental design overcomes many of the deficiencies of multiple regression. It also enables one to make forecasts for situations where actual data are not available. Once developed, the bootstrap model can provide forecasts at a low cost and make forecasts for different situations—e.g., by changing the features of a new product.

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 found that judgmental bootstrapping forecasts were more accurate than those from unaided judgments in eight of 11 comparisons, with two tests finding no difference and one finding a small loss in accuracy (Armstrong 2001a). The one failure occurred when the experts relied on an irrelevant variable that was not excluded from the bootstrap model. The typical error reduction was about 6% relative to unaided judgment.

Many universities taught the methods to their students, but we are aware of only one that adopted the method, despite the fact that one of the earliest validation tests showed that it provided a more accurate and less expensive way of predicting success in a PhD program (Dawes 1971).

In 2002, the Oakland Athletics baseball team adopted a version of judgmental bootstrapping. Attempts were made to block the use of the method by the experts who traditionally used their judgment to make these the selection decisions; the managers, owners, and scouts. But the new general manager persisted, and the team did well. Other professional sports teams subsequently adopted the method, improving both won-lost ratios and profitability (Armstrong 2012a).

Segmentation (14)

Segmentation involves structuring the forecasting problem in order to make best use of knowledge and data about parts, or sub-populations, that are expected to behave differently. Appropriate methods are used to make forecasts for each part, and the forecasts for the parts are then added to derive a forecast for the whole. Much attention was devoted to segmentation when it was used to forecast the Kennedy-Nixon 1960 election (Pool, Abelson and Popkin 1965).

The method was also used to forecast air travel demand in ten years' time by the Port of New York Authority in 1955. To do so, their analysts divided airline travelers into segments of 130 business traveler types and 160 personal traveler types. The personal travelers were segmented by age, then by occupation, income, and education: and the business travelers were segmented by occupation, then industry, then income. Data on each segment were 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).

To forecast using segmentation, identify important causal variables that can be used to define the segments and their priorities. Then determine cut-points—e.g., different age categories of people—for each variable such that more cut-points should be used when there are non-linearities in the relationships, and fewer cut points should be used when the samples of data are smaller. Next, forecast the population of each segment and the behavior of the population within each segment by using the typical behavior. Finally, combine the population and behavior forecasts for each segment and sum across segments. Clearly, this is a method that is most likely to be useful when large amounts of data are available.

A review of this literature on segmentation is provided in Armstrong (1985, Chapter 9). Segmentation has advantages over regression analysis for situations where variables are interrelated, the effects of variables are non-linear, prior causal knowledge is good, and the sample size for each of the segments is large. These conditions occurred to a reasonable extent in a study where data from 2,717 gas stations were used to estimate 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 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 mean absolute percentage error of the regression model forecasts of weekly gasoline sales volumes was 58%, while that of the segmentation model was 41% (Armstrong and Andress 1970).

While the evidence on predictive validity is not substantial, the method is sensible, as it is based on decomposition. While interest in segmentation fell away after the 1970s, we expect that it would be more useful now than ever before, given the wide availability of large databases.

Knowledge models (15)

Some forecasting problems are characterized by knowledge that many causal variables are important, or by known causal relationships for which little quantitative data are available. 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, or which applicants for immigration are most likely to pose a security risk. These are problems for which knowledge models, as opposed to data models, are suitable. Knowledge models are specified as simple linear models with equal weights unless knowledge is sufficient to specify differential weights and more complex relationships.

Benjamin Franklin proposed a form of knowledge model in a letter to his friend, Joseph Priestley, who had written to Franklin about a "vexing decision" he was struggling to make. His method was to list pros and cons for each alternative giving each a subjective weight, then to sum the lists to determine which alternative has the biggest net score in its favor. Franklin called his approach, "prudential algebra".

Apparently, Franklin's proposal attracted little attention. A similar approach, called "experience tables", was used in the early 1900s for deciding which prisoners should be given parole (Burgess 1936). Gough (1962) tested the method against a regression model for making predictions about parolee success, and found that regression provided no improvements in accuracy. Despite the finding, the experience table approach gave way to the allure of multiple regression analysis for parole prediction.

Another version of prudential algebra was "configural analysis," which came into limited use in the mid-1900s. The approach was found to have predictive validity (e.g., see Babst, Gottfredson and Ballard, 1968).

Formatted: No underline

Formatted: No underline

Formatted: No underline

More recently, Graefe and Armstrong (2011a) developed a version of Franklin's "prudential algebra" that they called the "index method." Substantial validation research has been published on the approach under the index method name.

To specify a knowledge model, use prior experimental evidence and domain knowledge to identify predictor variables, and to determine each variable's directional influence on the outcome of interest. If prior knowledge on a proposed variable's effect is ambiguous or not obvious, do not include the variable in the model. The disadvantage of knowledge models is that it often requires a time-consuming search for relevant evidence.

As with Franklin's pioneering prudential algebra, a basic knowledge model lists all important causal variables defined in such a way as to be positively correlated with whatever is being forecast. When knowledge on the relative effect sizes is weak, use equal weights and variable values as simple as yes, or no (1, or 0). A higher score for an alternative means that it is predicted to be better, or more likely to occur.

Where sufficient historical data are available—including quantitative data on the thing that is being forecast—one can estimate the relationship between the knowledge model scores and quantity being forecast using regression analysis. Quantitative forecasts are then obtained by applying the regression-estimated parameters—constant and score coefficient—to the knowledge model score for a new alternative or situation, such as a potential sports team recruit or call-center employee.

Knowledge 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) reviewed empirical studies that included comparisons of the two methods for forecasting problems in applied psychology, biology, economics, elections, health, and personnel selection. The knowledge model forecasts were more accurate than regression model forecasts for ten of the 13 studies.

One knowledge 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. See Armstrong and Cuzán (2006) for an analysis of this method. We are not aware of any other single model that has matched this level of accuracy in forecasting U.S. presidential elections.

Another test assessed the predictions from a knowledge 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 of which of the pairs was more effective. Judgmental predictions by novices were correct for 54% of the pairs, those with experience in advertising made 55% correct predictions. Copy testing (e.g., showing ads to subjects and assessing their likelihood of purchase) yielded 57% correct predictions. The knowledge model, by contrast, 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 *ex ante* forecast accuracy.

The basic rules for combining within and across methods are: (1) obtain forecasts from all valid evidence-based methods—and no other methods; (2) obtain forecasts from variations of each component method that are the product of diverse experts, data, procedures, and implementations; (3) equally weight forecasts from variations of the component methods; and then (4) equally weight the combined forecasts from each component method unless strong evidence is available that some methods provide more accurate forecasts than others for the problem at hand—in which case specify the weights *before* making the forecasts.

For important problems, we suggest obtaining forecasts from *at least* three component methods and from at least two variations of each component method—forecasts from six variations in total—in order to reduce the risk of biased forecasts. For more details on forecast combining methods, see Graefe, Armstrong, Jones, and Cuzán 2014; and Graefe (2015).

The combining procedures described *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 provides 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 member's 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 effect of combining on forecast accuracy is the consequences of bracketing: a situation in which forecasts lie on opposite sides of the actual value. Because bracketing is always possible, combining should always be used. Thus, when two or more forecasts *from evidence-based methods* can be obtained, the method of combining forecasts should always be used.

There is much research still to be done on combining forecasts. In particular, we need to learn more about (1) how to combine forecasts in order to produce the greatest gains in forecast accuracy, (2) whether and under what conditions some methods contribute more to increase the accuracy of a combined forecast than others and (3) the typical marginal effects on accuracy of adding extra methods and method variations to a forecast combination~~.~~

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, forecaster bias, and small sample sizes in any of the component forecasts. In other words, combining within a single method is likely to be most useful in improving the reliability of forecasts. Given that a particular method might tend to produce forecasts that are biased in a given direction, however, forecasts from a single method are less likely to bracket the true value than are forecasts from different methods, where the directional biases of the methods are more likely to differ.

One review identified 30 studies that compared combinations of forecasts mostly from a single method. The unweighted arithmetic mean error of the combined forecasts was 12.5% smaller than the average error of the typical 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. Moreover, by making use of more information about the situation, combining forecasts across methods is also likely to increase 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 testing the accuracy of combining forecasts across four to six different methods for the seven U.S. Presidential elections from 1992 to 2016. The individual method forecasts were themselves combinations of forecasts from variations of the method, or from different forecasters using the method. In other words, the PollyVote forecast is a product of combining within then across methods. Over the 100 days prior to the elections, the mean absolute error of the PollyVote forecast was, at 1.1 percentage points, smaller than the average errors of *all* of the component method combinations; which ranged from 1.2 to 2.6 percentage points, with a median of 1.8 (Graefe, Armstrong, Jones, and Cuzán 2017).

The PollyVote forecasts thus provided an error reduction of roughly 40% relative to the typical single method combination. Taken together with the finding from the previously mentioned study—that combining within a single provided an average error reduction of 12.5%—a crude estimate of the error reduction that might be expected from *combining within then across forecasting methods* suggests itself: forecast errors could be reduced by more than one-half.

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.

Exhibit 3: Golden Rule of Forecasting--- Checklist II

			Comparisons*		
			N	Error reduction	
		Guideline			
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 many 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 alternative causal models	5	5	22
5.	□	**Combine forecasts from diverse methods**	–	–	--
6.	□	**Avoid adjusting forecasts**	–	-	-
		Totals and Unweighted Average for Guidelines 1 through 4	**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).

The Golden Rule of Forecasting is like the traditional Golden Rule in the sense that it is an *ethical* principle that could be expressed as "forecast unto others as you would have them forecast unto you." The rule is especially useful when objectivity must be demonstrated, as is the case 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.*

We made changes to the previous—2015 version—of the Golden Rule. Most importantly, Guideline 5 now states that one should, "combine forecasts from diverse methods," and Guideline 6 now states that one should, "avoid adjusting forecasts". The support for revised combining guideline is described in the previous section. The adjustment guideline is now a prohibition, first because adjustments are prone to introduce bias away from accurate and towards desired forecasts, and second because following Guideline 1.1.2—to decompose the forecasting problem to make best use of knowledge information and judgment—and the revised Guideline 5—to combine forecasts from diverse methods—ensures that all relevant knowledge and information are included in the forecast, leaving no valid reason for adjusting forecasts.

Bias is a common problem. 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 research review was that forecasters should not make subjective adjustments to forecasts made by quantitative methods. Research in psychology since then continued to support Meehl's findings (see Grove, ~~Zald, Lebo,~~ et al. 2000). Research on adjusting forecasts from 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).

Following the Golden Rule guidelines to decompose the forecasting problem to make best use of what is known about the problem situation and to use diverse valid forecasting methods eliminates the only valid reason for adjusting a forecast. Unsurprisingly, then, 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 presented in this paper*.

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 sub-problems, separately forecasting the level, the trend, and the effect of promotions. 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.

Any stakeholder can use the Golden Rule of Forecasting Checklist. 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 even less time after becoming familiar with it. Forecasters must fully disclose their methods and clearly explain them (Guideline 1.3). To help improve the reliability of the checklist ratings, forecasters should ask at least three people, each working independently, to complete the ratings.

SIMPLICITY IN FORECASTING: OCCAM'S RAZOR

Occam's razor is a principle of science attributed to 14th-century scholar William of Ockham, but it was earlier proposed by Aristotle. The principle is that the simplest explanation is best. The principle also applies to scientific forecasting: forecasters should use methods that are no more complex than is necessary to provide a model that is consistent with the state of knowledge about the situation.

Do forecasters ascribe to Occam's razor? Apparently not: in 1978, when 21 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).

Starting in the 1950s, researchers developed complex statistical models to extrapolate time-series data. The authors described how their models were based on sound mathematical reasoning, and they reported on the ability of the models to fit the data used to estimate the models. The models were popular and widely used by academics and practitioners. But the question of which model provides the most accurate forecasts was not properly addressed until the end of the late-1970s.

At that time, researchers were invited enter their models in a competition to extrapolate 111 unidentified business and economic time-series of monthly, quarterly, and annual data up to six years ahead. Their accuracy of the forecasts from the different methods were assessed against that of the relevant no-change benchmark model forecasts. The simple "naïve models" performed well, and the differences in the accuracy of the forecasts from the better rival models and the naïve models was minor. The findings (Makridakis and Hibon 1979) were published with commentary by 14 respected statisticians in the field at the time. While the commentary was cordial, the study did not convince many of the statisticians of the power of simple models. Makridakis went on to conduct extensions of the competitions, which were referred to as the M-competitions (Makridakis *et al*. 1993, Makridakis and Hibon 2000), all of which concluded that simple methods provided extrapolation forecasts that were competitive with those from the complex methods.

A series of tests from across different kinds of problems—such as the 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 the ForecastingPrinciples.com site.

Exhibit 4: "Simple Forecasting" Checklist: Occam's Razor

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)	[__]

In that paper, our search identified 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 produced *ex ante* forecast errors that were 27% larger than those from the simpler methods. The finding was surprising, because the papers appeared to be proposing the more complex methods with the expectation that they would provide forecasts that were more accurate.

ASSESSING FORECAST UNCERTAINTY

A forecast's uncertainty affects its utility. For example, if demand for automobiles is forecast to increase by 20% next year, manufacturers 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.

There are several ways that one might assess forecast uncertainty. Exhibit 5 presents a checklist of four valid methods to use, along with warnings against the use of two invalid methods. Forecast uncertainty is an assessment of the likely range of forecast errors, so we first describe the measures of error that have been found to be most useful. We then discuss each of the checklist items.

Exhibit 5: Forecast Uncertainty Checklist

❑ **1.** Use empirical prediction intervals or likelihoods estimated from out-of-sample tests

❑ **2.** Decompose errors by source in order to estimate the uncertainty of each

❑ **3.** Use structured judgment to estimate prediction intervals or likelihoods

❑ **4.** Combine alternative valid estimates of uncertainty

❑ **5.** Avoid using statistical fit with historical data to assess uncertainty

❑ **6.** Avoid using tests of statistical significance to assess uncertainty

Error measures

In order to estimate prediction intervals, we suggest calculating the Mean Absolute Deviation (MAD) of forecasts as an obvious measure the importance—as in, the impact on welfare or life-expectancy—of forecast errors. The MAD has the advantages of being easy to calculate, relevant to decision makers and easily understood by them. These are qualities that the measure traditionally favored by statisticians—the Root Mean Square Error, or RMSE—does not possess.

For forecasting problems that are expected to involve asymmetric errors—i.e., negative errors are likely to be predominantly either smaller or larger than positive errors—calculate the logarithms of the forecast and actual values, calculate the errors using the logged values, use those errors to estimate prediction intervals, and then convert the bounds of the intervals back to actual values (Armstrong and Collopy 2001). The errors of time-series forecasts are often asymmetric, especially when the forecasting model applies an additive trend to a situation that is better characterized by constant elasticities.

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.

Use empirical prediction intervals or likelihoods estimated from out-of-sample tests (1)

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). Similarly, the confidence intervals estimated from the fit of a regression model are of no value to forecasters as estimates of prediction intervals (Pant and Starbuck 1990; Soyer and Hogarth 2012).

Uncertainty is most accurately represented using empirical prediction intervals based on out-of-sample forecast errors from the testing of each forecasting method (Chatfield 2001). To that end, simulate the actual forecasting procedure as closely as possible and use the distribution of the errors of the resulting forecasts to assess uncertainty. When analyzing time-series forecast errors, use successive updating to increase the number of predictions. If sufficient validation data are not available, consider using analogous situations.

Decompose errors by source in order to estimate the uncertainty of each (2)

In most situations, there are several sources of forecast error. Given that problem, consider decomposing the error by source, estimate the error due to each, then combine the estimates. For example, in polling to predict the outcomes of political elections, survey researchers report the error based only on sample size. They ignore response error and non-response bias. Thus, the reported 95% confidence intervals are about half as large as they really are, as was shown by Buchanan (1986).

When uncertainty is high—such as with surveying citizens to forecast the effects of a change in a government regulation—response error is likely to be high due to survey respondents' lack of self-knowledge about how they make decisions (see Nisbett and Wilson 1977). *Non*-response can also be a large source of error, because the people who are most affected by the topic of the survey are more likely to respond. While the latter error can be reduced to some extent by the "extrapolation-across-waves" method (Armstrong and Overton 1977), forecasters still need to consider that source of error in their assessment of uncertainty.

As with analyses of judgmental forecasts, regression models' diagnostic statistics ignore key sources 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 with historical data, and even more so when data mining techniques are used to achieve a close fit.

Use structured judgment to estimate prediction intervals or likelihoods (3)

One common judgmental approach to assessing uncertainty is to ask experts to express their confidence in their own judgmental forecasts in the form of 95% prediction intervals. One concern with this approach is that experts are typically overconfident about the accuracy of their forecasts. For example, an analysis of judgmental confidence intervals for economic forecasts from 22 economists over 11 years found the actual values were within the range of their own 95% confidence intervals only 57% of the time (McNees 1992). Another study tracked members of a ten-year panel that provided 13,300 estimates of expected stock market returns by company; the actual returns were within the executives' 80% confidence intervals only 36% of the time (Ben-David, et al. 2013).

There are a number of structured approaches to improve the calibration of judgmental forecasts. Ensure that the judgments are obtained from many experts and obtain independent anonymous estimates. The Delphi technique can be used for that purpose. Ask experts to list all sources of uncertainty and all reasons why they might be wrong. This was shown to be effective by Arkes (2001).

Finally, to improve the calibration of forecasters' estimates of uncertainty in the future, ensure that they receive timely, accurate, frequent, and well-summarized information on what actually happened, and reasons why their forecasts were right or wrong. For example, 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 rain falls 40% of the time (Murphy and Winkler 1984).

Combine alternative valid estimates of uncertainty (4)

The logic behind combining for estimating uncertainty is the same as we described above in relation to combining forecasts. Thus, for example, the estimates of uncertainty based on combined estimates can never be worse than the typical estimate, and the combined estimate will always be better than the typical estimate as long as bracketing of the uncertainty estimates occurs.

Avoid using statistical fit with historical data to assess uncertainty (5)

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 1970a).

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 fit measures, as was shown by Soyer and Hogarth (2012).

Avoid using tests of statistical significance to assess uncertainty (6)

Statistical significance tests do not provide valid estimates of forecast uncertainty. Attempts to use them in that way will likely lead to confusion and poor decision-making. That conclusion is supported by an extensive literature published over more than half a century, as is detailed by Ziliac and McCloskey (2008) and Hubbard (2016).

One experiment presented leading researchers with a treatment difference between two drugs, as well as a "*p*-value" for the difference, and asked them which of the drugs 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. By 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 (McShane and Gal 2015). Many of those teaching statistics also failed to draw logical conclusions as was shown in another experiment by McShane and Gal (2017).

Many of these errors in interpretation have led to decisions that have harmed people. Hauer (2004) described the harm caused by decisions related to automobile traffic safety, such as the "Right-turn-on-red decision." Ziliac and McCloskey (2008) provide many other examples.

To our knowledge, no scientific study has shown that statistical significance testing has led to better forecasts or decisions or scientific contributions. Schmidt (1996) offered this challenge: "Can you articulate even one legitimate contribution that significance testing has made (or makes) to the research enterprise (i.e., any way in which it contributes to the development of cumulative scientific knowledge)?" Schmidt and Hunter (1997) stated that no such cases have been reported, and they repeated the challenge, as we have, as we hereby do again.

DISCUSSION

What recourse do stakeholders have when inaccurate forecasts lead to poor decisions? The answer has traditionally been that there is none, because it has not been possible to distinguish between forecasts that were wrong due to random or unpredictable changes in the situation and forecasts that were wrong due to the incompetence of the forecaster. This paper offers stakeholders the means by which to hold *forecasters who fail to follow evidence-based* procedure to account. The account is in the form of checklists of evidence-based principles and methods. Stakeholders can use the checklists to become educated funders and consumers of forecasts—that is, to effectively implement the principle of *caveat emptor*, or "let the buyer beware."

Forecasters can use the checklists to improve the accuracy of their forecasts, and to protect themselves from claims against them by communicating that they have followed the checklists. Forecasters who follow the checklists might—as do medical practitioners—obtain additional protection against being ruined by claims of damages by arranging insurance on the understanding that they will follow the proper forecasting procedures listed in the checklists.

CONCLUSIONS

Experimental research over the past century has identified 17 evidence-based forecasting methods, and that the common approach of relying on the unaided judgments of experts leads to a never-ending stream of disastrous forecasts that delight the media (see, e.g., Cerf and Navasky 1998). We described the evidence-based methods, along with their estimated effects on the accuracy of *ex ante* forecasts.

The use of those methods substantially improved the accuracy of forecasts relative to the accuracy of forecasts from commonly used methods (Exhibit 1). We found error reductions ranged from approximately 5%—for damped-trend extrapolation, decomposition by seasonality, and judgmental forecasting—to over 50% when methods including simulated interaction and knowledge models were used. Additional improvements in accuracy—perhaps amounting to error reductions of one-half or more in some situations—are achievable by combining forecasts from diverse methods.

Gains in accuracy are larger for longer-term forecasts than for shorter-term forecasts, and for complex situations for which large forecast errors have been common.

Clients who are interested in accurate forecasts should require forecasters to adhere to the five evidence-based checklists provided in this paper. The checklists can help forecasters to follow evidence-based forecasting principles and to implement the evidence-based forecasting methods. Checklists have been successful for obtaining compliance in other domains. Clients, as well as other forecast stakeholders and commentators, should use the checklists to assess whether forecasts that are important to them were the product of scientific forecasting procedures.

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). Boston, MA: Kluwer Academic Publishers. FC

Armstrong, J. S. (2012a). Predicting job performance: The moneyball factor. *Foresight*, 25, 31-34. AO

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

Armstrong, J. S. (2010). *Persuasive Advertising: Evidence-based principles.* New York: Palgrave Macmillan. AO

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

Armstrong, J. S. (2007b), Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries, *International Journal of Forecasting*, 23, 335-336. 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

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, English (United States), Not Expanded by / Condensed by , Pattern: Clear

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

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

Armstrong, J.S. & Patnaik, S. (2009). Using quasi-experimental data to develop empirical generalizations for persuasive advertising. *Journal of Advertising Research*. 49, 170-175. AO

Armstrong, J. S. & Yokum, T. (1994), Effectiveness of monetary incentives: Mail surveys of multinational professional groups, *Industrial Marketing Management*, 23, 133-136. AO

Babst, D.V., Gottfredson, D.M., & Ballard, Jr., K.B. (1968). Comparison of multiple regression and configural analysis techniques for developing base expectancy tables. *Journal of Research in Crime and Delinquency*, 5, 72-80.

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

Ben-David, I., Graham, J. R., & Harvey. C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, 128, 1547-1584. FC~~.~~

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

Bradburn, N. M., ~~Seymour~~ Sudman, S., Wansink, B. (2004~~),~~). *Asking Questions: The Definitive Guide to Questionnaire Design -- For Market Research, Political Polls, and Social and Health Questionnaires*, 2nd Edition. New York: John Wiley & Sons. NS~~.~~

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

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Bold

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Font: 12 pt, No underline

Formatted: Font: 12 pt, No underline

Formatted: Font: 12 pt, No underline

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Check spelling and grammar

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

Chamberlin, T. C. (1890). The method of multiple working hypotheses. Reprinted in 1965 in *Science, 148*, 754-759. NF ~~.~~

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

Chen C, Twycross J, Garibaldi JM (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, 12(3): e0174202. https://doi.org/10.1371/journal.pone.0174202

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.

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

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

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

Einhorn, H. J. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly*, *36*, 367-378.

Einhorn, H. J. & R. Hogarth (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171-192. FC

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. FC ~~.~~

Gigerenzer, G., Todd, P.M., & T~~t~~he ABC Research Group (1999). *Simple Heuristics that Make us Smart*, New York: Oxford University Press. FC ~~.~~

Goodwin, P. (2017) Forewarned: A sceptic's guide to prediction. London: Biteback Publications. FC ~~.~~

Formatted: Default Paragraph Font, Font: 12 pt, Check spelling and grammar

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

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. NF~~.~~

Graefe, A. (2017a). Political markets. In Arzheimer, K., & 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. (2011a). 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. (2011b). 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*. AO~~.~~

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

Green, K. C., Armstrong, J. S., & Soon (2009). Validity of climate change forecasting for public policy decision making, *International Journal of Forecasting*, 25, 826-832. AO~~.~~

Grove, W. M., Zaid, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. ~~et. al.~~ (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 19-30.

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition*, 5, 215-231.

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Strikethrough

Formatted: Default Paragraph Font, Font: 12 pt, Strikethrough

Formatted: Strikethrough

Formatted: Strikethrough

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Field Code Changed

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

Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis and Prevention*, *36*, 495-500.

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, A. 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. NR~~.~~

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

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~~2017). The power of bias in economics research. ~~Deakin University Economics Series Working Paper SWP 2016/1~~The Economic Journal, 127, F236-F265. FC~~.~~

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)

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. FC~~.~~

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. FC

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. FC~~.~~

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

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

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., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J. Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting* 1, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M-2 Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9, 5–23.

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

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of the Royal Statistical Society A*, 142, 97–145.

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.~~

Formatted: Default Paragraph Font, Font: 12 pt, Check spelling and grammar

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Check spelling and grammar

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

~~Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of the Royal Statistical Society A*, 142, 97–145.~~

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]

McShane, B. B. & Gal, D. (2017). Statistical significance and the dichotomism of evidence. *Journal of the American Statistical Association* (forthcoming)

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]

Morgenstern, O. (1963). On the Accuracy of Economic Observations, Princeton: Princeton University Press.[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. . [NF]

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]

Pool, I. de S., Abelson, R. P., & Popkin, S. L. (1965). *Candidates, Issues and Strategies: A computer simulation of the 1960 and 1974 presidential elections*. Cambridge, M~~ass.~~A: MIT Press. [NF] .

Rhode, P.W. & Stumpf, 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. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115–129.

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.

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Field Code Changed

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Indent: Left: 0", Hanging: 0.33"

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Field Code Changed

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Field Code Changed

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in Harlow, L. L., Mulaik, S. A. & Steiger, J. H. *What if there were no Significance Tests?* London: Lawrence Erlbaun.

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

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

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. NF~~.~~

Vokurka, R. J., Flores, B. E., Pearce, S. L. (1996) Automatic feature identification and graphical support in rule-based forecasting: ~~A~~ a comparison. *International Journal of Forecasting*, 12, 495-512.

Wold, H. & Juréen, L. (1953). Demand Analysis: A study in econometrics. 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 19,000
Text only 15,200

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Italic

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt, Not Expanded by / Condensed by

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt

Formatted: Default Paragraph Font, Font: 12 pt