Golden Rule of Forecasting: Be conservative

Armstrong, J. Scott and Green, Kesten C. and Graefe, Andreas

Wharton School, University of Pennsylvania, and Ehrenberg-Bass Institute, University of South Australia School of Business, and Ehrenberg-Bass Institute, Department of Communication Science and Media Research, LMU Munich

6 February 2014

Golden Rule of Forecasting: Be Conservative

6 February 2014

Working Paper Draft

GoldenRule 364-R.doc

J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
and Ehrenberg-Bass Institute, Adelaide, Australia
armstrong@wharton.upenn.edu

Kesten C. Green
University of South Australia School of Business
and Ehrenberg-Bass Institute, Adelaide, Australia
kesten.green@unisa.edu.au

Andreas Graefe
Department of Communication Science and Media Research
LMU Munich, Germany
a.graefe@lmu.de

Abstract

This paper proposes a unifying theory of forecasting in the form of a Golden Rule of Forecasting. The Golden Rule is to *be conservative.* A conservative forecast is consistent with cumulative knowledge about the present and the past. To be conservative, forecasters must seek all knowledge relevant to the problem, and use methods that have been validated for the situation. A checklist of 28 guidelines is provided to implement the Golden Rule. This article's review of research found 150 experimental comparisons; all supported the guidelines. The average error reduction from following a single guideline (compared to common practice) was 28 percent. The Golden Rule Checklist helps forecasters to forecast more accurately, especially when the situation is uncertain and complex, and when bias is likely. Non-experts who know the Golden Rule can identify dubious forecasts quickly and inexpensively. To date, ignorance of research findings, bias, sophisticated statistical procedures, and the proliferation of big data have led forecasters to violate the Golden Rule. As a result, despite major advances in forecasting methods, evidence that forecasting *practice* has improved over the past half-century is lacking.

Key words: accuracy, analytics, bias, big data, causal forces, causal models, combining, complexity, contrary series, damped trends, decision-making, decomposition, Delphi, ethics, extrapolation, inconsistent trends, index method, judgmental bootstrapping, judgmental forecasting, nowcasting, regression, risk, shrinkage, simplicity, stepwise regression, structured analogies.

Introduction

Imagine that you are a manager who hires a consultant to predict profitable locations for stores. The consultant applies the latest statistical techniques to large databases to develop a forecasting model. You do not understand the consultant's procedures, but the implications of the forecasts are clear: invest in new outlets. The consultant's model is based on statistically significant associations in the data and fits the data closely. Your colleagues are impressed by the consultant's report, and support acting on it. Should you?

To answer that question, and the general question of how best to go about forecasting, this paper proposes a general rule. Further, to help forecasters make more accurate forecasts and to help decision makers assess whether forecasts were derived from proper procedures, guidelines are provided on how to implement the rule.

The proposed rule is a *Golden Rule of Forecasting*, because it applies to all forecasting problems. The Golden Rule is to *be conservative*. Conservatism requires a valid and reliable assessment of the forecasting problem in order to make effective use of cumulative knowledge about the historical situation, causality, and appropriate evidence-based forecasting procedures.

This paper is concerned with the effect of conservatism on point forecasts. Point forecasts are nearly always useful for decision-making. Conservatism is likely also to be useful for assessing uncertainty, but we do not address that issue.

The Golden Rule is relevant to all forecasting problems. It is especially important when bias is likely, and when the situation is uncertain and complex. Such situations are common in business, and in public policy, as with forecasts of the effects of economic policies and regulations.

The Golden Rule Checklist

The checklist in Exhibit 1 provides a set of guidelines for how and when to apply the Golden Rule. It can help forecasters to be conservative and decision makers to identify poor forecasts. Our intent was first that the guidelines follow logically from the principle of conservatism as defined in this paper. We then searched for research to test the guidelines. Most of the 28 guidelines are based on experimental evidence from comparative studies. Some of the guidelines were deduced from indirect evidence, and a few are based only on logic.

Exhibit 1 also shows the improvements in accuracy achieved by following a guideline relative to using a less-conservative approach. Percentage error reductions are provided for reasons of comparability across the studies and the guidelines. The average error reduction per guideline was 28 percent, so larger gains in accuracy are likely by using many guidelines.

Remarkably, no matter what the criteria, data set, forecast horizon, or type of problem, the authors of this paper were unable to find any studies in which following any of the Checklist guidelines harmed accuracy.

Insert Exhibit 1 about here

Evidence on the Golden Rule was obtained using computer searches of the literature, seeking help from key researchers, posting requests for relevant papers on the Internet, and investigating references in important papers. To ensure the evidence is properly summarized and to check whether any relevant evidence had been overlooked, the authors sent email messages to the lead authors of articles that were cited in substantive ways. Reminder messages were sent to authors who did not respond and to some co-authors. Responses were received for 84 percent of authors for whom valid email addresses were found.

Problem formulation (1)

Forecasters should first formulate the forecasting problem. Proper formulation allows for effective use of cumulative knowledge about the situation being forecast and about relevant evidence-based forecasting methods.

Obtain and use all important knowledge and information (1.1)

Forecasters should endeavor to use all relevant, reliable, and important information, and no more. To do so, they typically need to consult domain experts in order to acquire information on the situation to be forecast. One way is to ask a heterogeneous group of experts to independently list relevant variables, the directions and strengths of their effects, the support for their judgments, and recommendations on which data are relevant to the problem.

Forecasters should search the literature for evidence about causal relationships. Especially useful are meta-analyses, where structured procedures are used to summarize the findings of experimental studies.

Nonexperimental data might be useful in situations where experimental data are lacking, but should be used with great caution. Researchers often mistakenly conclude that statistical associations in non-experimental data show causality. Consider, for example, the many forecasts that eating certain foods will increase your life span, and exposure to tiny doses of certain chemicals will decrease it (see, e.g., Kabat, 2008, on health risk studies).

Exhibit 1: Golden Rule Checklist
(With evidence on percentage error reduction, and number of comparisons)

		Guideline	% error reduction	Comparisons* Size	Comparisons* All
1.		**Problem formulation**			
1.1		Obtain and use all important knowledge and information			
1.1.1	☐	Decompose to best use knowledge, information, and judgment	35	13	21
1.1.2	☐	Select evidence-based methods validated for the situation	16	4	8
1.2		Avoid bias, by…			
1.2.1	☐	concealing the purpose of the forecast		0	0
1.2.2	☐	specifying multiple hypotheses and methods	50	1	1
1.2.3	☐	obtaining signed ethics statements before and after forecasting		0	0
1.3	☐	Provide full disclosure for independent audits and replications		0	1
2.		**Judgmental methods**			
2.1	☐	Avoid unaided judgment		0	0
2.2	☐	Use alternative wording and pretest questions		0	0
2.3	☐	Ask judges to write reasons for and against the forecasts	8	2	3
2.4	☐	Use judgmental bootstrapping	6	1	11
2.5	☐	Use structured analogies	48	5	5
2.6	☐	Combine independent forecasts from judges	12	10	13
3.		**Extrapolation methods**			
3.1	☐	Use the longest time-series of valid and relevant data		0	0
3.2	☐	Decompose by causal forces	60	9	9
3.3		Be conservative when forecasting trends, if the…			
3.3.1	☐	series is variable or unstable	5	10	10
3.3.2	☐	historical trend conflicts with causal forces	30	10	10
3.3.3	☐	forecast horizon is longer than the historical series	43	1	1
3.3.4	☐	short and long-term trend directions are inconsistent		0	0
3.4		Estimate seasonal factors conservatively, when…			
3.4.1	☐	they vary substantially across years	25	3	3
3.4.2	☐	few years of data are available	15	14	15
3.4.3	☐	causal knowledge is weak		0	0
3.5	☐	Combine forecasts from alternative extrapolation methods, data	15	5	5
4.		**Causal methods**			
4.1	☐	Use prior knowledge to select variables and estimate effects	32	2	2
4.2	☐	Estimate variable weights conservatively	5	1	1
4.3	☐	Use all important variables	46	2	4
4.4	☐	Combine models that use different information, procedures	21	5	5
5.	☐	**Combine forecasts from diverse evidence-based methods**	18	16	20
6.	☐	**Avoid unstructured judgmental adjustments to forecasts**	72	1	2
			Total comparisons	**115**	**150**

* Size: number of comparisons with findings on effect sizes. All: number of comparisons with findings on effect direction.

Conservative forecasting requires knowing the current situation, and so forecasters should seek out the most recent data. For example, to forecast demand for ice cream in Sydney in the coming week, it would be important to know that a big cruise ship was due to arrive and that the most recent forecast was for a week of perfect beach weather. Similarly, to forecast the demand for building products in New Zealand, it would be important to know that an earthquake had leveled the city of Christchurch and to learn government policies on rebuilding.

The requirement to take account of recent information should not, however, be confused with claims that things are so different now that historical data and knowledge, are irrelevant or unimportant. Such claims should be met with demands for evidence. The mantra that the world in general or a particular situation is outside of previous experience is popular among CEOs and political leaders. U.S. president Dwight Eisenhower, for example, stated that, "Things are more like they are now than they ever were before." The belief that things are different now has led to disastrous forecasts by governments, businesses, and investors. The many and varied speculative bubbles from Dutch tulip bulbs to Dot.com stocks provide examples of the failed forecasts of investors who believed the situation was different from previous experience. See Schnaars (1989) for further examples.

Decompose the problem to best use knowledge, information, and judgment (1.1.1)

Decomposing the problem may enable forecasters to draw upon more knowledge, and to use the knowledge more effectively. Decomposition is conservative in part because the errors from forecasts of the parts are likely to differ in direction and thus to offset each other in the aggregate. Decomposition improves accuracy most when uncertainty is high.

Decomposition allows forecasters to better match forecasting methods to the situation, for example by using causal models to forecast market size, using data from analogous geographical regions to extrapolate market-share, and using information about recent changes in causal factors to help forecast trends. For some problems, however, paucity of knowledge or data may prevent decomposition.

Additive decomposition involves making forecasts for segments and then adding them, a procedure that is also known as segmentation, tree analysis, or bottom-up forecasting. Segments might be a firm's sales for different products, geographical regions, or demographic groups. Forecast each segment separately, and then add the forecasts.

One type of additive decomposition that can improve the accuracy of time-series forecasts is to estimate the current status or initial value—a process that is sometimes referred to as nowcasting—then add the trend forecast. The repeated revisions of official economic data suggest that uncertainty about the current level is common. For example, Runkle (1998) found that the difference between initial and revised estimates of quarterly GDP growth from 1961 to 1996 varied from 7.5 percentage points upward

to 6.2 percentage points downward. Zarnowitz (1967) found that about 20 percent of the total error in predicting GNP one-year-ahead in the U.S. arose from errors in estimating the current GNP. When data are subject to political interference, the problem is greater still.

Because data on the current level are often unreliable, forecasters should seek alternative estimates. Consider combining the latest survey data with estimates from exponential smoothing (with a correction for lag), or with a regression model's estimate of the level (at t=0). Armstrong (1970), for example, estimated a cross-sectional regression model using annual sales of photographic equipment in each of 17 countries for 1960-65. Current sales were estimated by combining econometric estimates with survey data on trade and production. Backcasts were then made for annual sales for 1955 to 1953. One approach started with the survey data and added the trend over time by using an econometric model. Another approach used a combination of the estimates from the survey data and the econometric estimates of the starting values and then added the trend. No matter what the weights, the combination was always more accurate than the use of only survey data. The *a priori* weights reduced the backcast errors for 14 of the 17 countries. On average across the countries, the MAPE was reduced from 30 percent to 23 percent, an error reduction of 23 percent.

Armstrong (1985, pp. 286–287) reports on nine studies on additive decomposition, all of which showed gains in forecast accuracy. Only one of the studies (Kinney Jr. 1971) included an effect size. That study, on company earnings, found that MAPE was reduced by 17 percent in one comparison and 3.4 percent in another.

Dangerfield and Morris (1992) used exponential smoothing models to forecast all 15,753 unique series derived by aggregating pairs of the 178 monthly time-series used in the M-Competition (Makridakis et al. 1982) that included at least 48 observations in the specification set. The additive decomposition forecasts derived by combining forecasts from exponential smoothing models of the individual series were more accurate for 74 percent of two-item series. The MAPE of the bottom-up forecasts was 26 percent smaller than for the top-down forecasts.

Jørgensen (2004) found that when seven teams of experts forecast project completion times, the errors of bottom-up forecasts were 49 percent smaller than the errors of direct forecasts.

Carson, Cenesizoglu, and Parker (2011) forecast total monthly U.S. commercial air travel passengers for 2003 and 2004. They estimated an econometric model using data from 1990 to 2002 in order to directly forecast aggregate passenger numbers. They used a similar approach to estimate models for forecasting passenger numbers for each of the 179 busiest airports using regional data, and then added across airports to get an aggregate forecast. The mean absolute error (MAE) from the recomposed forecasts was about half that from the aggregate forecasts, and was consistently lower over horizons from 1-month-ahead to 12-months-ahead.

A study on forecasting U.S. lodging market sales used an econometric model with successive updating to provide 28 forecasts from 1965 through 1971. The MAPE was reduced by 31 percent when the starting level was based on a combination of the survey data and the econometric forecast. A similar test, done with forecasts based on an extrapolation model, found the MAPE was reduced by 50 percent (Tessier and Armstrong 2014, this issue).

Additive decomposition enables forecasters to include information on many important variables when there are large databases. For example, Armstrong and Andress (1970) used data from 2,717 gas stations to estimate a stepwise regression model that included 19 variables selected based on domain knowledge (e.g. building age and open 24 hours). The model was then used to forecast sales for 3,000 holdout gas stations. Forecasts were also obtained from a segmentation model (Automatic Interaction Detector) that used 11 of the initial 19 variables. The segmentation model forecasts had a MAPE of 41 percent compared to 58 percent for the regression model's forecasts, an error reduction of 29 percent. The finding is consistent with the fact that segmentations can incorporate more information than regression analysis.

Multiplicative decomposition involves dividing the problem into elements that can be forecast and then multiplied. For example, multiplicative decomposition is often used to forecast a company's sales by multiplying forecasts of total market sales by forecasts of market share. As with additive decomposition, this is expected to be most useful when the decomposition allows a more effective use of information and when there is much uncertainty. If there is little uncertainty, then little gain is expected.

Perhaps the most widely used application of decomposition is to obtain separate estimates for seasonal factors for time-series forecasts. For forecasts over an 18-month horizon for 68 monthly economic series from the M-competition, Makridakis et al. (1982) showed that seasonal factors reduced the MAPE by 23 percent.

MacGregor (2001) tested the effects of multiplicative decomposition in three experimental studies of judgmental forecasting that involved 31 problems that involved high uncertainty. For example, how many pieces of mail were handled by the U.S. Postal service last year? The subjects made judgmental forecasts for each component. The averages of the forecasts for each component were then multiplied. Relative to directly forecasting global values, decomposition reduced median error ratios by 36 percent in one study, 50 percent in another, and 67 percent in the third (MacGregor's Exhibit 2).

Select evidence-based forecasting methods validated for the situation (1.1.2)

Forecasting methods that are suitable for one situation may not be suitable for another, and some commonly used methods are not suitable for any situations. Forecasters should therefore use only procedures that have been empirically validated under conditions similar to those of the situation being

forecast. Fortunately, there is much evidence on which forecasting methods are most accurate under which conditions. The evidence, derived from empirical comparisons of the out-of-sample accuracy of forecast from alternative methods, is summarized in *Principles of Forecasting* (Armstrong 2001c). The handbook is a collaborative effort by 40 forecasting researchers and 123 expert reviewers.

Despite the extensive evidence on forecasting methods, many forecasters overlook that knowledge. Consider the IPCC dangerous manmade global warming forecasts that have been used as the basis for expensive government policies (Randall et al. 2007). An audit found that procedures used to generate these forecasts violated 72 of the 89 relevant forecasting principles (Green and Armstrong 2007a).

Do not assume that published forecasting methods have been validated. Many statistical forecasting procedures have been proposed simply on the basis of experts' opinions or inadequate validation studies. An example of the latter is a published model for forecasting sales of high-technology products that was tested on only *six* holdout observations from three different products. A reanalysis of the model's performance using a more extensive dataset, consisting of 14 products and 55 holdout observations, found no evidence that the utility-based model yields more accurate forecasts than a much simpler evidence-based extrapolation model (Goodwin and Meeran 2012).

Further, do not assume that well-known and widely-used statistical forecasting techniques have been tested. For example, in a 1992 survey of 49 forecasting experts at the 1987 International Symposium on Forecasting, over half reported that the Box-Jenkins method was useful for forecasting (Collopy and Armstrong 1992a). However, little validation research had been done despite many journal articles and extensive applications.

When validation tests were done, Box-Jenkins procedures were less accurate than evidence-based procedures. The M2- and M3-Competitions compared the accuracy of Box-Jenkins forecasts against damped trend and combined forecasts, two conservative benchmark methods. The combined forecast was the simple average of three ways to use moving averages: exponential smoothing with no trend, Holt's linear exponential smoothing with full trend, and exponential smoothing with damped trend. The M2-Competition involved 29 series and 30 time horizons, and the M3-Competition involved 3,003 series and 18 time horizons (Makridakis, Chatfield, Hibon, Lawrence, Mills, Ord, and Simmons 1993, Exhibit 3; Makridakis and Hibon 2000, Table 6). Averaging across all time-series and all forecast horizons, the MAPE of the damped trend forecast was 28 percent smaller than the MAPE of the Box-Jenkins forecasts in the M2-Competition and 3 percent smaller in the M3-Competition. The combined forecast error was 27 percent smaller than the Box-Jenkins error in the M2-Competition and 4 percent smaller in the M3-Competition.

Statisticians have generally shown little interest in how well their proposed methods perform in empirical validation tests. A check of the Social Science and Science Citation Indices (SSCI and SCI) found that four key comparative validation studies on time-series forecasting were cited only three times per year between 1974 and 1991 in all the statistics journals indexed (Fildes and Makridakis 1995). Many thousands of empirical time-series studies were published over that period. In other words, most researchers ignored cumulative knowledge about forecasting methods.

Forecasters should validate any method they propose against evidence-based methods. Clients should ask about independent validation testing rather than assume that it was done. For example, independent evaluations of popular commercial programs sold by Focus Forecasting concluded that these forecasts were substantially less accurate than forecasts from exponential smoothing (Flores and Whybark 1986; Gardner and Anderson 1997) and damped smoothing (Gardner, Anderson-Fletcher, and Wickes 2001).

Avoid bias (1.2)

Forecasters sometimes depart from prior knowledge due to biases they may be unaware of, such as optimism, or using the most familiar method and most accessible data. Financial and other incentives, deference to authority, and confusing forecasting with planning can also cause forecasters to ignore prior knowledge or to choose unvalidated methods.

Avoid bias by concealing the purpose of the forecast (1.2.1)

Ensuring forecasters do not know the purpose of the forecast can avoid biasing them towards producing forecasts that promote the purpose. To implement this guideline, give the forecasting task to independent forecasters who are unaware of the purpose.

Avoid bias by specifying multiple hypotheses and methods (1.2.2)

Obtaining experimental evidence on multiple reasonable hypotheses is an ideal way to avoid bias. Following this guideline should help to overcome even unconscious bias by encouraging the forecaster to test unfavored alternatives. The approach has a long tradition in science as described by Chamberlin (1890, 1965). For example, to assess the effects of a medical treatment, one must show how it performs against alternative treatments, including no treatment. Prasad et al. (2013) summarized findings from the testing of a variety of medical procedures and found that "of the 363 articles testing standard of care, 146 (40.2%) reversed that practice, whereas 138 (38.0%) reaffirmed it" (p. 1).

Forecasters should generally consider using an appropriate no-change model as a benchmark hypothesis. In particular, the no-change model serves as a useful conservative approach for *complex and*

highly uncertain problems. The most famous application of this model is Julian Simon's 1980 bet with Paul Ehrlich on the prices of natural resources. Ehrlich claimed that resources are limited, and forecast mass starvation by the 1990s. Simon argued that the human ingenuity and effort caused resources to become more plentiful and thus cheaper. Given that trends over centuries have been consistent with these causal factors, Simon bet that real prices would not increase, and invited Ehrlich to pick resources and a time period for a bet. Ehrlich nominated five metals whose prices had been rising rapidly in recent years, and bet that their prices would be higher in 1990. Simon's no-change price forecasts were more accurate for all five metals over the ten-year period (see Tierney 1990).

Schnaars and Bavuso (1986) compared the accuracy of forecasts from the no-change model with forecasts from six full-trend extrapolation methods. These involved 180 weekly forecasts for each of fifteen economic time-series that included prices of resources, production, and indicators such as unemployment claims. On average, the no-change model yielded the most-accurate forecasts. The MAPE of forecasts from the no-change model was half that of the most complex extrapolation method tested (generalized adaptive filtering).

Consider the behavior of the stock market in the short-term. Researchers' attempts to make forecasts that beat the current market price have proven unsuccessful for those who lack inside information. Malkiel (2012) documents this phenomenon in a book first published over forty years ago and now in its tenth edition. In this case, for short-term forecasting the latest market price is a good summary of current knowledge.

The no-change model is not always conservative. There are many cases where cumulative knowledge calls for change. For example, consider that you sell baked beans and you have a tiny market share. You reduce your price by 10 percent. A no-change model would not be conservative. You should rely instead on knowledge about the price elasticity of similar products. In other words, forecasters should test alternative hypotheses, methods, and models such that a skeptical critic would not be able to point to a plausible and important alternative that was not tested.

The Relative Absolute Error (RAE) was developed to compare the accuracy of forecasts from alternative models. It is the error of a forecast from a proposed model relative to that of a forecast from a credible no-change model or other benchmark (Armstrong and Collopy 1992). Thus, a RAE less than 1 means the forecast is better than the benchmark forecasts, and a RAE greater than 1 means the forecast is worse than the benchmark forecast.

Avoid bias by obtaining signed ethics statements before and after forecasting (1.2.3)

Bias might be deliberate if the purpose of the forecasts is to serve strategic goals, such as with cost-benefit estimates for large-scale public works projects. For example, one study found that first-year demand forecasts for 62 large rail transportation projects were consistently optimistic, with a median overestimate of demand of 96 percent (Flyvbjerg 2013).

Bias is also common in business forecasting. One study analyzed more than 10,000 judgmental adjustments of quantitative model forecasts for one-step-ahead pharmaceutical sales forecasts. In 57 percent of 8,411 forecasts, the experts adjusted the forecast upwards, whereas downward adjustments occurred only 42 percent of the time. Optimism remained even after experts were informed about their bias, although the feedback decreased the rate of upward adjustments to 54 percent of 1,941 cases (Legerstee and Franses 2013).

To reduce deliberate bias, obtain signed ethics statements from all of the forecasters involved at the outset and again at the completion of a forecasting project. Ideally, these would state that the forecaster understands and will follow evidence-based forecasting procedures, and would include declarations of any actual or potential conflicts of interest. Laboratory studies have shown that when people reflect on their ethical standards, they behave more ethically (Armstrong 2010, pp. 89-94 reviews studies on this issue; also see Shu, Mazar, Gino, Ariely, and Bazerman 2012).

Provide full disclosure to encourage independent audits and replications (1.3)

Replications are fundamental to scientific progress. Audits are good practice in government and business, and might provide valuable evidence in a legal damages case. Even the possibility that a forecasting procedure might be audited or replicated is likely to encourage the forecaster to take more care to follow evidence-based procedures. To facilitate these benefits, forecasters should fully disclose the data and methods used for forecasting, and describe how they were selected.

Failures to disclose are often due to oversight, but are sometimes intentional. For example, in preparation for a presentation to a U.S. Senate Science Committee hearing, the first author requested the data used by the U.S. Fish and Wildlife Service researchers to prepare their forecasts that polar bears were endangered. The researchers refused to provide these data on the grounds that they were using them (Armstrong, Green, and Soon 2008).

Replications are important for detecting mistakes. Gardner (1984) found 23 books and articles, most of which were peer-reviewed, that included mistakes in the formula for the trend component of exponential smoothing model formulations. Gardner (1985) found mistakes in exponential smoothing programs used in two companies.

Weimann (1990) found a correlation (0.51) between comprehensive reporting of methodology (measured by the number of methodological *deficiencies* reported) and the *accuracy* of election polls. This is consistent with the notion that those who report more fully on the limitations of their methodology are more knowledgeable and careful in their forecasting procedures, and thus, their forecasts are more accurate.

Judgmental methods (2)

Judgmental forecasts are often used for important decisions such as whether to start a war, launch a new product, acquire a company, buy a house, select a CEO, get married, or stimulate the economy.

Avoid unaided judgment (2.1)

Structured judgments follow validated procedures in order to make effective use of available knowledge. In contrast, unaided judgment is not conservative because it is a product of faulty memories, inadequate mental models, and unreliable mental processing, to mention only a few of the shortcomings that act to prevent good use of knowledge. Moreover, when experts use their unaided judgment, they tend to more easily remember recent, extreme, and vivid events. As a result, they overemphasize the importance of such events when making judgmental forecasts, which leads them to overestimate change. These findings, from many years of experimental research, were supported by a study of 27,000 political and economic forecasts made over a 20-year period by 284 experts from different fields (Tetlock 2005).

Unaided judges tend to see patterns in the past and predict their persistence, despite lacking reasons for the patterns. Even forecasting experts are tempted to depart from conservatism in this way. For example, when two of the authors asked attendees at the 2012 *International Symposium on Forecasting* to forecast the annual global average temperature for the following 25 years on two 50-year charts, about half of the respondents drew zigzag lines (Green, Soon, and Armstrong 2014) probably to resemble the noise or pattern in the historical series (Harvey 1995)—a procedure that is almost certain to increase forecast error relative to a straight line.

Use alternative wording and pretest questions (2.2)

The way a question is framed can have a large effect on the answer. Hauser (1975, Chapter 15) provided examples of how the wording affects responses. One was the proportion of people who answered "yes" to alternatively worded questions about free speech in 1940. The questions and the percentage of affirmative responses are: (1) *"Do you believe in freedom of speech?"* 96 percent; (2) *"Do you believe in freedom of speech to the extent of allowing radicals to hold meetings and express their views to the community?"* 39 percent. Pose the forecasting question in a way that ensures the answer will

be unambiguous and useful to the decision maker. One way to reduce response errors is to pose the question in multiple ways, pre-test the different wordings, and then combine the responses.

Ask judges to write reasons for and against the forecast (2.3)

Ask judges to explain their forecasts in writing. This is conservative in that it encourages them to consider more information and that it also contributes to full disclosure. Asking for reasons is an important aspect and likely contributes to the accuracy of the Delphi method (discussed in guideline 2.6).

Koriat, Lichtenstein, and Fischhoff (1980) asked 73 subjects to pick the correct answer to each of ten general knowledge questions and then to judge the probability that their choice was correct. For ten further questions, the subjects were asked to make their picks and write down as many reasons for and against each pick that they could think of. Their errors were 11 percent less than when they did not provide reasons. In their second experiment, subjects predicted the correct answers to general knowledge questions and were asked to provide one reason to support their prediction (n=66), to contradict their prediction (55), or both (68). Providing a contradictory reason reduced error by 4 percent compared to providing no reason. Providing supporting reasons had only a small effect on accuracy.

Hoch (1985) asked students to predict the outcome of their job search efforts over the next nine months, particularly the timing of their first job offer, the number of job offers, and starting salaries. In general, students who wrote reasons why their desired outcome might not occur made more accurate forecasts.

Use judgmental bootstrapping (2.4)

People are often inconsistent in applying what they know about a problem. For example, they might suffer from information overload, boredom, fatigue, distraction, and forgetfulness. Judgmental bootstrapping protects against these problems by applying forecasters' implicit rules in a consistent way. Judgmental bootstrapping helps to ensure that the forecasts are more consistent with the forecasters' knowledge. In addition, the bootstrapping regression model is conservative in that it gives less weight to variables when uncertainty is high.

To use judgmental bootstrapping, develop a quantitative model to infer how an expert or group of experts makes the forecasts. To do so, first, present an expert with artificial cases in which the values of the causal factors vary independently of one another. Then, ask the expert to make forecasts for each case. Finally, estimate a simple regression model of the expert's forecasts against the variables. The key condition is that the model should not include variables whose actual causal effects are opposite to the effect expected by the experts.

Armstrong's (2001b) review found eleven studies using cross-sectional data from various fields, including personnel selection, psychology, education, and finance. The forecasts from judgmental

bootstrapping models were more accurate than those from unaided judgment in eight studies, there was no difference in two, and they were less accurate in one (in which an incorrect belief on causality was applied more consistently). Most of these studies reported accuracy in terms of correlations. One of them, however, reported an error reduction of 6.4 percent.

Use structured analogies (2.5)

A situation of interest, or target situation, is likely to turn out like analogous situations. Using evidence on behavior from analogous situations is conservative because it increases the knowledge applied to the problem.

To forecast using structured analogies, ask independent experts (e.g., 5 to 20) to identify analogous situations from the past, describe similarities and differences, rate each analogy's similarity to the current (target) situation, and then report the outcome of each. An administrator calculates a modal outcome for a set of experts by using each expert's top-rated analogy. That serves as the forecast for the target situation.

Structured analogies can provide easily understood forecasts for complex projects. For example, to forecast whether the California High Speed Rail (HSR) would cover its costs, a forecaster could ask experts to identify similar HRS systems worldwide and obtain information on their profitability. The Congressional Research Service did this and found that "Few if any HSR lines anywhere in the world have earned enough revenue to cover both their construction and operating costs, even where population density is far greater than anywhere in the United States" (Ryan and Sessions 2013).

In Jørgensen's (2004) study on forecasting the software development costs of two projects, the errors of the forecasts from two teams of experts who recalled the details of analogous projects were 82 percent smaller than the errors of top-down forecasts from five other teams of experts who did not recall the details of any analogous situation. The forecasts informed by analogies were also 54 percent smaller than the errors of seven bottom-up forecasts from seven teams of experts.

Research on structured analogies is in its infancy, but the findings of substantial improvements in accuracy for complex, uncertain situations are encouraging. In one study, eight conflict situations, including union-management disputes, corporate takeover battles, and threats of war were described to experts. Unaided expert predictions of the decisions made in these situations were little more accurate than randomly selecting from a list of feasible decisions. In contrast, by using structured analogies to obtain 97 forecasts, errors were reduced by 25 percent relative to guessing. Furthermore, the error reduction was as much as 39 percent for the 44 forecasts derived from data provided by experts who identified two or more analogies (Green and Armstrong 2007b).

Nikolopoulos, Litsa, Petropoulos, Bougioukos, and Khammash (this issue) test of a variation of the structured analogies method: structured analogies from an interacting group. The method reduced average percentage error relative to unaided judgment by 41 percent.

Combine independent forecasts from judges (2.6)

To increase the amount of information considered and to reduce the effects of biases, combine anonymous independent forecasts from judges. Judges can be a heterogeneous group who are experts about how others would behave, or a representative sample of people who can make valid predictions about how they will behave in the situation, such as in intentions surveys.

Armstrong (2001a) presented evidence from seven studies that involved combining forecasts of 4 to 79 experts. Combining forecasts reduced error by 12 percent compared to the typical expert forecast. Another study analyzed the accuracy of expert forecasts on the outcomes of the three U.S. presidential elections from 2004 to 2012. The error of the combined forecasts from 12 to 15 experts was 12 percent less than that of the forecast by the typical expert (Graefe, Armstrong, Jones, and Cuzán 2014).

Good results can be achieved by combining forecasts from eight to twelve experts whose knowledge of the problem is diverse and whose biases are likely to differ. Surprisingly, the expertise of the experts does not have to be high (Armstrong 1980; Tetlock 2005.)

The Delphi method is an established and validated structured judgmental forecasting method for combining experts' forecasts. Delphi is a multi-round survey that elicits independent and anonymous forecasts and reasons for them from a panel of experts. After each round, a summary of the forecasts and reasons is provided to the experts. The experts can then revise their own forecasts, free from group pressures, in later rounds. A review of the literature concluded that Delphi was more accurate than statistical groups (i.e., simple one-round surveys) in twelve studies and less accurate in two studies, with two ties. Compared to traditional meetings, Delphi was more accurate in five studies and less accurate in one; two studies showed no difference (Rowe and Wright 2001). Results from a laboratory experiment on estimation tasks that support these findings showed that Delphi not only was more accurate than prediction markets, it was also easier to understand (Graefe and Armstrong 2011).

Nikolopoulos, Litsa, Petropoulos, Bougioukos, and Khammash (this issue) obtained five forecasts about the outcomes of two government programs from a group of 20 experts using their unaided judgment, and from groups of experts using either semi-structured analogies or the Delphi method. The two structured approaches to combining forecasts reduced average percentage error relative to unaided judgment by 5 and 22.

Avoid combining forecasts in traditional group meetings. The risk of bias is high because group members can be reluctant to share their opinions in order to avoid conflict or ridicule. Managers often

rely on the unaided judgments of groups to make forecasts for important decisions, despite the approach's lack of predictive validity. Experimental evidence demonstrates that it is difficult to find a method that produces forecasts as inaccurate as unaided judgments from traditional group meetings (Armstrong 2006b).

Extrapolation methods (3)

Extrapolation is an inherently conservative approach to forecasting because it is based on data on past behavior. There are, however, a number of threats to conservatism from extrapolation because more is typically known about a situation than is contained in the time-series or cross-sectional data alone.

Use the longest time-series of valid and relevant data (3.1)

By selecting a particular starting point for estimating a time-series forecasting model or by selecting a specific subset of cross-sectional data, a forecaster has much influence over the resulting forecast. Such judgments allow people to make forecasts that support their prior beliefs. For example, those who believe in dangerous manmade global warming can select data to support their view, as can skeptics. Using the longest obtainable series or all obtainable cross-sectional data mitigates the problem.

Decompose by causal forces (3.2)

Causal forces that may affect a time series can be classified as growth, decay, supporting, opposing, regressing, and unknown (Armstrong and Collopy 1993). Growth, for example, means that the causal forces will lead the series to increase, irrespective of the historical trend. Ask domain experts to identify the effects of causal forces on the trend of the series to be forecast.

When forecasting a time-series that is the product of opposing causal forces such as growth *and* decay, decompose the series into the components affected by those forces and extrapolate each component separately. By doing so, the forecaster is being conservative by using knowledge about the expected trend in each component. Consider the problem of forecasting highway deaths. The number of deaths tends to increase with the number of miles driven, but to decrease as the safety of vehicles and roads improve. Because of the conflicting forces, the direction of the trend in the fatality rate is uncertain. By decomposing the problem into miles-driven-per-year and deaths-per-mile-driven, the analyst can use knowledge about the individual trends to extrapolate each component. The forecast for the total number of deaths per year is calculated as the product of the two components.

Armstrong, Collopy, and Yokum (2005) tested the value of decomposition by causal forces for twelve annual time-series for airline and automobile accidents, airline revenues, computer sales, and cigarette production. The authors expected decomposition to provide more accurate forecasts than those

from extrapolations of the global series if (1) each of the components could be forecast over a simulation period with less error than could the aggregate, or (2) the coefficient of variation about the trend line of each of the components would be less than that for the global series. They used successive updating to make 575 forecasts, some for forecast horizons from 1 to 5 years and some for horizons from 1 to 10 years. For the nine series that met one or more of the two conditions, forecasting the decomposed series separately reduced the MdRAE of the combined forecasts by 60 percent relative to forecasts from extrapolating the global series. (The original text of that paper has a typographical error as the text states the error reduction was 56 percent.)

Be conservative when forecasting trends (3.3)

Extrapolate conservatively by relying on cumulative knowledge about the trend. In many situations, conservatism calls for a reduction in the magnitude of the trend, commonly referred to as damping. This keeps the forecasts closer to the estimate of the current situation. However, damping might not be conservative if it were to lead to a substantial departure from a consistent long-term trend arising from well-supported and persistent causal forces, such as Moore's Law for improvements in computer performance. The doubling of performance roughly every two years has held up for over half a century, and there is reason to expect that the causal forces will continue to yield substantial improvements (Mollick 2006). Also, damping would not be conservative for situations in which a sharp change in causal forces has occurred, as might be caused by a substantial reduction in corporate taxes, elimination of a tariff, or introduction of a substantially improved product.

Be conservative when forecasting trends if the series is variable or unstable (3.3.1)

In a review of ten studies, damping the trend by using only statistical rules on the variability in the historical data yielded an average error reduction of about five percent (Armstrong 2006a). Improved accuracy was achieved in all but one study. In his review of research on exponential smoothing, Gardner (2006) concluded that "...it is still difficult to beat the application of a damped trend to every time series" (p. 637). Since the gains can be achieved easily and without any intervention, the adoption of the damped-trend exponential smoothing method would lead to immense savings for production and inventory control systems worldwide. Moreover, further gains in accuracy are possible by incorporating knowledge about the situation and judgment in structured ways as the following guidelines describe.

Be conservative when forecasting trends if the historical trend conflicts with causal forces (3.3.2)

If the causal forces acting on a time-series conflict with the observed trend in a time-series, a condition called a contrary series, damp the trend heavily toward the no-change forecast.

Judgment and prior knowledge should not, however, be abandoned. Causal forces may be sufficiently strong as to overwhelm a long-term trend, such as when a government decides to regulate an industry. In that case, one would expect the iron law of regulation to prevail (Armstrong and Green 2013) with consequent losses of consumer welfare as was found by Winston (2006).

To identify casual forces, ask a small group of experts (3 or more) for their assessment and adopt the majority opinion. Experts typically need only a minute or so to assess the causal forces for a given series, or for a group of related series.

Research findings to date suggest a simple guideline that works well for contrary series: *ignore trends*. Armstrong and Collopy (1993) applied this "contrary series rule" to forecasts from Holt's exponential smoothing (which ignores causal forces). They used 20 annual time-series from the M-Competition that were rated as contrary. By removing the trend term from Holt's model, the MdAPE was reduced by 18 percent for one-year-ahead forecasts, and by 40 percent for six-year-ahead forecasts. Additional testing used contrary series from four other data sets: annual data on (1) Chinese epidemics, (2) unit product sales, (3) economic and demographic variables, and (4) quarterly data on U.S. Navy personnel numbers. On average, the MdAPE for the no-trend forecasts was 17 percent less than Holt's for 943 one-step-ahead forecasts. For 723 long-range forecasts, which were 6-ahead for annual and 18-ahead for quarterly data, the error reduction averaged 43 percent over the five data sets.

Be conservative when forecasting trends if the forecast horizon is longer than the historical series (3.3.3)

Uncertainty is higher when the forecast horizon is longer than the length of the historical time-series. If making forecasts in such a situation cannot be avoided, consider (1) damping the trend toward zero as the forecast horizon increases, and (2) averaging the trend with trends from analogous series.

Wright and Stern (this issue) found that using an average of analogous sales growth trends for forecasting sales of new pharmaceutical products over their first year reduced the MAPE by 43 percent compared to forecasts from a standard marketing model, exponential-gamma, when 13 weeks of sales data were used for calibration.

U.S. Fish and Wildlife Service scientists overlooked the need for damping when they used only five years of historical data to forecast an immediate and strong reversal in the trend of the polar bear population. Moreover, they extended the forecast 50 years into the future (Armstrong, Green, and Soon 2008).

Be conservative when forecasting trends if the short- and long-term trend directions are inconsistent (3.3.4)

If the direction of the short-term trend is inconsistent with that of the long-term trend, the short-term trend should be damped towards the long-term trend as the forecast horizon lengthens. Assuming no major change in causal forces, a long-term trend represents more knowledge about the behavior of the series than does a short-term trend.

Estimate seasonal factors conservatively (3.4)

For situations clearly affected by causal factors, such as monthly sales of sunscreen or furnace oil, seasonal factors typically improve forecast accuracy. When the situation is uncertain, damp the estimated seasonal effects. Another approach is to combine the estimate of a seasonal factor with those for the time period before and the period after. One should also damp to adjust for the uncertainty regarding the causes of the seasonal factors. Still another approach is to combine the seasonal factors estimated for the series of interest with those estimated from analogous series.

Estimate seasonal factors conservatively when they vary substantially across years (3.4.1)

If estimates of the size of seasonal factors differ substantially from one year to the next, this suggests uncertainty. This might be due to shifting dates of major holidays, strikes, natural catastrophes, irregular marketing actions such as advertising or price reductions, and so on. To deal with this, damp the estimated seasonal factors or take an average based on each seasonal factor and those from the time periods immediately before and after.

Miller and Williams (2004) damped the seasonal factors for the 1,428 monthly series of the M3-Competition based on the degree of variability. Forecasts based on damped seasonal factors were more accurate for 59 to 65 percent of the series, depending on the horizon. For series where the tests of variability called for damping, MAPEs were reduced by about 4 percent.

Chen and Boylan (2008) tested the Miller and Williams damping procedures by analyzing 111 monthly series from the M-competition; the error reductions were similar to those obtained by Miller and Williams. They then damped the seasonal factors for 216 monthly series on light bulbs and again found that the two damping procedures reduced the error (symmetrical MAPE) of cumulative forecasts for horizons out to nine periods by 67 percent on average (from Chen and Boylan's Table 7).

Estimate seasonal factors conservatively when few years of data are available (3.4.2)

Lacking strong evidence on the causes of seasonality, damp seasonal factors strongly (or perhaps avoid using them) unless there are sufficient years of historical data from which to estimate

them. Chen and Boylan (2008) found that seasonal factors harmed accuracy when they were estimated from fewer than three years of data.

To compensate for a lack of information, consider estimating seasonal factors from analogous series. For example, for a new ski field development, one could combine seasonal factors from time-series on analogous fields with those from the new field. Using analogous data in that way, Withycombe (1989) reduced forecast errors in a test using 29 products from six product lines from three different companies. Combining seasonal factors across the products in each product line provided forecasts that were more accurate than those based on estimates of seasonality for the individual product in 56 percent of 289 one-month-ahead forecasts. Combining seasonal factors from analogous series reduced the mean squared error of the forecasts for each of the product lines, with error reductions ranging from 2 to 21 percent.

In an analysis of 44 series of retail sales data from a large U.K. department store chain, Bunn and Vassilopoulos (1999) found that forecasts from models that used seasonal factors estimated from analogous series were consistently more accurate than forecasts from models that used seasonal factors calculated in the traditional way from the target series data. When analogies were from the same business class as the target series the error reductions (MADs) compared to forecasts from standard seasonal adjustment were between 8 and 25 percent, depending on the model used. [

Gorr, Olligschlaeger, and Thompson (2003) combined seasonal crime rates from six precincts in Pittsburgh. The combined-seasonality forecast errors were about 8 percent smaller than the individual seasonality forecast errors.

Estimate seasonal factors conservatively when causal knowledge is weak (3.4.3)

Without prior knowledge on the causes of seasonality in the series to be forecast, seasonal factors are likely to increase forecasting error. To the extent that the causal knowledge is weak, damp the factors. If there is no causal basis, do not use seasonal factors. For example, it makes little sense to look for seasonable variations in the stock market.

Combine forecasts from alternative extrapolation methods or alternative data (3.5)

Armstrong (2001, page 428) found error reductions from combining forecasts from different extrapolation methods in five studies. The error reductions ranged from 4.3 to 24.2, with an average of 15 percent.

Analogous time-series can provide useful information for extrapolation models. The information is relevant for levels (or base rates for cross-sectional data), and for trends. For example, consider that one wishes to forecast sales of the Hyundai Genesis automobile. Rather than relying only

on the Genesis sales trend data to forecast, use the sales trend data for all luxury cars to forecast the trend, and then combine the two forecasts.

Causal methods (4)

Regression analysis is currently the most common approach for developing and estimating causal models. It is conservative in that it regresses to the mean value of the series in response to unattributed variability in the data. However, a regression model is not sufficiently conservative because it does not reflect uncertainty in predicting the causal variables, or in changes in causal relationships, and can include mistaken causality if any variable in the model correlates with important excluded variables over the estimation period. Another problem occurs when forecasters use statistical significance tests and sophisticated statistical methods to select predictor variables, a problem that is exacerbated when large databases are used. Sophisticated statistical techniques and an abundance of observations tend to seduce forecasters and their clients away from using cumulative knowledge and evidence-based forecasting procedures. In other words, they lead forecasters to ignore the Golden Rule. For a more detailed discussion of problems with using regression analysis for forecasting, see Armstrong (2012) and Soyer and Hogarth (2012).

Use prior knowledge to select variables and estimate effects (4.1)

Scientific discoveries about causality were, of course, made prior to the availability of regression analysis. For example, John Snow discovered the cause of cholera in London in the 1850s as a result of "the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data" (Freedman 1991, p. 298). Until around the late 1960s, data collection and statistical analyses remained expensive, and forecasters were also expected to develop their models using *a priori* analyses.

Nowadays, economists and other social scientists concerned with measuring relationships use elasticities to summarize prior knowledge. Elasticities are unit-free and easy to understand. They represent the percentage change that occurs in the variable to be forecast in response to a one-percent change in the causal variable. For example, a price elasticity of demand of -1.5 would mean that if the price increased by 10 percent, unit sales would go down by 15 percent. Forecasters can examine prior research in order to estimate elasticities and their plausible lower and upper bounds. For example, in forecasting sales, one can find income, price, and advertising elasticities for various product types in published meta-analyses. If little prior research exists, obtain estimates by surveying domain experts.

Armstrong (1970) tested the value of an *a priori* analysis by forecasting international camera sales. A fully specified model was developed from prior knowledge about causal relationships before analyzing data. Data from 1960 to 1965 for 17 countries were then used to estimate regression model

coefficients. The final model coefficients were calculated as an average of the *a priori* estimates and regression coefficients, a process later referred to as a poor man's Bayesian regression analysis. To test the predictive value of the approach, the updated model was used to backcast 1954's camera sales. Compared to a benchmark model with only statistically estimated coefficients, the model that included *a priori* knowledge reduced the MAPE by 23 percent. Another test estimated models using 1960-1965 data for 19 countries that were then used to predict market size in 11 holdout countries. The models that used *a priori* knowledge in estimating coefficients reduced forecast MAPE by 40 percent.

A priori analyses are time consuming, expensive and difficult, as they require considerable effort and good judgment by people with expertise in the field. Perhaps unsurprisingly, then, over the past half-century, forecasters have looked to sophisticated statistical procedures such as stepwise regression and data mining along with large databases and high-speed computers in the hope that these would replace the need for *a priori* analyses.

While leading econometricians have expressed support for the belief that complex statistical procedures yield greater forecast accuracy (see, for example, a survey by Armstrong 1978), a number of researchers have been skeptical of this trend. In his examination of four complex analytical techniques—automatic interaction detection, multiple regression analysis, factor analysis, and nonmetric multidimensional scaling—Einhorn (1972) concluded, "Just as the alchemists were not successful in turning base metal into gold, the modern researcher cannot rely on the 'computer' to turn his data into meaningful and valuable scientific information" (p. 378). Research since then supports Einhorn's assessment (Armstrong 2012).

Estimate variable weights conservatively (4.2)

Damping is often useful for making causal model forecasts more conservative. One strategy is to damp estimates of each variable's coefficient (weight) toward zero, a process also referred to as shrinkage. Shrinkage reduces the amount of change that a model will predict in response to changes in the causal variables, and is thus conservative when predicting change. A related strategy is to adjust the weights of the variables so that they are more equal with one another. To do this, express the variables as differences from their mean divided by their standard deviation (i.e., as normalized variables), estimate the model, and then adjust the estimated coefficients toward equality. When uncertainty about relative effect sizes is high, consider assigning equal weights to all normalized variables.

As summarized by Graefe (this issue), much experimental evidence since the 1970s has found that equal-weights models often provide more accurate *ex ante* forecasts than those from regression models. That paper also provides evidence for U.S. presidential election forecasting. Equal-weights variants of nine established regression models yielded forecasts that were more accurate for six of the

nine models. On average, the equal-weights model reduced the MAE of the original regression models by 5 percent.

Use all important variables (4.3)

When estimating relationships using non-experimental data, regression models can properly include only a subset of variables—typically about three—no matter the sample size. However, many practical problems involve more than three important variables. For example, the long-run economic growth rates of nations are affected by more than fifty important variables. In addition, many causal variables may not vary over historical periods, so regression models cannot provide estimates of the causal relationships for these variables.

Index methods allow for the inclusion of all important knowledge about causal relationships into a single model. The approach draws on an insight from Benjamin Franklin's "method for deciding doubtful matters" (Sparks 1844). Franklin suggested listing all relevant variables, identifying their directional effect, and weighting them by importance. Index models might also be called *knowledge models*, because they can represent all knowledge about factors affecting the thing being forecast.

To develop an index model, use prior knowledge to identify all relevant variables and their expected directional influence on whatever is being forecast (e.g., job performance). Ideally one should develop an index model based on knowledge gained by reviewing experimental studies. In fields where experimental studies are scarce, survey experts with diverse knowledge and hypotheses. Calculate an index score by determining the values of variables for a situation of interest and then add the values. This can be done by simply assigning equal weights to all variables, but consider using different weights for the variables if there is strong prior evidence that the variables have differential effects. The index score is then used to calculate the forecast. For selection problems, the option with the highest score is favored. For numerical forecasts, use a simple linear regression model to estimate the relationship between the index score and the variable to be predicted (e.g., sales of a new movie).

The index method has been used to forecast U.S. presidential elections, a situation with knowledge about a large number of causal variables. An index model based on 59 biographical variables correctly predicted the winners in 28 of 30 U.S. presidential elections up through 2012 (Armstrong and Graefe 2011). Another index model was based on surveys of how voters expected U.S. presidential candidates to handle up to 47 important issues. The model correctly predicted the election winner in ten of the eleven elections up to 2012 (Graefe and Armstrong 2013). Another study (Graefe, this issue) created an index model by adding the standardized values of all 29 variables that were used by nine established U.S. presidential election forecasting models. Across the ten elections to 2012, the forecast

error of this index model was 48 percent lower than the error of the typical individual regression model and 29 percent lower than the error of the most accurate individual model.

A recent study develops an index model to predict the effectiveness of advertisements based on the use of up to 195 evidence-based persuasion principles. Advertising novices were asked to rate the how effectively each relevant principle was applied for each ad in 96 pairs of print ads. The ad with the highest index score was predicted to be the most effective. The index-score predictions were compared to the advertising experts' unaided judgments. Expert unaided judgment is the typical approach for such forecasts. The experts were correct for 55 percent of the pairs whereas the index scores were correct for 75 percent, an error reduction of 43 percent (Armstrong, Du, Green, Graefe, and House, 2014).

Combine models that use different information and procedures (4.4)

One way to deal with the limitations of regression analysis is to develop different models with different variables and data, and to then combine the forecasts from each model. In a study on 10-year-ahead forecasts of population in 100 counties of North Carolina, the average MAPE for a set of econometric models was 9.5 percent. In contrast, the MAPE for the combined forecasts was only 5.8 percent, an error reduction of 39 percent (Namboodiri and Lalu 1971). Armstrong (2001a, page 428) found error reductions from combining forecasts from different causal models in three studies. The error reductions were 3.4 percent for GNP forecasts, 9.4 percent for rainfall runoff forecasts, and 21 percent plant and equipment.

Another test involved forecasting U.S. presidential election results. Most of the well-known regression models for this task are based on a measure of the incumbent's performance in handling the economy and one or two other variables. The models differ in the variables and in the data used. Across the six elections from 1992 to 2012, the combined forecasts from all of the published models in each year—the number of which increased from 6 to 22 across the six elections—had a mean absolute error that was 30 percent less than that of the typical model (Graefe, Armstrong, Jones Jr., and Cuzán 2014).

Combine forecasts from diverse evidence-based methods (5)

Combining forecasts from evidence-based methods is conservative in that more knowledge and data are used, and the effects of biases and mistakes such as data errors, computational errors, and poor model specification are likely to offset one another. Consequently, combining forecasts reduces the likelihood of large errors. Equally weighting component forecasts is conservative in the absence of strong evidence on large differences in out-of-sample forecast accuracy from different methods.

Interestingly, the benefits of combining are not intuitively obvious. In a series of experiments with highly qualified MBA students, a majority of participants thought that averaging estimates would deliver only average performance (Larrick and Soll 2006).

Most studies on the value of combining have used equal weights. A meta-analysis by Armstrong (2001a p. 428) found 11 studies on the error reductions of combining forecasts from different methods. On average, the errors of the combined forecasts were 11.5 percent lower than the average error of the component forecasts. More recent research on U.S. presidential election forecasting (Graefe, Armstrong, Jones, and Cuzán, 2014) finds much larger gains when forecasts are combined from different evidence-based methods that draw upon different data. Averaging forecasts within and across four established election-forecasting methods (polls, prediction markets, expert judgments, and regression models) yielded forecasts that were more accurate than those from each of the component methods. Across six elections, the average error reduction compared to the typical component method forecast error was 60 percent.

Many scholars have proposed methods for how to best weight the component forecasts. However, Clemen's (1989) review of over 200 published papers from the fields of forecasting, psychology, statistics, and management science concluded that using equal weights often provides the best forecast when combining. Mancuso and Werner's (2013) update of Clemen's review covering 174 articles reinforces his conclusion.

If evidence suggests that some methods provide more accurate forecasts than others for the given situation, specify the combining procedure (i.e., the weights on the component forecasts) prior to making the forecasts. Doing so will reduce the effects of any biases. One method, rule-based forecasting, uses prior evidence on the relative accuracy of forecasts from different methods under different conditions. For example, it varies the weights on extrapolation forecasts based on the horizon, causal forces, and variability of the historical data. Rule-based forecasting provided the most accurate forecasts for annual data in the M-Competition. There was a reduction in the MdAPE of 18 percent for one-year ahead forecasts compared to that for the equal-weights combined forecast. For six-year ahead forecasts, the error reduction was 42 percent (Collopy and Armstrong 1992b). Vokurka, Flores and Pearce (1996) provide additional support for differential weights in rule-based forecasting. They used automatic rule selection and found that errors for 6-year-ahead forecasts of M-Competiton data were 15 percent less than those for the equal-weights combined forecasts.

The evidence for differential weights must be strong, however, when the weights are estimated from data rather than based on prior knowledge. For example, as summarized by Graefe, Küchenhoff, Stierle and Riedl (2014), 2 of 3 studies on economic forecasting found that simple averages provided more accurate forecasts than Bayesian combining methods while one study provided mixed evidence. Their study also provides new evidence for U.S. presidential election forecasting, where the error of the simple average forecasts were 25 percent less than the error of the Bayesian Model Averaging forecasts. A study that tested the range of theoretically possible combinations found that easily understood and

implemented heuristics, such as take-the-average, will in most situations, perform as well as the rather complex Bayesian approach (Goodwin, this issue).

Avoid unstructured judgmental adjustments to forecasts (6)

Judgmental adjustments can lead to a loss of objectivity and introduce biases and random errors. For example, a survey of 45 managers in a large conglomerate found that 64 percent of them believed that "forecasts are frequently politically motivated" (Fildes and Hastings 1994).

Unfortunately, forecasters and managers are often tempted to make unstructured adjustments to forecasts from quantitative methods. One study found that 91 percent of more than 60,000 statistical forecasts made in four companies were judgmentally adjusted (Fildes, Goodwin, Lawrence, and Nikolopoulos 2009). Consistent with this finding, a survey of forecasters at 96 U.S. corporations found that about 45 percent of the respondents claimed that they always made judgmental adjustments to statistical forecasts, while only 9 percent said that they never did (Sanders and Manrodt 1994). Legerstee and Franses (2014) found that 99.7 percent of 8,411 one-step-ahead sales forecasts for pharmaceutical products made by 21 experts in 21 countries were adjusted. Providing experts with feedback on the harmful effects of their adjustments only slightly reduced the rate of adjustments (to 98.4 percent).

Most forecasting practitioners expect that judgmental adjustments will lead to error reductions of between five and ten percent (Fildes and Goodwin 2007). Yet little evidence supports that belief. For example, Franses and Legerstee (2010) analyze the relative accuracy of original model forecasts and expert adjusted forecasts for 194 combinations of one-step-ahead forecasts in 35 countries and across 7 pharmaceutical product categories. On average, the adjusted forecasts were less accurate than the model forecasts in 57 percent of the 194 country-category combinations.

In psychology, extensive research on cross-sectional data led to the conclusion that one should not make subjective adjustments to forecasts from a quantitative model. For example, a summary of research on personnel selection revealed that employers should rely on forecasts from validated statistical models. They should *not* meet job candidates, because doing so leads them to adjust the forecasts to the detriment of accuracy (Meehl 1954).

Adjustments that follow structured procedures are less harmful. In an experiment by Goodwin (2000), 48 subjects made adjustments to one-period ahead statistical sales forecasts. When no specific instructions were provided, subjects adjusted 85 percent of the statistical forecasts; the revised forecasts had a median absolute percentage error (MdAPE) of 10 percent. In comparison, when subjects were asked to justify any adjustments by picking a reason from a pre-specified list, they adjusted only 35 percent of the forecasts. The MdAPE was 3.6 percent and thus 64 percent less than the error of the unstructured adjustment. In both cases, however, the judgmental adjustments yielded less accurate

forecasts than the original statistical forecasts, which had a MdAPE of 2.8 percent. In other words, the unadjusted statistical forecasts provided an error reduction of 72 percent relative to the more expensive approach of making adjustments, and 22 percent even when structured adjustments were used.

Adjustments should only be considered when the conditions for successful adjustment are met and when bias can be avoided (Goodwin and Fildes 1999; Fildes, Goodwin, Lawrence, and Nikolopoulos 2009). Judgmental adjustments of forecasts are best confined to experts' estimates of the effects of important influences not included in the forecasting model (Sanders and Ritzman 2001) such as when experts have good knowledge of the effects of special events and changes in causal forces (Fildes and Goodwin 2007). The estimates should be made in ignorance of the model forecasts, but with knowledge of what method and information the model is based upon (Armstrong and Collopy 1998; Armstrong, Adya, and Collopy 2001). The experts' estimates should be derived in a structured way (Armstrong and Collopy 1998), and the rationale and process documented and disclosed (Goodwin 2000). In practice, documentation of the reasons for adjustments is uncommon (Fildes and Goodwin 2007). The final forecasts should be composed from the model forecasts and the experts' adjustments.

Discussion

The Golden Rule provides a unifying theory of forecasting: Be conservative by adhering to cumulative knowledge. The theory is easy to understand and provides the basis for a checklist to help forecasters and decision makers.

Checklists have been shown to be of enormous value as a tool to help practitioners and decision-makers follow standard practice and, even better, evidence-based guidelines. Checklists are useful because unaided human brains are maladapted for solving complex problems with many variables. Think of operating a nuclear power plant, flying an airplane, or drafting a regulation.

Arkes, Shaffer, and Dawes (2006) provide a review of evidence on the efficacy of checklists. For example, an experiment on avoiding infection in intensive care units of 103 Michigan hospitals required physicians to follow five rules when inserting catheters: (1) wash hands, (2) clean the patient's skin, (3) use full-barrier precautions when inserting central venous catheters, (4) avoid the femoral site, and (5) remove unnecessary catheters. Adhering to this simple checklist reduced the median infection rate from 2.7 per 1,000 patients to zero after three months. Benefits persisted sixteen to eighteen months after the checklist was introduced, as infection rates decreased by 66 percent (Pronovost et al. 2006). Another study reports on the application of a 19-item checklist to surgical procedures on thousands of patients in eight hospitals in cities around the world. Following the introduction of the checklist, death rates declined by almost half (from 1.5 to 0.8 percent), and complications declined by over one-third

(from 11 to 7 percent) (Haynes, Weiser, Berry, Lipsitz, Breizat, and Dellinger 2009). Gawande (2010) provides further evidence of the usefulness of checklists in medicine, aviation, finance, and other fields.

Forecasting experts apparently concur with the guidelines of the Golden Rule Checklist. In a survey of forecasting experts conducted while this paper was being written, most respondents stated that they typically follow or would consider following all but three of the guidelines. The guidelines that most experts disagreed with were 1.2.1/1.2.2 (which were originally formulated as one guideline: "specify multiple hypotheses or conceal the purpose of the forecast") and 2.6 ("use structured analogies"). The survey questionnaire and responses are available at goldenruleofforecasting.com.

The Checklist items were derived from evidence from forecasting research. Exhibit 2 summarizes the evidence to date. All the evidence is consistent with the guidelines provided in the Checklist, and the gains in accuracy are large on average. (Details on how these improvements were assessed are provided in the Spreadsheet "Error reductions for Golden Rule Guidelines" in the Research Repository at ForPrin.com.)

Exhibit 2: Evidence on the 28 Golden Rule Guidelines

Evidence available on	21
Effect size reported	20
More than one effect size comparison	15
Average error reduction	28%
Range of error reductions	2.% to 82%

There are gaps in the evidence. For example, no evidence was found for seven of the guidelines, and four guidelines were based on only single comparisons. It was difficult to track down relevant studies, so there are likely to be more than the 150 experimental comparisons identified in this paper. Surely, then, new or improved ways of being conservative will be found and the improvements will be made in how and when to apply the guidelines.

Current forecasting practice

Great advances have been made in the development and validation of useful forecasting procedures over the past century. This is evident, for example, in the astonishing improvements summarized in Exhibit 2. However, it is difficult to find evidence that forecasting has improved *in practice*. Ascher (1978) concluded that forecast accuracy had not improved over time in his review of forecasting for population, economics, energy, transportation, and technology. In his review of the research on agriculture forecasting, Allen (1994) was unable to find evidence that forecasting practice in

economics had improved over time; He then compared the accuracy of forecasts from 12 studies (22 series) before 1985 and 11 studies after 1985, finding only trivial differences in accuracy. Based on her review of 25 years of population forecasting research, Booth (2006) finds no evidence that the accuracy of population forecasts has improved over time. When McCarthy, Davis, Golicic and Mentzer (2006) replicated two surveys on sales forecast accuracy conducted 20 years earlier, they concluded that accuracy had fallen.

There are a number of reasons why forecasting practice has not improved. Originally, this might have been due to ignorance of evidence-based forecasting procedures. For the past 15 years, evidence-based procedures have been readily available, and at no cost, at ForPrin.com. Yet, the ignorance remains. One important reason is that forecasters are biased to satisfy the client. Another reason, which emerged decades ago, is the reliance of forecasters on big data and sophisticated statistical procedures. Given the lack of experimental evidence to support this movement, the authors are reminded of rain dancers and the belief in magic. The Golden Rule is proposed as an aid for overcoming these remaining obstacles to improving forecasting practice.

How to use the Golden Rule to improve forecasting practice

The Golden Rule Checklist was developed to improve forecasting practice. The primary way for forecasters to use the Checklist is to help them derive their forecasts. Forecasting audits can help to ensure that this is done. Forecasting software providers could help to encourage the adoption of the Checklist by implementing the guidelines in their products.

Use the Golden Rule Checklist to audit forecasting reports

Ideally, forecasting audits involve two or more experts who were not part of the team that prepared the forecast and who have no biases about the subject of the forecast. The Golden Rule Checklist provides an evidence-based standard against which forecasting procedures can be examined. Using it requires little training. Intelligent people who have no background in forecasting but who have read this paper can use the Golden Rule Checklist. With about two hours of preparation, analysts should be able to conduct audits that would help them to guard against inaccurate forecasts. Such audits can be done at little cost. A person who is familiar with a forecasting report can quickly assess which Golden Rule guidelines are relevant to the forecasting task and whether the forecasters followed them. If the description of the forecasting procedure in the report is inadequate for assessing whether the guidelines were followed, be conservative and ignore the report and its forecasts.

In a test of the Checklist, two of the authors (Green and Armstrong) audited the forecasting procedures used for the IPCC's Fourth Assessment Report (Randall et al. 2007). Given their familiarity

with the report, it took them each only ten minutes to do audits. They agreed on the ratings, so no time was needed to resolve differences. They concurred that 25 of the 28 guidelines were relevant and that all of the 25 relevant guidelines were violated. Not surprisingly, then, Green, Armstrong, and Soon's (2009) validation study of the IPCC global warming projections found the error for long-term forecasts (those for 91 to 100 years into the future) was 12 times larger than the no-change forecast. The no-change forecast was conservative for this problem due to the lack of strong evidence for the dangerous long-term global warming hypothesis (see, e.g., Idso, Carter, and Singer, 2013) and the very long history of trend reversals on all time scales.

Implement conservative guidelines as defaults in forecasting software

Software providers could implement the Golden Rule Checklist as defaults in forecasting software. In case of resistance by buyers, providers could allow their users to opt out if they do not want to use a default. For example, it would be a simple and inexpensive matter to include the contrary-series rule (3.3.2) and to avoid using seasonal factors if there are fewer than 3 years of data (3.4.2). Another simple procedure would be to combine a forecast with an appropriate no-change forecast for uncertain and complex situations. More weight should be placed on the no-change model when uncertainty is high. Uncertainty typically increases with the complexity of the problem and the length of the forecast horizon.

Forecasting software providers may be unaware of the latest experimental evidence on the accuracy of forecasting methods. Therefore, the clients might need to request implementation of the Golden Rule Checklist. Clients could simply provide the Checklist to their software providers and ask.

Hold forecasters to account when they fail to follow conservative guidelines

When bad outcomes occur in medicine and engineering, doctors and engineers are often sued because they failed to follow proper evidence-based procedures. Should this recourse also be available for forecasting? To increase the chances of obtaining valid and useful forecasts, clients could insist that forecasters use the evidence-based Golden Rule Checklist, and require that they sign a document to certify that they did so.

An expectation of perfect forecast accuracy is unreasonable. Perhaps as a consequence, there have been few lawsuits claiming damages arising from poor forecasts. In these few cases, the plaintiffs almost always failed. A recent Italian lawsuit against seismologists' non-prediction of an earthquake is an exception, but the case may yet be overturned. Stronger cases for damages could, however, be made by showing that forecasters' practices did not follow evidence-based guidelines. The Golden Rule Checklist could provide the basis for such cases. One would hope that this possibility would motivate

forecasters to use conservative forecasting procedures. To ensure objectivity, forecasters would be advised to use independent auditors.

Conclusions

The first paragraph of this paper asked how a decision maker should evaluate a forecast. The answer is to assess whether the forecasting process followed the Golden Rule Checklist (Exhibit 1). Following the guidelines in the Golden Rule Checklist improved forecast accuracy substantially and consistently no matter what was being forecast, how the guidelines were applied, how many guidelines were used, how long the forecast horizon, how much data were available, how good the data, or what criteria were used for accuracy. The error reductions, based on 115 experimental comparisons, ranged from 2 to 82 percent. Moreover, using the Golden Rule is likely to reduce the risk of large errors.

The evidence-based Golden Rule Checklist presented in this article provides simple and easily understood guidance on how to make conservative forecasts. The guidelines enable non-forecasters to judge the value of forecasts by assessing the validity of the forecasting process that gave rise to them. That assistance is especially important in situations in which non-forecasters are likely to be intimidated by experts. One such situation is when experts claim that things are different now.

The Golden Rule makes scientific forecasting comprehensible and accessible to all. It can be readily understood, and those who use the checklist can easily spot violations. Following the Golden Rule improves the accuracy of forecasts substantially, which helps decision makers to make better decisions. The Golden Rule faces the traditional enemies of evidence-based forecasting: politics and the belief in magic.

Acknowledgments

Fred Collopy, Jason Dana, Peter Fader, Robert Fildes, Everette Gardner, Paul Goodwin, Nigel Harvey, Robin Hogarth, Michael Lawrence, Mike Metcalf, Don Peters, Fotios Petropoulos, Steven P. Schnaars, and Eric Stellwagen provided reviews. Kesten Green presented a version of the paper at the University of South Australia in May 2013 and at the International Symposium on Forecasting in Seoul in June 2013. Geoff Allen, Hal Arkes, Bill Ascher, Bob Clemen, Shantayanan Devarajan, Magne Jørgensen, Geoffrey Kabat, Peter Pronovost, Lisa Shu, Jean Whitmore, and Clifford Winston made suggestions for improvements. We thank the many authors who provided suggestions on our summaries of their research. Hester Green and Jennifer Kwok edited the paper. Responsibility for any errors remains with the authors.

References

Allen, P. G. (1994). Economic forecasting in agriculture. *International Journal of Forecasting, 10*(1), 81–135.

Arkes, H. R., Shaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making, 19*(5), 429–439.

Armstrong, J. S. (1970). An application of econometric models to international marketing. *Journal of Marketing Research, 7*(2), 190–198.

Armstrong, J. S. (1978), Forecasting with econometric methods, *Journal of Business*, *51* (4), 549–564.

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

Armstrong, J. S. (1985). *Long-range Forecasting: From Crystal Ball to Computer*. New York: Wiley.

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 417–439). New York: Springer.

Armstrong, J. S. (2001b). Judgmental bootstrapping: Inferring experts' rules for forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 171–192). New York: Springer.

Armstrong, J. S. (2001c). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. New York: Springer.

Armstrong, J. S. (2006a). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting, 22*(3), 583–598.

Armstrong, J. S. (2006b). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting, 5*(2006), 3–8.

Armstrong, J. S. (2010). *Persuasive Advertising*. New York: Palgrave MacMillan.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting, 28*(3), 689–694.

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: A Handbook for Researchers and Practitioners* (pp. 259–282). New York: Springer.

Armstrong, J. S., & Andress, J. G. (1970). Exploratory Analysis of Marketing Data: Trees vs. Regression, *Journal of Marketing Research, 7,* 487–492.

Armstrong, J. S., & Collopy, F. (1992). Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, 8, 69–80.

Armstrong, J. S., & Collopy, F. (1993). Causal forces: Structuring knowledge for time-series extrapolation. *Journal of Forecasting, 12*(2), 103–115.

Armstrong, J. S. & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: Principles from empirical research. In G. Wright & P. Goodwin (Eds.), *Forecasting with Judgment* (pp.263–393). Chichester: Wiley.

Armstrong, J. S., Collopy, F., & Yokum, J. T. (2005). Decomposition by causal forces: a procedure for forecasting complex time series. *International Journal of Forecasting, 21*(1), 25–36.

Armstrong, J. S., & Graefe, A. (2011). Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research, 64*(7), 699–706.

Armstrong, J. S. & Green, K. C. (2013). Effects of corporate social responsibility and irresponsibility policies: Conclusions from evidence-based research. *Journal of Business Research*, *66*, 1922–1927.

Armstrong, J. S., Green, K. C., & Soon, W. (2008). Polar bear population forecasts: A public-policy forecasting audit. *Interfaces, 38*(5), 382–405.

Armstrong, J. S., Du, R., Green, K. C., Graefe, A., & House, A. (2014). Predictive validity of evidence-based advertising principles. Annual Meeting of the International Communication Association, May 2014, Seattle. Available at https://marketing.wharton.upenn.edu/files/?whdmsaction=public:main.file&fileID=6794

Ascher, W. (1978). *Forecasting: An Appraisal for Policy-makers and Planners*. Baltimore: The Johns Hopkins University Press.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting* 22 (3), 547-581.

Bunn, D. W., & Vassilopoulos, A. I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting, 15*(4), 431–443.

Carson, R. T., Cenesizoglu, T., & Parker, R. (2011). Forecasting (aggregate) demand for US commercial air travel. *International Journal of Forecasting,* 27, 923–94.

Chamberlin, T. C. (1890, 1965). The method of multiple working hypotheses. *Science,* 148, 754–759. (Reprint of an 1890 paper).

Chen, H. & Boylan, J. E. (2008). Empirical evidence on individual, group and shrinkage indices. *International Journal of Forecasting*, *24*, 525–543.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559–583.

Collopy, F., & Armstrong, J. S. (1992a). Expert opinions about extrapolation and the mystery of the overlooked discontinuities. *International Journal of Forecasting, 8*(4), 575–582.

Collopy, F., & Armstrong, J. S. (1992b). Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science, 38*(10), 1394–1414.

Dangerfield, B. J., & Morris, J. S. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting, 8*(2), 233–241.

Devarajan, S. (2013). Africa's statistical tragedy. *The Review of Income and Wealth*, 59, Special Issue, S9–S15.

Einhorn, H. J. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly, 36*(3), 367–378.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37*(6), 570–576.

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*(1), 3–23.

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

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review / Revue Internationale de Statistique, 63*(3), 289–308.

Flores, B. E, & Whybark, C. D. (1986). A comparison of focus forecasting with averaging and exponential smoothing. *Production and Inventory Management*, *27*(3), 96–103.

Flyvbjerg, B. (2013). Quality control and due diligence in project management: Getting decisions right by taking the outside view. *International Journal of Project Management, 31*(5), 760–774.

Franses, P. H. & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *29*(3), 331–340.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology, 21*(1), 201–313.

Gardner, E. S., Jr. (1984). The strange case of the lagging forecasts. *Interfaces*. *14*(3), 47–50.

Gardner, E. S., Jr. (1985). Further notes on lagging forecasts, *Interfaces*. *15*(5), 63.

Gardner, E. S. Jr. (2006). Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting*, *22*, 637–666.

Gardner, E. S. Jr. & Anderson E. A. (1997). Focus forecasting reconsidered, *International Journal of Forecasting*, *13*(4), 501–508.

Gardner, E. S. Jr,. Anderson-Fletcher, E. A., & Wickes, A. M. (2001). Further results on focus forecasting vs. exponential smoothing. *International Journal of Forecasting*, *17*(2), 287–293.

Gawande, A. (2010). *The Checklist Manifesto: How to Get Things Right*. New York: Metropolitan Books.

Goodwin, P. (this issue). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. *Journal of Business Research*, XXXX.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, *16*(1), 85–99.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*(1), 37–53.

Goodwin, P. & Meeran, S. (2012) Robust testing of the utility-based high-technology product sales forecasting methods proposed by Decker and Gnibba-Yukawa (2010). *Journal of Product Innovation Management*, *29*(S1), 211–218.

Gorr, W., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting, 19*(4), 579–594.

Graefe, A. (this issue). Improving forecasts using equally weighted predictors. *Journal of Business Research*, XXXX.

Graefe, A., & Armstrong, J. S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, 27(1), 183–195.

Graefe, A., & Armstrong, J. S. (2013). Forecasting elections from voters' perceptions of candidates' ability to handle issues. *Journal of Behavioral Decision Making*, *26*(3), 295–303.

Graefe, A., Küchenhoff, H., Stierle, V. & Riedl, B. (2014). Conditions of Ensemble Bayesian Model Averaging for political forecasting, Working paper, Available at: http://ssrn.com/abstract=2266307.

Graefe, A., Armstrong, J. S., Jones Jr., R. J., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43–54.

Green, K. C., & Armstrong, J. S. (2007a). Global warming: Forecasts by scientists versus scientific forecasts. *Energy & Environment, 18*(7-8), 997–1021.

Green, K. C., & Armstrong, J. S. (2007b). Structured analogies for forecasting. *International Journal of Forecasting, 23*(3), 365–376.

Green, K. C., Armstrong, J. S., & Soon, W. (2009). Validity of climate change forecasting for public policy decision making. *International Journal of Forecasting, 25*(4), 826–832.

Green, K. C., Soon, W., & Armstrong, J. S. (2014). Evidence-based forecasting for climate change. *Working paper*.

Harvey, N. (1995). Why are judgments less consistent in lss predictable task situations. *Organizational Behavior and Human Decision Processes*, *63*, 247–263.

Hauser, P. M. (1975). *Social Statistics in Use*. New York: Russell Sage.

Haynes, A. B., Weiser, T. G., Berry, W. R., Lipsitz, S. R., Breizat, A. H. S., & Dellinger, E. P., in Lapitan, M. C. M. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine, 360*(5), 491–499.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 719–731.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*(1), 40–46.

Idso, C. D., Carter, R. M. & Singer, S. F. (2013). *Climate Change Reconsidered II: Physical Science*. Chicago, Il: The Heartland Institute.

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

Kabat, G. C. (2008). *Hyping Health Risks.* New York: Columbia University Press.

Kinney, W. R., Jr. (1971). Predicting earnings: Entity versus subentity data. *Journal of Accounting Research*, *9*, 127–136.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118.

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

Legerstee, R. & Franses, P. H. (2014), Do experts' SKU forecasts improve after feedback? *Journal of Forecasting*, *33*, 66-79.

MacGregor, D. (2001). Decomposition for judgmental forecasting and estimation. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 107–123). New York: Springer.

Makridakis, S., 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*(2), 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting, 9*(1), 5–22.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting, 16*(4), 451–476.

Malkiel. B. G. (2012)**.** *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing (Tenth Edition).* New York: W.W. Norton.

Mancuso, A. C. B., & Werner, L. (2013). Review of combining forecasts approaches. *Independent Journal of Management & Production*, *4*(1), 248–277.

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

Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.

Miller, D. M., & Williams, D. (2004). Damping seasonal factors: Shrinkage estimators for the X-12-ARIMA program. *International Journal of Forecasting. 20*(4), 529-549. (Published with commentary, pp 551–568).

Mollick, E. (2006). Establishing Moore's Law. IEEE *Annals of the History of Computing, 28*, 62-75.

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 estimation: A case study. *Rural Sociology, 36*, 187–194.

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukosa, V., & Khammash, M. (2014). Relative performance of methods for forecasting special events. *Journal of Business Research*, this issue, XXXX.

Peacock, E., Taylor, M. K., Laake, J., Stirling, I. (2013). Population ecology of polar bears in Davis Strait, Canada and Greenland. *The Journal of Wildlife Management*, 77, 463–476.

Prasad, V., Vandross, A., Toomey, C., Cheung, M., Rho, J., Quinn, S., Chako, S.J., Borkar, D., Gall, V., Selvaraj, S., Ho, N., & Cifu, A. (2013). A decade of reversal: An analysis of 146 contradicted medical practices. *MayoClinicProceedings.org*, 790–798. Available from http://www.senyt.dk/bilag/artiklenframayoclinicproce.pdf

Pronovost, P., Needham, D., Berenholtz, S., Sinopoli, D., Chu, H., Cosgrove, S., Sexton, B., Hyzy, R., Welsh, R., Roth, G., Bander, J., Kepros, J., & Goeschel, C. (2006). An intervention to decrease catheter-related bloodstream infections in the ICU. *New England Journal of Medicine*, *355*, 2725–2732.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A., & Taylor, K.E. (2007). Climate Models and Their Evaluation. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, & H. L. Miller (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 589–662). Cambridge, UK and New York, USA: Cambridge University Press.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 125–144). New York: Springer.

Runkle, D. E. (1998). Revisionist history: how data revisions distort economic policy research. *Federal Reserve Bank of Minneapolis Quarterly Review, 22*(4), 3–12.

Ryan, P., & Session, J. (2013). Sessions, Ryan Call For Halt On Taxpayer Funding For Risky High-Speed Rail Project. *U.S. Senate Budget Committee,* Available from http://www.budget.senate.gov/republican/public/index.cfm/2013/3/sessions-ryan-call-for-halt-on-taxpayer-funding-for-risky-high-speed-rail-project.

Sanders, N. R., & Manrodt, K. B. (1994). Forecasting practices in US corporations: Survey results. *Interfaces, 24*(2), 92–100.

Sanders N. R., & Ritzman L. P. (2001). Judgmental adjustment of statistical forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 405–416). New York: Springer.

Schnaars, S. P. (1989). *Megamistakes: Forecasting and the Myth of Rapid Technological Change*. The Free Press: New York.

Schnaars, S. P., & Bavuso, R. J. (1986). Extrapolation models on very short-term forecasts. *Journal of Business Research, 14*(1), 27–36.

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the begining makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences*, *109*(38), 15197–15200.

Soyer, E., & Hogarth, R. M. (2012). Illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting, 28*(3), 695–711.

Sparks, J. (1844). *The Works of Benjamin Franklin* (Vol. 8). Boston: Charles Tappan Publisher.

Tessier, T. H., & Armstrong, J. S. (2014). Decomposition of Time-Series Forecasts by Current Status and Change: Effects on Accuracy. Working Paper (provide URL)

Tetlock, P. C. (2005). *Expert political judgment*. Princeton: Princeton University Press.

Tierney, J. (1990). Betting on the planet. *New York Times*. Available from http://www.nytimes.com/1990/12/02/magazine/betting-on-theplanet.html?pagewanted=all&src=pm.

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

Weimann, G. (1990). The obsession to forecast: Pre-election polls in the Israeli press. *Public Opinion Quarterly, 54*, 396–408.

Winston, C. (2006). *Government Failure versus Market Failure: Microeconomics Policy Research and Government Performance.* Washington, D.C.: AEI-Brookings Joint Center for Regulatory Studies. Available from http://www.brookings.edu/press/Books/2006/governmentfailurevsmarketfailure.aspx.

Withycombe, R. (1989). Forecasting with combined seasonal indices. *International Journal of Forecasting*, 5, 547–552.

Wright, M., & Stern, P. (2014). Forecasting new product trial with analogous series. *Journal of Business Research*, *this issue*, XXXX.

Zarnowitz, V. (1967). An appraisal of short-term economic forecasts, *NBER Occasional Paper 104*, New York: National Bureau of Economic Research.

Word count: 15,800
Text only: 13,100