Golden Rule of Forecasting: Be Conservative

November 10, 2013

Working Paper Draft

GoldenRule 261.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

We propose the Golden Rule of Forecasting: Be Conservative when Forecasting. A conservative forecast in consistent with cumulative knowledge about the present and the past. Forecasters should incorporate all knowledge relevant to the forecasting problem and use forecasting methods that have been validated for the type of situation. Guidelines, in the form of a Golden Rule Checklist, are provided as an aid to following the Golden Rule. The guidelines are the product of a review of experimental evidence. In all of the prior studies the authors found, forecasts derived in conservative ways were more accurate than forecasts derived in less conservative ways under all conditions. Such forecasts also reduce the risk of large errors. Gains from conservatism are greater when the situation is uncertain and complex, and when bias is likely. Conservative procedures are simple to understand and to implement. Those who are not forecasting experts should be able to understand the guidelines well enough to be able to identify doubtful forecasts. The guidelines can help forecasters avoid temptations to violate the Golden Rule presented by access to large databases and complex statistical analyses. Given that all forecasting involves uncertainty, the Golden Rule is applicable to all forecasting.

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 forecasting, judgmental bootstrapping, nowcasting, regression, risk, simplicity, shrinkage, stepwise regression, structured analogies.

Introduction

Imagine that you are a manager who hires a respected consultant to predict profitable locations for stores. You provide access to large databases, and the consultant responds with a model that applies the latest statistical techniques to data on the international economy and on your company. You do not understand the procedures, but the implications of the forecasts are clear: you should make major investments in opening new outlets. The forecasts are based on statistically significant relationships identified in the data, and the consultant's story is convincing. Your colleagues are impressed by the consultant's report, and support acting on it. What should you do? To answer that question, and the general question of how to improve forecasting, this paper provides evidence-based guidelines to help forecasters make better forecasts and to help decision makers assess whether forecasts were derived from proper forecasting procedures.

We propose the *Golden Rule of Forecasting*: Be conservative when forecasting. Conservatism in forecasting is the expectation that the future will be like the past and that future changes will be circumscribed by past changes and relationships. Conservatism requires a valid and reliable assessment of the problem, use of cumulative knowledge about the current and historical situation as well as about causality, and the application of appropriate validated forecasting procedures. In short, it is necessary to know the past and present to predict the future. Claims that things are different now should be met with skepticism.

The Golden Rule is relevant to all forecasting problems. The rule is especially important when bias is likely, and when the situation is uncertain and complex. Such situations are common in public policy, as with forecasts of the economy, environmental impacts, and expenditures on welfare, transportation, education, and medicine. Situations in businesses can be complex and uncertain too, as with new product introductions. Bias can promote expenditures or pessimistic forecasts to increase the chance of good news.

The Golden Rule Checklist in Exhibit 1 can help forecasters be conservative when making forecasts. Some of the 28 guidelines in the Checklist are inherently conservative, and some encourage use of conservative procedures. To the extent possible, the guidelines are based on experimental evidence. Additional guidelines are implied by the evidence. Exhibit 1 also shows the error reductions that have been achieved by following each of the guidelines. (Some of the studies used other error measures, such as percentage of the series for which

forecast accuracy was improved by use of the rule.) This paper does not investigate the effect on error reduction of following *all* relevant guidelines, but it seems plausible to assume that the total error reduction from doing so would be at least somewhat greater than the error reduction from following only the guideline associated the greatest expected error reduction. *What is remarkable is that no matter what the criteria, data set, forecast horizon, or type of problem, the authors of this article were unable to find any studies where following any of the checklist items harmed accuracy.*

This paper summarizes evidence on the effects of each guideline. In some cases, the guidelines are deduced from indirect evidence. The evidence was obtained by using computer searches of the literature, seeking help from key researchers, and following up on references in key papers.

To ensure the evidence is properly summarized and to check whether any relevant evidence had been overlooked, email messages were sent to the lead authors articles. that were cited in substantive ways. Reminder messages were sent to authors who did not respond. Responses were received for 67 percent of the papers.

Insert Exhibit about here

On the Value of Checklists

Unaided human brains are maladapted for solving complex problems with many variables. Think of operating a nuclear power plant or flying an airplane. For such tasks, checklists remind users to follow standard practice and, better yet, they can provide evidence-based guidelines.

Arkes, Shaffer, and Dawes (2006) provide a review of evidence on the efficacy of checklists. For example, an experiment regarding infections 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).

Exhibit: Golden Rule Checklist
(With evidence on percentage error reduction)

Guideline	*Done or N/A (✓ or ✗)*	*% error reduction*
1. Problem formulation		
1.1. Obtain and use all important knowledge and information		
1.1.1. Obtain important knowledge and information, including analogies	☐	
1.1.2. Decompose the problem to best use knowledge, information, judgment	☐	(26–49)
1.1.3. Select evidence-based forecasting methods validated for the situation	☐	(88)
1.2. Avoid bias		
1.2.1. Conceal the purpose of the forecast	☐	
1.2.2. Specify multiple hypotheses and methods	☐	
1.2.3. Obtain signed ethics statements before and after forecasting	☐	
1.2.4. Structure adjustments for important knowledge outside the model	☐	(4–64)
1.3. Provide full disclosure to encourage independent audits and replications	☐	
2. Judgmental methods		
2.1. Use structured judgmental procedures	☐	
2.2. Frame questions in various ways	☐	
2.3. Combine independent forecasts from heterogeneous experts	☐	(12)
2.4. Write reasons why the forecast might be wrong, then revise the forecast	☐	(15)
2.5. Use judgmental bootstrapping	☐	(6)
2.6. Use structured analogies	☐	(25-82)
3. Extrapolation methods		
3.1. Use all relevant, reliable, and important data	☐	(2–40)
3.2. Decompose by causal forces	☐	(56)
3.3. Be conservative when forecasting trends if the…		
3.3.1. historical trend conflicts with causal forces	☐	(17–43)
3.3.2. series is variable or unstable	☐	(5)
3.3.3. forecast time horizon is longer than the historical series	☐	
3.3.4. short- and long-term trend directions are inconsistent	☐	
3.4. Estimate seasonal factors conservatively when…		
3.4.1. causal knowledge is weak	☐	
3.4.2. few years of data are available	☐	
3.4.3. factors vary substantially across years	☐	(5)
4. Causal methods		
4.1. Use prior knowledge to select variables and estimate effects	☐	(20)
4.2. Estimate weights conservatively	☐	(4)
4.3. Use diverse information and models	☐	(3–39)
4.4. Use all important variables	☐	(48)
5. Combine forecasts from validated methods and diverse data	☐	(60)

Another study reports on the application of a 19-item checklist for 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.

Formulating the Problem

Forecasters must first formulate the forecasting problem. Proper formulation allows for a more effective use of prior knowledge.

Obtain and use all important knowledge and information (1.1)

Forecasts should be based on cumulative knowledge about the forecasting problem and relevant evidence-based forecasting methods. Forecasters typically need to work with domain experts in order to acquire the relevant prior knowledge and data.

Obtain all important knowledge and information, including analogies (1.1.1): Conservative forecasting requires the forecaster to obtain all relevant information, and sufficient understanding of the relevant theories and evidence in order to realistically represent the situation in forecasting models. A superficial knowledge of the situation can lead to large errors if the forecaster, for example, fails to recognize that the state of knowledge is contested and uncertain, or is ignorant of key relationships or of special events that may have major impacts.

One way to obtain knowledge and information is to ask a heterogeneous group of experts to independently list relevant and important variables, the directions and strengths of their effects, the support they have for their judgments, and their recommendations of relevant data. Nevertheless, experts may not be familiar with what happened in the past or their recollections may be unreliable, and so forecasters should search the research literature for evidence about causal relationships. Meta-analyses (where authors summarize relevant prior studies in structured ways) are especially useful. Computer searches are vital, but contacting key researchers in the field and to following-up on references in their papers is generally more effective.

Information from analogous situations can be used to supplement information about the target situation, especially when data is sparse. Data from analogs can be used in following many of the Golden Rule Guidelines.

Decompose the problem to best use knowledge, information, and judgment (1.1.2): Decompose the problem in ways that help in the collection of knowledge and relevant data about the situation. Decomposing the problem also enables forecasters to draw upon diverse expertise and more data. Decomposition is conservative because the errors from forecasts of the parts are unlikely to all be in the same direction. Decomposition improves accuracy most when the situation being forecast is uncertain.

Decomposition allows forecasters to better match forecasting methods to the situation by, for example, using causal models to forecast market size, using data from analogous geographical regions to extrapolate market-share, using extrapolation to forecast business-as-usual demand, and using judgmental adjustments to forecast the effects of special events such as advertising campaigns or product design changes.

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

Additive decomposition is conservative because it allows the use of more information, and is likely to be more accurate than a direct forecast of the aggregate to the extent that the forecaster is able to obtain knowledge about the segments. A combination of conditions that is likely to be particularly favorable to additive decomposition is when many variables are known to have important effects on the aggregate being forecast, and large databases on the variables are available.

A form of additive decomposition that can be important for time-series forecasting is to break the problem down by considering the current status and the trend separately, then adding the estimates. The procedure is sometimes referred to as "nowcasting". Nowcasting is important when current levels are uncertain. The repeated revisions of official economic data suggests that uncertainty about data is a common problem. For example, one study analyzed deviations between initial and revised estimates of quarterly GDP growth from 1961 to 1996 (Runkle 1998). The figures revealed upward revisions of as much as 7.5 percentage points and

downward revisions of as much as 6.2 percentage points. These problems get worse if data collection is difficult, as is the case in many African countries. An update of the system used in Ghana, for example, resulted in a GDP estimate that was 62 percent higher than the previous estimate (Devarajan 2013). When adjustments to data are secret or subject to political interference, the problem is greater still. A review of industrial output figures for the Chinese town of Henglan found that the town's economic development bureau had overstated output by almost four times (McMahon 2013). Such errors harm forecast accuracy. The problem occurs also in developed countries. One study 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 (Zarnowitz 1967). ~~With fewer resources available for their collection, unoffical data are unlikely to be immune to such estimation problems~~. [Hard to understand and the reason makes no sense to me.]

With data often uncertain, forecasters should seek alternative estimates. There are often many ways to assess the current situation. Consider, for example, estimating the current level by combining the latest observation with estimates from evidence-based methods such as exponentially smoothed levels with a correction for lag, or with the constant from a regression model of the series against time, or with survey data.

A study on forecasting U.S. lodging market sales provides a demonstration of the importance of starting with an accurate estimate of the current level. An econometric model was used to provide 7 one-year-ahead forecasts, 6 two-ahead, and so on for the 7-year period from 1965 through 1971; 28 forecasts in total. The Mean Absolute Percentage Error (MAPE) was 15.7 when the starting level was based on the official data available at the time, but only 10.7 when the starting level was an average of the official data and an econometric estimate, an error reduction of 32% (Tessier and Armstrong 1977).

Jørgensen (2004) found that when seven teams of experts forecast the number of hours needed to complete software projects, the errors of bottom-up forecasts were 49 percent smaller than the errors of direct forecasts.

Dangerfield and Morris (1992) used exponential smoothing models to forecast all 15,753 unique series that can be 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.

Armstrong and Andress (1970) used data from 2,717 gas stations to estimate a stepwise regression model that included 19 variables out of a possible 37, which was then used to forecast sales for 3,000 holdout gas stations. Forecasts were also obtained from a segmentation model that included 11 of the same initial 37 variables (e.g. building age, and open 24 hours). The segmentation model forecasts had a MAPE of 41 percent compared to 58 percent for the regression model's forecasts. This represents an error reduction of 29 percent.

Carson, Cenesizoglu, and Parker (2011) forecast total monthly U.S. commercial air travel passengers for 2003 and 2004. They estimated a multivariate 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, then added the forecasts across the airports to get an aggregate forecast. The mean absolute error 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.

Multiplicative decomposition involves breaking the problem down to elements that can be multiplied. The method is commonly used to forecast a company's sales by multiplying forecasts of the total market by forecasts of market share.

The multiplicative decomposition approach was used to structure 15 highly uncertain situations in three experiments, with subjects making judgmental forecasts of each component. The averages of the forecasts for each component were then multiplied. The procedure reduced median error ratios by 42 percent relative to directly forecasting sales (MacGregor 2001, Exhibit 2).

Select evidence-based forecasting methods validated for the situation (1.1.3): Use only procedures that have been empirically validated under conditions similar to those in the situation being forecast. Fortunately, there is much evidence on which forecasting methods work best under which conditions. The evidence derives from empirical comparisons of the out-of-sample forecast accuracy of alternative methods. The evidence was summarized in as

evidence-based principles (condition/action statements) in the *Principles of Forecasting* handbook, a collaborative effort of 40 forecasting researchers and 123 expert reviewers (Armstrong 2001c). To our knowledge, this is the only published summary of evidence-based forecasting principles. The principles have been available at forprin.com since 2000.

The *Forecasting Method Selection Tree* at forprin.com summarizes how to select which of 15 forecasting methods are appropriate to the conditions. There is rarely a single best method.

Despite the extensive evidence on forecasting methods and the availability of evidence-based principles for forecasting, many forecasters overlook this knowledge. Consider the forecasts that are the basis of manmade global warming policies (Randall et al. 2007). An audit found that the procedures used to predict dangerous warming violated 72 of 89 relevant forecasting principles (Green and Armstrong 2007a).

Similarly, do not assume that published methods have been validated. Many statistical forecasting procedures have been proposed simply on the basis of experts' opinions or defective validation studies. A recent example is a model for forecasting sales of high-technology products proposed by Decker and Gnibba-Yukawa (2010). The authors tested the accuracy of their model on only *six* holdout observations from three different products and concluded that the model provides highly accurate forecasts. However, a reanalysis of the model's performance using a more extensive dataset, consisting of fourteen products and 55 holdout observations, found no evidence that the utility-based model yields more accurate forecasts than a much simpler extrapolation model (Goodwin and Meeran, 2012).

In general, statisticians have 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 time-series studies were published over that time.

Further, do not assume that well-known and widely-used methods have been validated. Box and Jenkins (1970) provide an example of a popular but unsupported statistical method proposed for forecasting. 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

(Collopy and Armstrong1992a). However, little validation research had been done despite many journal articles and widespread applications. When validation tests were done, Box-Jenkins procedures performed poorly relative to evidence-based procedures. For example, the M2- and M3-Competitions compared the accuracy of Box-Jenkins forecasts to the accuracy of damped trend and combining 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 included 29 series and 30 time horizons, and the M3-Competition included 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 Box-Jenkins forecasts was 38 percent larger than the MAPE of the damped trend forecast in the M2-Competition and 3 percent larger in the M3-Competition. The Box-Jenkins forecast error was 37 percent larger than the combined forecast error in the M2-Competition and 4 percent larger in the M3-Competition

Professional forecasters should validate their methods against evidence-based methods. Clients should inquire about testing rather than assumng that it was done, and esepecially ask about independent evaluations. For example, independent evaluations of popular commercial programs sold by Focus Forecasting concluded that the 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)

Biased data and methods lead forecasters to depart from the cumulative knowledge about a problem. The government sector is particularly vulnerable to biased forecasting because forecasting failures typically go unpunished. Instead, governments often reward poor forecasts by providing additional funding when outcomes are disappointing, arguing that the policy intervention was not large enough. Bias also occurs in firms; for example in forecasting the effects of medical treatments.

Biases may arise from political pressures. For example, the U.S. Environmental Protection Agency hearings held over 1971 and 1972 concluded that DDT did not present a danger to human health and wildlife. However, William Ruckelshaus, the Agency's first

Administrator, overruled the finding and banned the pesticide based on his unaided judgment influenced by public protests. The ban is estimated to have led to the deaths of millions of people (Edwards 2004). The use of DDT has picked up again in recent years due to its effectiveness in eradicating malaria with few harmful side effects (Roberts, Tren, Bate, and Zambone 2010).

***Conceal the purpose of the forecast (1.2.1)*:** Forecasters sometimes depart from prior knowledge due to biases they may be unaware of, such as optimism or using the most easily available data. Financial incentives and deference to authority can also cause forecasters to ignore prior knowledge. For example, polar bear population forecasting reports requested by the U.S. Fish and Wildlife Service included title pages that read "USGS [U.S. Geological Survey] Science Strategy to Support U.S. Fish and Wildlife Service Polar Bear Listing Decision." A listing decision required forecasts of declining polar bear numbers. The State of Alaska commissioned an audit of the forecasting procedures (Armstrong, Green, and Soon 2008). Consistent with this guideline (1.2.1), the Alaskan government officials provided no guidance on the purpose of the audit or the desired outcome.

To implement this guideline, one can hire independent analysts who are unaware of the forecasts' purpose. They can be provided with written instructions on how to select, clean, transform, and adjust data, and then to obtain the forecasts using the prescribed methods.

Specify multiple hypotheses and methods (1.2.2): The use of experimental evidence on multiple reasonable hypotheses is the ideal way to reduce bias. It has a long tradition in science (e.g., see Chamberlin 1890. 1965). For example, an unbiased formulation of the polar bear population problem would be to forecast the most likely outcome and thus to consider whether the evidence is most consistent with a decreased, or increased, population. At the time of the Senate Hearing "Examining Threats and Protections for the Polar Bear" on January 30, 2008, the polar bear population had been growing for over 3 decades due to hunting restrictions. With the restrictions still in place, one might expect the upward trend to continue, at least in the short-term. However, as noted above, the government forecasters were asked to prepare forecasts to "Support U.S. Fish and Wildlife Service Polar Bear Listing Decision." In response, a rapid *decline* in the polar bear population was forecast. The testing of alternative reasonable approaches is vital. For example, to assess the effects of a medical treatment, one must show how it performs against alternative treatments, including doing nothing. 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, 40 percent reversed that practice, whereas 38 percent reaffirmed it."

Obtain 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 cost-benefit estimates for large-scale public works projects. In such cases, benefit forecasts are commonly biased upwards and cost forecasts downwards in order to meet a government's benefit-cost ratio criteria. 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).

To reduce bias, obtain signed ethics statements from all of the forecasters involved at the onset 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 relationships that might lead to a conflict of interest. Laboratory studies have shown that when people reflect on their ethical standards, they behave more ethically (Armstrong 2010, pp. 89-94; Shu, Mazar, Gino, Ariely, and Bazerman 2012).

Structure adjustments for important knowledge outside the model (1.2.4): Adjustments often introduce biases. 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). In psychology, extensive research on cross-sectional data led to the conclusion that one should not subjectively adjust forecasts obtained from a quantitative model. For example, a summary of research on personnel selection revealed that employers should rely on statistical models and they should not meet job candidates because this leads them to adjust the forecasts from statistical models to the detriment of accuracy (Meehl 1954). Armstrong (1985, pp. 235-238) summarizes seven studies on this issue.

Unfortunately, forecasters and managers are often tempted to adjust forecasts from quantitative methods. In a survey of forecasters at 96 U.S. corporations, 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. The main reasons the respondents gave for revising quantitative forecasts were to incorporate knowledge of the environment (39 percent), product knowledge (30 percent), and past experience (26 percent); Sanders and Manrodt

(1994). In addition, Fildes et al. (2009) found that among the four companies they studied, up to 91 percent of more than 60,000 statistical forecasts were judgmentally adjusted.

Not surprisingly, the vast majority of forecasting practitioners believe that judgmental adjustments improve the accuracy of their forecasts and most expect the error reductions to range between five and ten percent (Fildes and Goodwin 2007). Evidence does not appear to support this belief. Adjustments that follow structured procedures, however, offer some promise. In one experiment, 48 subjects made adjustments to one-period ahead sales forecasts. When no specific instructions were provided, subjects adjusted 85% of the statistical forecasts and the revised forecast had a MdAPE of 10 percent. In comparison, when being asked to justify the adjustment by picking a reason from a pre-specified list, subjects adjusted only 35% of the statistical forecasts; the MdAPE was 3.6 percent and thus 64% below 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 (Goodwin, 2000). Unfortunately, documentation of the reasons for adjustments is uncommon (Fildes and Goodwin 2007).

In view of the potential for judgmental adjustments to introduce bias (see Exhibit 1, guideline 2.1), adjustments should only be made when the conditions for successful adjustment are met and when bias can be avoided (Goodwin and Fildes 1999; Fildes, Goodwin, Lawrence, and Nikolopoulos 2009). The adjustment procedure should be structured. The adjustment and the procedures should be documented. 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). The estimates should be made in ignorance of the forecasts from the model, but with knowledge of what variables and other information the model uses (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). Structured judgmental adjustments can improve accuracy when experts have good knowledge of the effects of special events and changes in causal forces (Fildes and Goodwin 2007). Compose the final forecast from the model forecast and the experts' adjustments. If the Golden Rule guidelines are followed, judgmental adjustments should be rare.

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

Fully disclose the data and methods used for forecasting, and describe how they were selected. This is important not only to identify and avoid biases, but also to facilitate independent audits and replications.

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. One study found 23 books and articles, most of which were peer-reviewed, that included mistakes in the trend component of exponential smoothing model formulations (Gardner 1984). A follow-up study found mistakes in exponential smoothing programs used in two companies (Gardner 1985).

Judgmental Methods

Judgment is 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. Unfortunately, judgmental methods are prone to biases.

Use structured judgmental procedures (2.1)

By structured judgmental procedures, we mean procedures that are formal and evidence-based. Unaided judgment is not conservative because it is a product of faulty memories, inadequate mental representation of complex phenomena, and unreliable record keeping, to mention only a few of the shortcomings. Moreover, when experts use their unaided judgment, they tend to more easily remember recent, extreme, and vivid events (Kahneman 2011). As a result, they overemphasize the importance of such events when making judgmental forecasts, which leads them to overestimate change. Indeed, in a study of 27,000 political and economic forecasts made over a 20-year period, 284 experts from different fields tended to did just that (Tetlock 2005, pp. 83).

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 apparently to resemble the noise or apparent pattern in the historical series (Harvey 1995)—a procedure that is virtually certain to increase forecast error relative to a straight line. Green, Soon, and Armstrong (2013) provide details on this study.

The biases of unaided judgment tend to be increased when forecasts are made in group settings. Group members may be reluctant to share their opinions because they wish to avoid conflict or ridicule. When making forecasts for important decisions, people tend to rely on the unaided judgments of groups despite that common approach's lack of predictive validity. Experimental evidence demonstrates that it is difficult to find a method that produces forecasts as poor as unaided judgments from group meetings (Armstrong 2006b).

Frame questions in various ways (2.2)

The way that a problem is phrased can have a large effect on how experts answer. The long-standing approaches for dealing with this problem are to pose the forecasting question in multiple ways and to pre-test the different wordings to ensure the questions are understood in the way that the forecaster intends. The final questions should be written and followed exactly. Responses to the variant questions should be combined.

Combine independent forecasts from heterogeneous experts (2.3)

To increase the amount of information considered and to reduce the effects of biases, combine anonymous forecasts from a heterogeneous group of independent experts. Surprisingly, contributions from experts with only modest domain knowledge can improve the accuracy of a combined forecast (Armstrong 1980a). Good results can be achieved by combining forecasts from eight to twelve experts whose knowledge of the problem is heterogeneous and whose biases are likely to differ. Adding forecasts from more experts helps, although the rate of improvement in forecast accuracy diminishes substantially (Hogarth 1978).

One review presented evidence from seven studies that involved combining forecasts of 4 to 79 experts. Combining forecasts reduced error between 7 to 19 percent compared to the typical expert forecast (Armstrong 2001a). 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 Jr., and Cuzán 2013).

Write reasons why the forecast might be wrong, then revise the forecast (2.4)

To make judgmental forecasts more conservative, ask experts to write reasons why their forecasts might be wrong. Doing so will encourage them to use more information. It will also help to avoid optimism and overconfidence.

In one experiment, researchers 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, along with a description and numerical rating of each reason's strength. Their predictions were more accurate than when they did not provide reasons, reducing error by 11 percent. In addition, subjects in the no-reason conditions were much more overconfident (Koriat, Lichtenstein, and Fischhoff, 1980, Table 1).

In a 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 and overconfidence by 30 percent compared to providing no reason. However, providing supporting reasons versus providing both supporting and contradicting reasons made little difference to accuracy (Koriat, Lichtenstein, and Fischhoff 1980).

Another study asked students to predict the outcome of their job search efforts over the next 9 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 (Hoch 1985).

The Delphi method elicits independent and anonymous forecasts, along with reasons, from experts. It then summarizes the forecasts and reasons and provides them as feedback to the experts. The experts can revise their own forecasts, free from group pressures, in later rounds.

Delphi tends to provide more accurate forecasts than alternative group methods. A review of the literature concluded that Delphi outperformed statistical groups (i.e., simple one-round surveys) in twelve studies and was less accurate in two studies, with two ties. Compared

to traditional meetings, Delphi was more accurate in five studies and less accurate in one study; two studies showed no difference (Rowe and Wright 2001). Results from a laboratory experiment on estimation tasks confirm these findings, showing that the Delphi method not only outperformed prediction markets, but also was easier to understand (Graefe and Armstrong 2011). The Delphi method is well suited when relevant knowledge is distributed among experts and thus their initial opinions vary.

Use judgmental bootstrapping (2.5)

People are often inconsistent in applying what they know to a problem. They might get overloaded with information, forgetful, tired, distracted, or irritable. Judgmental bootstrapping offers a method for applying forecasters' implicit rules in a consistent way to routine forecasting problems.

To forecast using judgmental bootstrapping, develop a quantitative model to infer how an expert or a group of experts makes the forecasts. First, present an expert with 20 or more 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. This is the judgmental bootstrapping model.

A review of eleven studies using cross-sectional data from various fields (e.g., personnel selection) found that forecasts from judgmental bootstrapping models were more accurate than forecasts from unaided judgment in eight studies, there was no difference in two studies, and bootstrap forecasts were less accurate in one in which an incorrect belief on causality was applied more consistently (Armstrong 2001b). Most of these studies reported accuracy in terms of correlations. One of them, however, reported an error reduction of 6.4 percent.

Judgmental bootstrapping can also reveal when forecasters rely on irrelevant information. For example, beauty and height are not predictive of performance as a computer programmer. Forecasters can eliminate spurious variables from the model and thereby improve the forecasts.

Use structured analogies (2.6)

A target situation is likely to turn out like analogous situation. For example, to forecast trolley ridership in a U.S. city, one might examine the performance of similar trolley systems around the world (see Scheib 2012 for examples). Obtaining evidence on behavior from analogous situations is consistent with the conservative practice of ensuring forecasting is consistent with cumulative knowledge. To forecast using analogous data, ask independent experts to identify analogous situations from the past, to rate each analogy's similarity to the current (target) situation, and to identify the outcome of each. Then calculate the average of the outcomes implied by each expert's top-rated analogy. This typical outcome is the forecast for the target situation. The method is known as structured analogies.

This intuitively appealing method can provide easily understood forecasts for proposed large and complex projects. For example, to forecast whether the California High Speed Rail would cover its costs, a forecaster could ask experts to identify similar high-speed rail systems (HSR) and obtain information on their profitability. The Congressional Research Service did just 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 Session 2013).

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

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 were described to experts, including union-management disputes, corporate takeover battles, and warfare. 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).

Extrapolation methods

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 abuses of extrapolation.

Use all relevant. reliable, and important data (3.1)

Conservative forecasting requires that forecasters use all of the relevant, reliable, and important data. This includes data about analogies. In addition, the forecaster should avoid data that do not meet these criteria.

Practitioners often violate this guideline. For example, a forecaster has much influence over the resulting forecast by selecting a particular starting point for estimating a time-series forecasting model or by selecting a specific subset of cross-sectional data. Such judgments allow people to make forecasts that support their prior beliefs.

When analogous situations are identified in an unbiased way, they can provide useful data for estimating extrapolation models. This is relevant for forecasting trends as well as seasonal factors. For example, consider that one wishes to forecast sales of the Hyundai Genesis automobile. Rather than relying only on the Genesis data, look at the trends for all luxury cars, and then combine the two estimates.

Now assume an estimate is needed for seasonal factors. For the Hyundai Genesis, one would combine seasonal factors from time-series on analogous car sales with those from the Genesis series. Using analogous data in that way 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 161 (56%) 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 (Withycombe 1989).

In an analysis of 44 series of retail sales data from a large U.K. department store chain, forecasts from models that used seasonal factors estimated from analogous series were consistently more accurate with error reductions of up to 40 percent (Bunn and Vassilopoulos 1999). Another study combined seasonal crime rates from six precincts in Pittsburgh and found

that the combined-seasonality forecast errors were about 8 percent smaller than the individual-seasonality forecast errors (Gorr, Olligschlaeger, and Thompson 2003).

Decompose by causal forces (3.2)

Ask domain experts to assess whether the trend of a series will be due to causal forces. These have been defined as growth, decay, supporting, opposing, regressing, or unknown (Armstrong and Collopy 1993). Growth, for example, means that the causal forces will lead the series to increase, irrespective of the historical trend. When forecasting a time-series that is the product of conflicting causal forces such as growth *and* decay, decompose the series by these forces and extrapolate each component separately. By doing so the forecaster is being conservative in adhering to 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 tends to decrease as the safety of vehicles and roads increases. 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.

One study tested the value of decomposition by causal forces for twelve annual time-series for airline and automobile accidents, airline revenues, personal computer sales, and cigarette production. The researchers expected decomposition to yield forecasts that were more accurate than those from simple 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. Successive updating was used 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 of the two conditions, forecasting the decomposed series separately reduced the median absolute percentage error (MdAPE) of the combined forecasts by 56 percent relative to forecasts from extrapolating the global series. For the three remaining series, decomposition reduced error by 17 percent (Armstrong, Collopy, and Yokum 2005).

Be conservative when forecasting trends (3.3)

Extrapolate conservatively by relying on causal information about the trend. One common conservative approach is to damp the magnitude of the trend. This brings the trend closer to the estimate of the current situation. However, damping might not be conservative if it resulted in a substantial departure from a consistent long-term trend arising from well-supported and persistent causal forces. In such cases, consider whether support for departing from a trend is provided by evidence from analogous series.

***Be conservative when forecasting trends if the historical trend conflicts with causal forces (3.3.1)*:** 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. Research findings to date suggest a simple guideline that works well: *ignore trends for contrary series.*

One study compared the performance of this "contrary series rule" to forecasts from Holt's exponential smoothing, a method that ignores causal forces. It used annual data from twenty series from the M-Competition (Makridakis et. al. 1982) that were rated as contrary. By removing the trend term from Holt's model, the Median Absolute Percentage Error (MdAPE) was reduced by 18 percent for one-year-ahead forecasts, and by 40 percent for six-year-ahead forecasts. Additional testing involved contrary series from four other data sets: Chinese epidemics, unit product sales, U.S. Navy personnel numbers, and economic and demographic variables. On average, the MdAPE for the no-trend forecasts was 17 percent less than for Holt's forecast errors for 943 one-year-ahead forecasts. For 723 long-range forecasts, the error reduction was 43 percent (Armstrong and Collopy 1993).

Armstrong, Green and Soon (2008) encountered a violation of the contrary series in their study of a government agency's polar bear population forecast. The historical data suggested a long-term upward trend in the polar bear population as a consequence of hunting restrictions. The agency's commissioned report forecast the trend would reverse immediately and that there would be a rapid decline toward extinction of polar bears unless the government acted. Evidence since 2007 indicates that the upward trend in population has continued ,

Be conservative when forecasting trends if the series is variable and unstable (3.3.2). In a review of ten studies, damping the trend by using statistical rules on the variability in the historical data yielded an average error reduction of about five percent (Armstrong 2006a).

Be conservative when forecasting trends if the forecast horizon is longer than the historical series (3.3.3): Uncertainty is high if 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) increasing the damping of the trend toward zero as the forecast horizon increases and (2) avareaging the trend with trends from analogous series. The U.S. Fish and Wildlife Service scientists in the aforementioned polar bear population study overlooked the need for damping when they used only five years of historical data to forecast an immediate and sharp decline in population extending 50 years into the future.

***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 change in causal forces, along-term trend represents more knowledge about the behavior of the series than does a short-term trend, and thus provides a more conservative forecast.

Estimate seasonal factors conservatively (3.4)

For situations clearly affected by the season, such as monthly sales of sunscreen or furnace oil, seasonal factors improve forecast accuracy. When the situation is uncertain as well as seasonal, one conservative procedure is to damp the estimated seasonal effects: toward 1.0 for multiplicative factors or toward zero for additive factors. Another approach is to combine the seasonal factors estimated for the series of interest with those estimated from analogous series. Still another approach is to combine the estimate of a seasonal factor with those from the time period before and the period after.

Estimate seasonal factors conservatively when causal knowledge is weak (3.4.1): Without prior knowledge and theory 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.

Estimate seasonal factors conservatively when few years of data are available (3.4.2): For situations lacking strong evidence on the causes of seasonality, avoid using seasonal factors unless there are at least three years of historical data from which to estimate them. Chen and Boylan (2008) found that, in general, seasonal factors harmed accuracy when they were estimated from fewer than three years of data. Alternatively, consider estimating seasonal factors from three or more analogous series.

Estimate seasonal factors conservatively when the factors vary substantially across years (3.4.3): If estimates of the size of seasonal factors differ substantially from one year to the next, this suggests either that the factors do indeed vary—as from large responses to unusually warm or cool seasons—or that the estimates are capturing other, non-seasonal, variations in the data, shifting dates of major holidays, strikes, natural catastrophes, and so on. To avoid the harm to forecast accuracy that would arise from using factors thus affected, forecasters should reduce the magnitude (damp) the estimated seasonal factors or combine the factors with those from adjacent time periods. Miller and Williams (2004) did this for the 1,428 monthly series of the M3-Competition by applying statistical rules that damped estimated seasonal factors for each time-series based on the degree of variability. Forecasts based on damped seasonal factors were more accurate for 59 to 65 percent of the series and gains in accuracy were larger for short-term forecasts, one-to-three-months ahead. For series where the tests of variability called for damping, MAPEs were reduced by up to five percent.

A follow-up study by Chen and Boylan (2008) first tested the Miller and Williams rules for damping using 111 monthly series from the M-competition (Makridakis and Hibon 1979). The gains in accuracy were similar to those obtained by Miller and Williams (2004). They then tested the seasonal factor damping rules on 218 monthly series on light bulbs manufactured in the U.K. and found that the damping improved forecast accuracy.

Causal methods

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 by damping the coefficients for the predictor variables in response to unexplained variability in the historical data.

A regression model is not sufficiently conservative because it does not reflect uncertainty in predicting the causal variables or changes in causal factors, nor if any variable in

the model correlates with important excluded variables over the estimation period. But the primary problem is that occurs when forecasters depart from conservatism by searching for the "best" predictor variables, a problem that intensifies when large databases are used. These conclusions are consistent with those from a review of why the effects of newly discovered relationships are commonly inflated in the medical field (Ioannidis 2008). 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, pp. 298). Until the latter part of the 20th Century, data collection and statistical analyses remained expensive, and forecasters had little choice but to develop their models using well-supported theories and *a priori* analysis.

In econometrics, *a priori* analysis uses elasticities to summarize prior knowledge. Elasticities are unit-free and easy to interpret. They represent the ratio of the percentage change that occurs in the variable to be forecast in response to a percentage 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 should examine prior research in order to determine the expected elasticities and their plausible lower and upper bounds. For problems in economic forecasting, one can find estimates of income, price, and advertising elasticities in published meta-analyses. If little prior research exists, use estimates from domain experts and from data on the specific situation. To obtain elasticity estimates from available data, formulate models in multiplicative terms by calculating the logarithm of the data values and then run the regression analysis.

One study tested the value of such an a priori analysis for forecasting international camera sales by fully specifying a model based on prior knowledge about causal relationships and before analyzing the data (Armstrong 1970). The final model coefficients were then calculated as a weighted average of the *a priori* estimates and coefficients derived from a regression analysis, a process later referred to as a "poor man's Bayesian regression analysis"

(Armstrong and Grohman 1972). In a test of predictive value, data from 1960 to 1965 for 17 countries were used to estimate the model and calculate a "backcast" of 1954's camera sales. Compared to a benchmark model with statistically estimated coefficients and no use of a priori information, the model that included a priori knowledge reduced the MAPE by 23 percent. Another test estimated models using 1960-1965 data for 19 countries and to predict market size in 11 holdout countries. Here, the model using a priori information reduced error by 40 percent.

Bayesian analysis provides another way to incorporate prior knowledge into a forecasting model. There is an enormous number of journal articles on Bayesian forecasting, leading to about 6,000 hits for "Bayesian forecasting" on Google Scholar as of October 2013. Despite the volume of research, the authors of this article are not aware of evidence that forecasts from Bayesian forecasting are more accurate than those from alternative evidence-based forecasting methods. One study finds that a simple average of forecasts from six established regression models for forecasting U.S. presidential elections produced errors that were 19 percent lower than those from a Bayesian approach to combining forecasts (Graefe 2013a). The lack of evidence that Bayesian analysis provides cost-effective improvements in accuracy for practical forecasting situations is a blessing for those who prefer simple straightforward methods.

Since the 1960s, the stress on prior knowledge in developing causal models has given way to a preference for statistical procedures in the belief that they would replace the need for an *a priori* analyses, which can be time consuming and expensive, and which must be done by people with high expertise in the field. As computers make it possible to handle more and more data, analysts increasingly rely on ever more sophisticated statistical procedures to select predictor variables and estimate the relationships among them. The belief that complex statistical procedures yield greater forecast accuracy is widespread but by no means new. A survey of leading econometricians in the mid-1970s showed support for the belief that complex statistical procedures yield greater forecast accuracy (Armstrong 1978bIn a discussion of the limitations of four complex analytical techniques (i.e., automatic interaction detection, multiple regression analysis, factor analysis, and nonmetric multidimensional scaling), Einhorn's (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" (pp. 378). Research since then supports Einhorn's assessment

(Armstrong 2012). Despite this, atheoretical approaches keep appearing under different names such as stepwise regression, data mining, and analytics. We view these techniques as minefields where spurious relationships can cause much harm as, for example, in predicting the effects of various health treatments, environmental regulations, economic policies, and business strategies.

Despite the theoretical and empirical objections, much current practice continues to ignore prior knowledge in favor of selecting models that show statistically significant relationships in a given set of data. Journal editors and reviewers favor statistically significant results, and researchers oblige them by searching for the model that best fits available data, a procedure that typically harms accuracy (Armstrong 2007).

Estimate weights conservatively (4.2)

As with extrapolation forecasts, damping is useful for making causal model forecasts more conservative. One strategy is to damp estimates of each variable's coefficients (weights) toward zero. The process is sometimes referred to as shrinkage. It reduces the amount of change that a model will predict in responses to changes in the causal variables, and is thus conservative.

Another 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 move 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 (2013b) in the present issue, a large body of analytical and empirical evidence since the 1970s has found that equal-weights models often provide more accurate forecasts than those from regression models. That paper also provides further evidence for U.S. presidential election forecasting, a field that is dominated by the use of regression analysis. When calculating equal-weights variants of nine established regression models, the equal-weight models yielded more accurate forecasts for six of the nine models. On average, the error of the equal-weights model forecasts was four percent lower than the error of the regression models' forecasts.

Use diverse information and models (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 might be affected by fifty or more variables. In addition, many of causal variables may not vary over long periods, and reliable data may be difficult to obtain. In such situations, regression models ignore important knowledge.

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. For example, 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 forecast was only 5.8 percent, an error reduction of 39 percent (Namboodiri and Lalu 1971). 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 and Cuzán 2014).

Use all important variables (4.4)

Another approach to developing conservative causal models is to incorporate all important knowledge about causal relationships into one model. This solution draws on an insight from Benjamin Franklin's "method for deciding doubtful matters" (Sparks 1844). Franklin suggested listing all important variables, identifying the directional effect of each variable, and assigning weights when possible. We refer to his approach as the *index method*. The resulting index models might also be called *knowledge models*, because they include all knowledge about factors affecting the thing being forecast.

To develop an index model, first identify all relevant variables by asking experts and by reviewing experimental evidence. Code the expected directional influence of the variables on whatever is being forecast (e.g., job performance). Some of the prior validation studies have

assigned a coefficient of +1 for variables with a positive influence and zero otherwise. Subjectively weighted variable can also be considered if one has strong evidence on the relative predictive ability of the variables.

To use the index model for forecasting, rate the subject of the forecasts (e.g., a job candidate, an advertisement, a country, etc.) against each variable. If the reliability of the ratings is low, use several raters and average their ratings. Finally, add the variable ratings to calculate a single index score. For selection problems, the option with the highest index score is predicted to be the best. 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). Note that the inclusion of all important variables is possible because the method does not require the estimation of variable coefficients (weights) from data.

The index method has been used to forecast U.S. presidential elections, a situation with knowledge on a large number of predictor variables but few observations. An index model based on 59 biographical variables correctly predicted the winners of 28 of the 30 U.S. presidential elections from 1896 to 2012 (Armstrong and Graefe 2011). Another index model was based on surveys of how voters expected U.S. presidential candidates to handle important issues. The number of issues varied from 23 to 47. The model correctly predicted the election winner in ten of the eleven elections from 1972 to 2012 (Graefe and Armstrong 2013).

Finally, one study created an index model by adding up the standardized values of all 29 variables that are used by nine established US presidential election forecasting models. Across the ten elections from 1976 to 2012, the forecasts error of this index model was 48% lower than the error of the typical individual regression model and 29% lower than the error of the most accurate individual model (Graefe 2013b). This shows the benefits of using cumulative prior knowledge about causal factors.

Combine forecasts from validated methods and diverse data

Combining forecasts from evidence-based methods is a conservative strategy in that it relies on more information. Combining also reduces the effects of mistakes such as data errors, computational errors, or using a poorly specified model. Combining forecasts across methods is particularly valuable if the biases of the forecasts from the different methods are expected to differ from one another.

To reduce the effects of biases, specify the combining procedure (i.e., how to weight the forecasts) prior to making the forecasts. Many scholars have proposed methods for how to best weight the component forecasts. However, a review of over 200 published papers from the fields of forecasting, psychology, statistics, and management science concluded that using equal weights usually provides the best forecast when combining (Clemen 1989). An updated review reinforces Clemen's conclusion (Mancuso and Werner 2013).

Combine forecasts by calculating simple averages unless strong evidence suggests that forecasts from some methods are consistently more accurate in the given situation. Differential weights are appropriate under certain conditions. For example, rule-based forecasting (which varies the weights on extrapolation methods with the horizon, causal forces, variability of the historical data and so on) provided the most accurate forecasts for annual data in the M-Competition (Collopy and Armstrong 1992b). Vokurka, Flores and Pearce (1996) provided additional support. They used automatic rule selection and found errors for 6-year-ahead forecasts on M-Competiton data that were 15 percent less than those for the equal weights combined model.

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). The gains from combining can, however, be considerable, in particular when different evidence-based forecasting methods are used and when the forecasts draw upon different data. This proposition was tested in a study involving forecasts of the popular vote shares in the six U.S. presidential elections from 1992 to 2012. Averaging forecasts within and across four established election-forecasting methods (polls, prediction markets, expert judgments and quantitative models) yielded forecasts that were more accurate than those from each of the component methods. The error reduction compared to the typical component method forecast error was as much as 60 percent (Graefe, Armstrong, Jones Jr. and Cuzán 2014).

Suggestions on the use of the Golden Rule

One way to improve forecasting practice would be to implement conservative forecasting procedures as default options in forecasting software products. For example, it would be a simple and inexpensive matter to include the contrary-series rule (3.3.1) and to

avoid calculation of seasonal factors if there are fewer than 3 years of data (3.4.2). In addition, it would be helpful if software programs provided users with convenient access to the Golden Rule checklist.

Given the failures of sophisticated statistical techniques to improve forecasting, we suggest ignoring them until such time as experimental evidence demonstrates their benefits. That recommendation includes avoiding statistical significance and model fit testing as criteria for developing a forecasting model.

While large databases tend to mislead, they can provide useful data for forecasting through decomposition. Decomposition enables forecasters to make proper use of large bodies of data with information on many important variables where problems exist with respect to interactions, multicollinearity, causal priorities among predictor variables, and non-linear effects.

Do not draw causal inferences from observational data. Associations in observational studies are often presented by researchers and reported in the media as if they provided evidence for causal relationships. Consider, for example, the many forecasts that avoiding certain foods will increase your life span, and exposure to tiny doses of certain chemicals will decrease it. In practice, such exciting associations are often due to chance, confounding by extraneous factors, biased sampling, biased design, or biased reporting of positive findings. There are numerous examples of associations claimed by the authors of individual studies and of overviews that experimental studies find to have been spurious (e.g., see Kabat 2008 on health risk studies). Statistical analysis and reanalysis of large datasets has become a self-perpetuating academic activity with further research either challenging or supporting a previous study, without resolution, until experimental studies are conducted to establish the existence or otherwise of proposed causal relationships.

Do not use forecasts as a motivational tool. Organizations often view forecasts as motivational tools, rather than as attempts to know what will happen given the adoption of a particular strategy. They may believe that an unusual forecast will help them take advantage of an apparent new trend. This is a mistake, because it confuses forecasting with planning. Forecasters should confine themselves to forecasting what is likely to happen if a given strategy is adopted.

Use the no-change model as a benchmark. The no-change model is the ultimate in conservative forecasting. Proposed forecasting models should be validated against this benchmark model. The no-change model differs depending on the situation. It may be based on the long-run level, the extrapolation of a long-run trend, a base rate, or the usual behavior in similar situations.

Consider the complex and uncertain behavior of the stock market in the short-term. One would expect forecasts from the no-change model would be difficult to beat. Unsurprisingly, researchers' attempts to make forecasts that beat the current market price have proven unsuccessful for those who do not have inside information. Malkiel (2012) has documented this phenomenon in a book first published over forty years ago and that is now in its tenth edition.

One study compared the accuracy of forecasts from the no-change model with forecasts from six full-trend extrapolation methods. These involved 180 forecasts 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" (Schnaars and Bavuso 1986). Of course, a more conservative strategy would have combined forecasts from the two methods.

When comparing the efficacy of alternative models, use the Relative Absolute Error (RAE). The measure gives the error of a forecast from a proposed model relative to that of a forecast from the no-change model (Armstrong and Collopy 1992). Thus, a RAE of 1.0 means that the proposed model is no better than the benchmark, a RAE less than 1.0 is better than the benchmark, and a RAE greater than 1.0 is worse than the benchmark.

One inexpensive approach to implementing the Golden Rule is to combine a forecast with the appropriate no-change forecast. More weight should be placed on the no-change model if uncertainty is high. Uncertainty typically increases with the complexity of the problem and the length of the forecast horizon.

Forecasters may resist employing the no-change model because its simplicity and the lack of novelty in the forecasts makes it hard for them to justify high consulting fees. Moreover, a forecast that implies inaction is often unattractive to decision makers.

Use the Golden Rule Checklist. The checklist provides an evidence-based standard against which forecasting procedures can be examined. Applying it requires no training; a normally intelligent person familiar with a forecasting report can quickly assess which Golden Rule guidelines are relevant and whether forecasters followed them. If the description of the forecasting procedure is unintelligible, be conservative and ignore the report—even if it was expensive.

The first two authors used the Checklist to assess the forecasting procedures described in the IPCC's Fourth Assessment Report (Randall et al. 2007). Given their earlier familiarity with the report, it took them each only ten minutes to do the ratings. They concurred that 25 of the 28 guidelines were relevant. None of the 25 relevant guidelines were followed. Consistent with this, a validation study of the global warming projections used by governments around the world, the error for long-term forecasts (91 to 100 years into the future) was 12 times larger that the conservative no-change forecast (Green, Armstrong, and Soon 2009).

Organizations can use the Golden Rule Checklist to audit in-house or consultant forecasting procedures. A failure to follow the evidence-based guidelines could be the basis for penalties within the firm or for obtaining damages from a forecast provider. Such failures could also provide a basis for challenging government policies and regulations.

On the cost side, forecasters must devote effort to learn about the past (though easier now with Internet searches) and about evidence-based forecasting methods (information is available at no cost at forprin.com).

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 be made by showing evidence that forecasters' practices were incompetent, in that they did not follow evidence-based guidelines. The Golden Rule Checklist and the Forecasting Audit could provide the basis for such cases. One would hope that this possibility would motivate forecasters to learn how to use evidence-based forecasting procedures. To ensure objectivity, forecasters would be advised to use independent auditors and to provide such audits with their forecasts.

The first paragraph of this paper asked how a client should evaluate a forecast. The answer is to use the Golden Rule Checklist from Exhibit 1. One approach would be to ask the consultant to sign a document describing how he applied the guidelines in the Golden Rule checklist. Another would be to obtain Golden Rule checklist audits of the procedures described in the consultant's reports. To be fair, and to increase the chances of obtaining valid and useful forecasts, make the Golden Rule checklist audit a condition of the consultant's engagement.

Barriers to the Progress in Forecasting

This review of experimental evidence finds that the use of conservative forecasting practices leads to forecasts that are consistently and substantially more accurate than forecasts from non-conservative practices. In addition, for about a century, much experimental evidence has been conducted related to which methods are most useful in which situations. Also during this time, new methods have been validated. Surprisingly, however, the practice of forecasting does not however appear to have improved.

Ascher (1978) concluded that forecast accuracy in practice had not improved over time in his review of forecasting for population, economics, energy, transportation, and technology. In a comprehensive review of the research on agriculture forecasting, Allen (1994) was unable to find any evidence that forecasting practice had improved over time. He then compared accuracy of forecasts from 12 studies (22 series) before 1985 and 11 studies after 1985, finding only trivial differences in accuracy. A 20-year follow-up study found evidence that sales forecasts have become less accurate over time (McCarthy, Davis, Golicic and Mentzer 2006).

There are many possible explanations as to why forecasters and their clients fail to follow the Golden Rule. One explanation is the belief that advanced methods for regression analysis, along with very large databases can provide superior forecasts. We have been unable to find evidence to support that belief. Over the past half-century, complex statistical methods, computers, and large databases have seduced forecasters and their clients away from cumulative knowledge and evidence-based forecasting procedures. In short, it has led them to ignore the Golden Rule.

Another problem is that forecasts are often used as a motivational tool to justify a strategy or policy. Economic forecasts commonly show systematic biases depending on the political agenda of the institution publishing the forecast. Politicians want support for public

works projects. Managers of firms want to invest other people's money. These biases are likely to be particularly prevalent among long-term forecasts in the public sector, since the institutions and politicians are rarely evaluated on the accuracy of such forecasts.

Then there is the persistent belief that decision makers are capable of making valid predictions in complex and uncertain situations. They do not trust formal evidence-based forecasting procedures. This belief is common for important problems. For example, on August 4, 2011, the front page of *The New York Times* announced a new educational program, partly funded by New York City, for disadvantaged, young black and Latino males. Confident predictions of the program's success were based on expert opinions. Prior experimental evidence suggests otherwise. Between 1939 and 1944 the Cambridge-Somerville experiment gave counseling and training to a randomly selected half of a group of more than 500 troubled young men. Thirty years later in a follow-up study, those in the experimental program reported that they had benefited immensely from the program, as expected. However, the follow-up also found that program participants had less-prestigious jobs, lower job-satisfaction, and higher crime rates. They also had higher rates of alcoholism, sickness, and mortality (McCord 1978; additional evidence is provided in McCord 2003).

Conclusions

This article proposes that the Golden Rule of Forecasting. Following the Golden Rule guidelines described in this article increased forecast accuracy substantially and consistently no matter what was being forecast, how the guidelines were applied, how many guidelines were used, when the studies were done, how long the forecast horizon, the amount and quality of the data, or what criteria were used for accuracy. Following the Golden Rule also reduces the risk of large errors. The Golden Rule is applicable to all forecasting problems because they all involve uncertainty.

Conservative procedures should be adopted when formulating the forecasting problem, and when making forecasts with judgmental, extrapolative, and causal methods. Forecasts from different evidence-based methods should be combined.

The evidence-based *Golden Rule Checklist* presented in this article provides simple and easily understood guidance on how to make conservative forecasts. That assistance is especially useful when "big data" are available, which might otherwise encourage forecasters

to ignore the need for conservatism. Finally, the Checklist can help non-forecasters judge the value of forecasts by assessing the quality of the forecasting process that gave rise to them.

Acknowledgements: Fred Collopy, Jason Dana, Peter Fader, Everette Gardner, Paul Goodwin, Nigel Harvey, Robin Hogarth, and Don Peters 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, Shantayanan Devarajan, Robert Fildes, Magne Jørgensen, Geoffrey Kabat, Peter Pronovost, Lisa Shu, and Jean Whitmore made suggestions for improvements. Jennifer Kwok copy-edited the paper. Responsibility for the final article remains the authors alone.

References

Allen, P. Geoffrey (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. (1980). The Seer-Sucker theory: The value of experts in forecasting. *Technology Review, 82*(7), 16-24.

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. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, 23, 321-327.

Armstrong, J. S. (2010). *Persuasive Advertising*: 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., & Soon, W. (2008). Polar bear population forecasts: A public-policy forecasting audit. *Interfaces, 38*(5), 382-405.

Armstrong, J. S., & Grohman, M. C. (1972). A comparative study of methods for long-range market forecasting. *Management Science*, *19*(2), 211-221.

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

Box, G. E., & Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

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.

Decker, R. & Gnibba-Yukawa, K. (2010). Sales forecasting in high-technology markets: A utility-based approach. *Journal of Product Innovation Management*, 27(1), 115-129.

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

Edwards, J. G. (2004). DDT: A case study in scientific fraud. *Journal of American Physicians and Surgeons*, 9(3), 83-88.

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 farket 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.

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

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

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

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

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.

Goodwin, P. & Wright, G (1993). Improving judgmental time series forecasting: a review of the guidance provided by research, *International Journal of Forecasting*, 9, 147-161.

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.

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

Graefe, A. (2013a). Conditions of Ensemble Bayesian Model Averaging for political forecasting, Working paper, Available at: http://ssrn.com/abstract=2266307.

Graefe, A. (2013b). Improving forecasts using equally weighted predictors, Working paper, Available at: http://ssrn.com/abstract=2311131.

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., 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. (2013). Evidence-based forecasting for climate change. *Working paper*.

Harvey, N. (1995). Why are judgments less consistent . ..

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. (2012). When simple is hard to accept. In *Ecological Rationality: Intelligence in the World*, edited by Peter M. Todd, Gerd Gigerenzer and ABC Research Group, 61-79. Oxford: Oxford University Press.

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

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology, 19*(5), 640-648.

IPCC (1990). In J. T. Houghton, G. J. Jenkins, & J. J. Ephraums (Eds.), *Climate change: The IPCC scientific assessment*. Cambridge, United Kingdom: Cambridge University Press.

IPCC (1992). In J. T. Houghton, B. A. Callander, & S. K. Varney (Eds.), *Climate change 1992: The supplementary report to the IPCC scientific assessment*. Cambridge, United Kingdom: Cambridge University Press.

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

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrier, Straus and Giroux.

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.

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.

Makridakis, S. & Hibon, M. (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society. Series A (General), 142*(2), 97-145.

Malkiel. Burton 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., and Mentzer. J. T. (2006). The evolutions of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting,* 25. 303-324.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist,* 33 (3) 284-290.

McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime. *The Annals of the American Academy of Political and Social Science,* 587, 16-30.

McMahon, D. (2013). How China fudges its numbers. *The Wall Street Journal*, 19 June. http://blogs.wsj.com/chinarealtime/2013/06/19/a-rare-look-into-how-china-fudges-its-numbers/

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.

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.

Prasad, Vinay et al. (2013), A decade of reversal: An analysis of 146 contradicted medical practices. *MayoClinicProceedings,org*, 790-798.

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, NY, USA: Cambridge University Press.

Roberts, D., Tren, R., Bate, R., and Zambone, J. (2010). *The excellent powder: DDT's political and scientific history*. Indianapolis, IN: Dog Ear.

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, 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.

Scheib, S. L. (2012). The streetcar swindle, *Reason*, http://reason.com/archives/2012/09/27/the-streetcar-swindle.

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 regressions 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. (1977). *Improving current sales estimates with econometric models*. Working paper, Available from http://www.forecastingprinciples.com/paperpdf/improvingsalesestimates.pdf.

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

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.

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

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

Word count: 15,200