Golden Rule of Forecasting: Be Conservative

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

March 2015

Abstract

This article proposes a unifying theory, or Golden Rule, of forecasting. The Golden Rule of Forecasting is to *be conservative.* A conservative forecast is consistent with cumulative knowledge about the present and the past. To be conservative, forecasters must seek out and use all knowledge relevant to the problem, including knowledge of methods validated for the situation. Twenty-eight guidelines are logically deduced from the Golden Rule. A review of evidence identified 105 papers with experimental comparisons; 102 support the guidelines. Ignoring a single guideline increased forecast error by more than two-fifths on average. Ignoring the Golden Rule is likely to harm accuracy most when the situation is uncertain and complex, and when bias is likely. Non-experts who use the Golden Rule can identify dubious forecasts quickly and inexpensively. To date, ignorance of research findings, bias, sophisticated statistical procedures, and the proliferation of big data, have led forecasters to violate the Golden Rule. As a result, despite major advances in evidence-based forecasting methods, forecasting practice in many fields has failed to improve over the past half-century.

Keywords: analytics, bias, big data, causality, checklists, combining, elections, index method, judgmental bootstrapping, structured analogies, uncertainty.

This paper is forthcoming in *Journal of Business Research* in 2015. This working paper version is available from *GoldenRuleofForecasting.com*.

Acknowledgments: Kay A. Armstrong, Fred Collopy, Jason Dana, Peter Fader, Robert Fildes, Everette Gardner, Paul Goodwin, David A. Griffith, Nigel Harvey, Robin Hogarth, Michael Lawrence, Barbara Mellers, Mike Metcalf, Don Peters, Fotios Petropoulos, Nada R. Sanders, Steven P. Schnaars, and Eric Stellwagen provided reviews. This does not imply that all reviewers were in agreement with all of our conclusions. Geoff Allen, Hal Arkes, Bill Ascher, Bob Clemen, Shantayanan Devarajan, Magne Jørgensen, Geoffrey Kabat, Peter Pronovost, Lisa Shu, Jean Whitmore, and Clifford Winston suggested improvements. Kesten Green presented a version of the paper at the University of South Australia in May 2013, and at the International Symposiums on Forecasting in Seoul in June 2013 and in Rotterdam in June 2014. Thanks also to the many authors who provided suggestions on our summaries of their research. Hester Green, Emma Hong, Jennifer Kwok, and Lynn Selhat edited the paper. Responsibility for any errors remains with the authors.

Contact information:

J. Scott Armstrong, The Wharton School, University of Pennsylvania, 700 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, U.S.A., and Ehrenberg-Bass Institute, Adelaide; armstrong@wharton.upenn.edu.

Kesten C. Green, University of South Australia Business School, and Ehrenberg-Bass Institute, GPO Box 2471, Adelaide, SA 5064, Australia; kesten.green@unisa.edu.au.

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

Introduction

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

To answer that question, and the general question of how best to go about forecasting, this paper proposes a general rule: a *Golden Rule of Forecasting*. The short form of the Golden Rule is to *be conservative*. The long form is to *be conservative by adhering to cumulative knowledge about the situation and about forecasting methods.* Conservatism requires a valid and reliable assessment of the forecasting problem in order to make effective use of cumulative knowledge about the situation, and about evidence-based forecasting procedures.

The Golden Rule applies to all forecasting problems, but is especially important when bias is likely and when the situation is uncertain and complex. Such situations are common in physical and biological systems—as with climate, groundwater, mine yield, and species success—business—as with investment returns—and public policy—as with the effects of government projects, laws, and regulations.

Work on this paper started with a narrow conception of the application of conservatism to forecasting: reduce the amount of change that is forecast in the presence of uncertainty. That philosophy is the basis of regression analysis, which regresses toward the mean. The narrow conception created its own contradictions, however, because reducing the amount of change predicted is not conservative when a larger change is more consistent with cumulative knowledge. Consider, for example, that it would not be conservative to reduce growth forecasts for a less-developed nation that has made big reductions in barriers to trade and investment, and in the regulation of business. Deliberations on this point led to the definition of conservatism proposed for the Golden Rule. To the authors' knowledge, the foregoing definition of conservatism has not been used in the forecasting literature, but it is consistent with Zellner's description of a "sophisticatedly simple model" being one that "takes account of the techniques and knowledge in a field and is logically sound" (Zellner, 2001, p. 259).

The Golden Rule Checklist

The checklist of 28 operational guidelines provided in this article follows logically from the definition of conservatism. The checklist can help forecasters to be conservative by applying the Golden Rule.

Subsequent searches for papers with comparative evidence relevant to the 28 guidelines involved Internet literature searches, investigating references in important papers, asking key researchers, and posting requests on the Internet. Email messages were then sent to the lead authors of articles cited in substantive ways in order to check whether any relevant evidence had been overlooked and to ensure the evidence is properly summarized. Reminder messages were sent to authors who did not respond and to some co-authors. Eighty-four percent of authors for whom valid email addresses were found responded.

The unit of analysis for assessing evidence on the guidelines is the paper or chapter. While findings from individual studies are sometimes mentioned in this article, where a paper includes more than one relevant comparison these are averaged before calculating any summary figures. Averages are geometric means of error reductions. Where they are available, error reductions are based on appropriate evidence-based and intuitive measures of error (Armstrong, 2001c); often median-absolute-percentage-errors.

Table 1 shows the improvements in accuracy achieved by following a guideline relative to using a less-conservative approach. The guidelines are each denoted by a check box. The total number of papers, the number of papers providing evidence included in the calculation of the average percentage error reduction, and the error reduction accompanies each guideline. For example, for the first checklist item, seven comparisons were identified. Of those, three studies provided evidence on the relative accuracy of forecasts from evidence-based methods validated for the situation. The average error reduction is 18 percent. Almost all of the evidence identified in the searches supports the guidelines, and each of the 21 guidelines for which evidence was identified is supported by the overwhelming balance of that evidence. The last row of the Table shows that the weighted average error reduction per guideline was 31 percent. The balance of this section describes each of the guidelines and the evidence.

Table 1: Golden Rule Checklist with evidence on error reduction

		Guideline	Comparisons* N	Error reduction n	Error reduction %
1.		**Problem formulation**			
1.1		Use all important knowledge and information by…			
1.1.1	☐	selecting evidence-based methods validated for the situation	7	3	18
1.1.2	☐	decomposing to best use knowledge, information, judgment	17	9	35
1.2		Avoid bias by…			
1.2.1	☐	concealing the purpose of the forecast	–		
1.2.2	☐	specifying multiple hypotheses and methods	–		
1.2.3	☐	obtaining signed ethics statements before and after forecasting	–		
1.3	☐	Provide full disclosure for independent audits, replications, extensions	1		
2.		**Judgmental methods**			
2.1	☐	Avoid unaided judgment	2	1	45
2.2	☐	Use alternative wording and pretest questions	–		
2.3	☐	Ask judges to write reasons against the forecasts	2	1	8
2.4	☐	Use judgmental bootstrapping	11	1	6
2.5	☐	Use structured analogies	3	3	57
2.6	☐	Combine independent forecasts from judges	18	10	15
3.		**Extrapolation methods**			
3.1	☐	Use the longest time-series of valid and relevant data	–		
3.2	☐	Decompose by causal forces	1	1	64
3.3		Modify trends to incorporate more knowledge if the…			
3.3.1	☐	series is variable or unstable	8	8	12
3.3.2	☐	historical trend conflicts with causal forces	1	1	31
3.3.3	☐	forecast horizon is longer than the historical series	1	1	43
3.3.4	☐	short and long-term trend directions are inconsistent	–		
3.4		Modify seasonal factors to reflect uncertainty if…			
3.4.1	☐	estimates vary substantially across years	2	2	4
3.4.2	☐	few years of data are available	3	2	15
3.4.3	☐	causal knowledge is weak	–		
3.5	☐	Combine forecasts from alternative extrapolation methods, data	1	1	16
4.		**Causal methods**			
4.1	☐	Use prior knowledge to specify variables, relationships, and effects	1	1	32
4.2	☐	Modify effect estimates to reflect uncertainty	1	1	5
4.3	☐	Use all important variables	5	4	45
4.4	☐	Combine forecasts from dissimilar models	5	5	22
5.	☐	**Combine forecasts from diverse evidence-based methods**	15	14	15
6.	☐	**Avoid unstructured judgmental adjustments to forecasts**	4	1	64
		Totals and **Unweighted Average**	109	70	**31**

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

Problem formulation (1)

Forecasters should first formulate the forecasting problem. Proper formulation calls for use of cumulative knowledge about the situation and the selection of relevant evidence-based forecasting methods.

Use all important knowledge and information (1.1)

Use all relevant, reliable, and important information, and no more. It is important to exclude unimportant and dubious variables because their use will harm predictive validity. That is one of the major objections to the application of complex statistical techniques, or analytics, to "big data"—see, e.g. Sanders, 2014, pp. 195–196 and 204, for illustrations of the problems with this approach. To identify important information, ask a heterogeneous group of experts to independently list data sources, relevant variables, directions and strengths of the variables' effects. In addition, ask experts to justify their judgments. Search the literature for evidence about causal relationships. Especially useful are meta-analyses, where structured procedures are used to summarize the findings of experimental studies. Non-experimental data might be useful in situations where experimental data are lacking, but should be used with great caution—see Kabat 2008 on health risk studies for illustrations of problems with analysis of non-experimental data.

Conservative forecasting requires knowing the current situation, and so forecasters should seek out the most recent data. For example, to forecast demand for ice cream in Sydney in the coming week, it would be important to know that a big cruise ship was due to arrive and that a week of perfect beach weather was expected.

The need to weight recent history more heavily should not, however, lead one to conclude that things are so different now that historical data and knowledge should be ignored. Such claims should be met with demands for evidence. The mantra that the world in general or a particular situation is outside of previous experience is popular among CEOs and political leaders. U.S. President Dwight Eisenhower, for example, stated that, "Things are more like they are now than they ever were before." The belief that things are different now has led to disastrous forecasts by governments, businesses, and investors. The many and varied speculative bubbles from Dutch tulip bulbs to Dot-com stocks provide examples of the failed forecasts of investors who believed the situation was different from previous experience.

Schnaars (1989) provides many further examples. If need be, conduct experiments to assess the effects of recent changes, or identify and analyze the outcomes of analogous situations.

Use all important knowledge and information by selecting evidence-based methods validated for the situation (1.1.1)

Forecasters should use only procedures that have been empirically validated under conditions similar to those of the situation being forecast. Fortunately, there is much evidence on which forecasting methods provide the most accurate forecasts under which conditions. Evidence, derived from empirical comparisons of the out-of-sample accuracy of forecasts from alternative methods, is summarized in *Principles of Forecasting* (Armstrong, 2001d). The handbook is a collaborative effort by 40 forecasting researchers and 123 expert reviewers. Updates since the book's publication are provided at ForecastingPrinciples.com.

Do not assume that published forecasting methods have been validated. Many statistical forecasting procedures have been proposed without adequate validation studies, simply on the basis of experts' opinions. An example is a published model for forecasting sales that was tested on only six holdout observations from three different products. A reanalysis of the model's performance using a more extensive dataset, consisting of 14 products and 55 holdout observations, found no evidence that the complex utility-based model yields more accurate forecasts than a much simpler evidence-based extrapolation model (Goodwin and Meeran, 2012).

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

Forecasters should validate any method they propose against evidence-based methods. Clients should ask about independent validation testing rather than assume that it was done. For example, independent evaluations of popular commercial programs sold by Focus Forecasting concluded that these forecasts were substantially less accurate than forecasts from exponential

smoothing (Flores and Whybark, 1986; Gardner and Anderson, 1997) and damped smoothing (Gardner, Anderson-Fletcher, and Wickes, 2001).

One validated approach, Rule-based Forecasting (RBF), embodies existing knowledge about which methods work best under what conditions in the form of rules. RBF involves 99 rules for how to forecast given up to 28 conditions of time series data. For example, the method varies the weights on alternative extrapolation forecasts depending on the forecast horizon, causal forces, and variability of the historical data. The conditions also allow for the incorporation of experts' domain knowledge. RBF provided the most accurate forecasts for annual data in the M-Competition. There was a reduction in the Median Average Percentage Errors (MdAPE) of 18 percent for one-year-ahead forecasts compared to that for the equal-weights combined forecast—the next most accurate method. For six-year ahead forecasts, the error reduction versus equal-weights combining was 42 percent (Collopy and Armstrong 1992). Vokurka, Flores and Pearce (1996) provide additional support for differential weights RBF. They used automated procedures for rule selection and found that errors for 6-year-ahead forecasts of M-Competiton data were 15 percent less than those for the equal-weights combined forecasts.

Fildes and Petropoulos (this issue) also provide evidence that forecasters can reduce forecast error by using basic knowledge about the characteristics of series being forecast along with simple evidence-based rules for selecting extrapolation methods and weighting them. That approach led to a five percent error reduction in their study.

Despite the extensive evidence on forecasting methods, many forecasters overlook that knowledge. Consider the U.N. Intergovernmental Panel on Climate Change's (IPCC's) forecasts of dangerous manmade global warming (Randall et al., 2007). An audit of the procedures used to generate these forecasts found that they violated 72 of the 89 relevant forecasting principles such as "compare track records of various forecasting methods" (Green and Armstrong, 2007a). As a consequence of overlooking evidence on what forecasting procedures should be used in the situation, the IPCC used invalid forecasting methods to generate the forecasts that have been used as the basis for costly government policies.

Use all important knowledge and information by decomposing to best use knowledge, information, judgment (1.1.2)

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

There are two types of decomposition: additive and multiplicative.

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

Another additive decomposition procedure is to estimate the current status or initial value of a time series—a process that is sometimes referred to as nowcasting—and to then add a forecast of the trend. The repeated revisions of official economic data suggest much uncertainty about initial levels. For example, Runkle (1998) found that the difference between initial and revised estimates of quarterly GDP growth from 1961 to 1996 varied from 7.5 percentage points upward to 6.2 percentage points downward. Zarnowitz (1967) found that about 20 percent of the total error in predicting one-year-ahead GNP in the U.S. arose from errors in estimating the current GNP.

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

Dangerfield and Morris (1992) used exponential smoothing models to forecast all 15,753 unique series derived by aggregating pairs of the 178 monthly time-series used in the M-Competition (Makridakis et al., 1982) that included at least 48 observations in the specification set. The additive decomposition forecasts derived by combining forecasts from exponential smoothing models of the individual series were more accurate for 74 percent of two-item series. The MAPE of the bottom-up forecasts was 26 percent smaller than for the top-down forecasts. Similarly, Jørgensen (2004) finds that when seven teams of experts forecast project

completion times, the errors of bottom-up forecasts were 49 percent smaller than the errors of direct forecasts.

Carson, Cenesizoglu, and Parker (2011) forecast total monthly U.S. commercial air travel passengers for 2003 and 2004. They estimated an econometric model using data from 1990 to 2002 in order to directly forecast aggregate passenger numbers. They used a similar approach to estimate models for forecasting passenger numbers for each of the 179 busiest airports using regional data, and then added across airports to get an aggregate forecast. The mean absolute error (MAE) from the recomposed forecasts was about half that from the aggregate forecasts, and was consistently lower over horizons from one-month-ahead to 12-months-ahead. Additive decomposition enables forecasters to include information on many important variables when there are large databases. For example, Armstrong and Andress (1970) used data from 2,717 gas stations to derive a segmentation model that used 11 of an initial 19 variables selected based on domain knowledge—e.g. building age, and open 24 hours. They used the same data to estimate a stepwise regression model that included all 19 variables. The two models were used to forecast sales for 3,000 holdout gas stations. The segmentation model forecasts had a MAPE of 41 percent and provided an error reduction of 29 percent compared to the 58 percent MAPE of the regression model's forecasts. The finding is consistent with the fact that segmentations can properly incorporate more information than regression analysis.

Because data on the current level are often unreliable, forecasters should seek alternative estimates. Consider combining the latest survey data with estimates from exponential smoothing—with a correction for lag—or with a regression model's estimate of the level at t=0. Armstrong (1970), for example, estimated a cross-sectional regression model using annual sales of photographic equipment in each of 17 countries for 1960 to 1965. Backcasts were made for annual sales from 1955 to 1953. One approach started with the survey data and added the trend over time by using an econometric model. Another approach used a combination of survey data and econometric estimates of the starting values, and then added the trend. No matter what the weights, forecasts based on the combined estimates of the starting values were more accurate than forecasts based on survey data estimates of the starting values alone. The a priori weights reduced the backcast errors for 14 of the 17 countries. On average across the countries, the mean absolute percentage error (MAPE) was reduced from 30 percent to 23 percent, an error reduction of 23 percent.

Another study, on forecasting U.S. lodging market sales, examined the effect of estimating the current level and trend separately. An econometric model provided 28 forecasts from 1965 through 1971 using successive updating. The MAPE was reduced by 29 percent when the current level was based on a combination of survey data and the econometric forecast. Another test, done with forecasts from an extrapolation model, found the MAPE was reduced by 45 percent (Tessier and Armstrong, this issue).

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

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

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

Avoid bias (1.2)

Forecasters sometimes depart from prior knowledge due to unconscious biases such as optimism. Financial and other incentives, deference to authority, and confusing forecasting with planning can also cause forecasters to ignore prior knowledge or to choose methods that have not been validated.

Bias might be deliberate if the purpose of the forecasts is to further an organizational or a political objective, such as with profit forecasts to help raise capital for a risky venture, or cost-benefit forecasts for large-scale public works projects. For example, one study analyzed more

than 10,000 judgmental adjustments of quantitative model forecasts for one-step-ahead pharmaceutical sales forecasts. In 57 percent of 8,411 forecasts, the experts adjusted the forecast upwards, whereas downward adjustments occurred only 42 percent of the time. Optimism remained even after experts were informed about their bias, as the feedback decreased the rate of upward adjustments only slightly to 54 percent of 1,941 cases (Legerstee and Franses, 2013). Another 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).

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

By ensuring that forecasters are unaware of the purpose of the forecast, one can eliminate intentional biases. To implement this guideline, give the forecasting task to independent forecasters who are not privy to the purpose of the forecast.

Avoid bias by specifying multiple hypotheses and methods (1.2.2)

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

Forecasters should consider using an appropriate no-change model as a benchmark hypothesis. The no-change model is a reasonable conservative approach for many complex and highly uncertain problems. The no-change model is, however, not always conservative: There are cases where cumulative knowledge calls for change. For example, consider that you sell baked beans and have a small market share. You reduce your price by 20 percent. A no-change model for forecasting unit sales would not be conservative. You should rely instead on knowledge about the price elasticity of similar products. In other words, forecasters should test

alternative hypotheses, methods, and models to the extent that a skeptical critic would not be able to point to a plausible and important alternative that was not tested.

Given the power of the no-change model in many situations, the Relative Absolute Error (RAE) was developed to compare the accuracy of forecasts from alternative models (Armstrong and Collopy, 1992). It is the error of a forecast from a proposed model relative to that of a forecast from a credible no-change or other benchmark model. Thus, a RAE less than 1.0 means the forecasts are more accurate than the benchmark forecasts, and a RAE greater than 1.0 means the forecasts are worse than the benchmark forecasts.

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

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

Provide full disclosure for independent audits, replications, extensions (1.3)

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

Cumulative knowledge, and hence full disclosure, is vital to the Golden Rule. Failures to disclose are often due to oversights, 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 U.S. Fish and Wildlife Service researchers as the basis of their forecasts that polar bears are endangered. The researchers refused to provide the data on the grounds that they were "using them" (Armstrong, Green, and Soon 2008).

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

Finally, Weimann (1990) finds a correlation of 0.51 between comprehensive reporting of methodology—as measured by the number of methodological deficiencies reported—and the accuracy of election polls. The finding is consistent with the notion that those who report more fully on the limitations of their methodology are less biased, and thus their forecasts are more accurate.

Judgmental methods (2)

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

Avoid unaided judgment (2.1)

Use structured, validated procedures to make effective use of knowledge that is available in the form of judgment. Unaided judgment is not conservative because it is a product of faulty memories, inadequate mental models, and unreliable mental processing, to mention only a few of the shortcomings that prevent good use of judgment. As a result, when the situation is complex and uncertain, forecasts by experts using their unaided judgment are no more accurate than those of non-experts (Armstrong, 1980). Green (2005) finds that forecasts of the decisions that would be made in eight conflict situations that were obtained by using simulated interaction—a form of role-playing that involves structuring judgment—reduced error relative to unaided judgment forecasts by 45 percent.

Moreover, when experts use their unaided judgment, they tend to more easily remember recent, extreme, and vivid events. Thus, they overemphasize the importance of recent events, as was shown in a study of 27,000 political and economic forecasts made over a 20-year period by 284 experts from different fields (Tetlock 2005).

Unaided judges tend to see patterns in the past and predict their persistence, despite lacking reasons for the patterns. Even forecasting experts are tempted to depart from conservatism in this way. For example, when attendees at the 2012 *International Symposium on Forecasting*

were asked to forecast the annual global average temperature for the following 25 years on two 50-year charts, about half of the respondents drew zigzag lines (Green and Armstrong, 2014). They likely drew the zigzags to resemble the noise or pattern in the historical series (Harvey, 1995)—a procedure that is almost certain to increase forecast error relative to a straight line.

Use alternative wording and pretest questions (2.2)

The way a question is framed can have a large effect on the answer. Hauser (1975, Chapter 15) provides examples of how wording affects responses. One example was the proportion of people who answered "yes" to alternatively worded questions about free speech in 1940. The questions and the percentage of affirmative responses are: (1) *"Do you believe in freedom of speech?"* 96 percent; (2) *"Do you believe in freedom of speech to the extent of allowing radicals to hold meetings and express their views to the community?"* 39 percent. To reduce response errors, pose the question in multiple ways, pre-test the different wordings to ensure they are understood as intended, and combine the responses to the alternative questions.

Ask judges to write reasons against the forecast (2.3)

Asking judges to explain their forecasts in writing is conservative because it encourages them to consider more information and contributes to full disclosure.

Koriat, Lichtenstein, and Fischhoff (1980) asked 73 subjects to pick the correct answer to each of ten general knowledge questions and then to judge the probability that their choice was correct. For ten additional questions, the subjects were asked to make their picks and write down as many reasons for and against each pick that they could think of. Their errors were 11 percent less when they provided reasons. In their second experiment, subjects predicted the correct answers to general knowledge questions and provided one reason to support their prediction (n=66), to contradict their prediction (n=55), or both (n=68). Providing a single contradictory reason reduced error by 4 percent compared to providing no reason. Providing supporting reasons had only a small effect on accuracy.

Additional evidence was provided in an experiment by Hoch (1985). Students predicted the timing of their first job offer, the number of job offers, and starting salaries. Those who wrote reasons why their desired outcome might not occur made more accurate forecasts.

Use judgmental bootstrapping (2.4)

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

To use judgmental bootstrapping, develop a quantitative model to infer how an expert or group of experts makes forecasts. To do so, ask an expert to make forecasts for artificial cases in which the values of the causal factors vary independently of one another. Then, estimate a regression model of the expert's forecasts against the variables. A key condition is that the final model must exclude any variable that affects the forecast in a way that is opposite to what is known about causality from prior knowledge, especially experimental evidence.

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

Use structured analogies (2.5)

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

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

Research on structured analogies is in its infancy, but the findings of substantial improvements in accuracy for complex uncertain situations are encouraging. In one study, eight

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

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

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

Nikolopoulos, Litsa, Petropoulos, Bougioukos, and Khammash (this issue) test a variation of the structured analogies method: structured analogies from an interacting group. Their approach reduced the average percentage error relative to unaided judgment by 54 percent.

Combine independent forecasts from judges (2.6)

To increase the amount of information considered and to reduce the effects of biases, combine anonymous independent forecasts from judges. For example, experts can make useful predictions about how others would behave in some situations. Avoid using traditional group meetings to combine experts' forecasts. The risk of bias is high in face-to-face meetings because people can be reluctant to share their opinions in order to avoid conflict or ridicule. Managers often rely needlessly on the unaided judgments that emerge from group meetings as forecasts for important decisions. Experimental evidence demonstrates that it is easy to find

structured combining methods that produce forecasts from expert judgments that are more accurate than those from traditional group meetings (Armstrong, 2006b).

The Delphi technique is one established and validated structured judgmental forecasting method for combining experts' forecasts. Delphi is a multi-round survey that elicits independent and anonymous forecasts and reasons for them from a panel of experts. After each round, a summary of the forecasts and reasons is provided to the experts. The experts can revise their forecasts, free from group pressures, in later rounds. A review of the literature concluded that Delphi provided forecasts that were more accurate than forecasts from traditional face-to-face meetings in five studies and less accurate in one; two studies showed no difference (Rowe and Wright, 2001). A laboratory experiment involving estimation tasks found that Delphi is easier to understand than prediction markets (Graefe and Armstrong, 2011).

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

Good results can be achieved by combining forecasts from eight to twelve experts with diverse knowledge of the problem and biases that are likely to differ. Surprisingly, expertise does not have to be high, and often has little impact on forecast accuracy (Armstrong, 1980; Tetlock, 2005). Graefe (2014) finds that voters' combined expectations of who will win provided forecasts that were more accurate than the expectations of the typical individual expert, with errors 32 percent smaller across six U.S. presidential elections. Combined voter expectations were also more accurate than the single-expert complex statistical forecasts for the 2012 U.S. presidential election at FiveThirtyEight.com for all of the 100-day period leading up to Election Day. Combined voter expectations reduced MAE by an average of 38 percent.

Nikolopoulos, Litsa, Petropoulos, Bougioukos, and Khammash (this issue) obtained five forecasts about the outcomes of two government programs from a group of 20 experts using their unaided judgment, and from groups of experts using either semi-structured analogies or the Delphi method. The two structured approaches to combining judgmental forecasts reduced average percentage error relative to unaided judgment by eight and 27 percent.

In some situations, people are experts about their own behavior. The standard method for combining judgments of one's likely behavior is an intentions survey. There is a close relationship between intentions and behavior as shown in the meta-analysis by Kim and Hunter (1993), especially for high-involvement decisions (Morwitz, 2001). Here, again, it is harmful to make judgmental revisions (Wright and McRae, 2007).

Forecasts from intentions surveys are more accurate when they are very short-term predictions about important events. For example, while polls that ask people who they intend to vote for have no predictive value for long-term forecasts, they are highly accurate shortly before Election Day (Erikson and Wlezien, 2012).

Extrapolation methods (3)

Extrapolation for forecasting is in part conservative because it is based on data about past behavior. Extrapolation can be used with time-series data or cross-sectional data. For an example of the latter, behavioral responses to gun law changes in some states can be used to predict responses in other states.

Extrapolation ceases to be conservative when knowledge about the situation that is not contained in the time-series or cross-sectional data is at odds with the extrapolation. Thus, there have been attempts to incorporate judgments into extrapolation. This section examines approaches to incorporating more knowledge into extrapolations.

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

This guideline is based on the logic of the Golden Rule. The alternative of selecting a particular starting point for estimating a time-series forecasting model, or of selecting a specific subset of cross-sectional data, allows the forecaster considerable influence over the forecast that will result. For example, McNown, Rogers, and Little (1995) showed that an extrapolation model predicted increases in fertility when based on five years of historical data, but declines in fertility when based on 30 years of data. Similar findings had been published earlier. For example, Dorn's (1950) review of research on population forecasts led him to conclude that they were insufficiently conservative due to an over emphasis on recent trends.

By using the longest obtainable series, or all obtainable cross-sectional data, one reduces the risk of biasing forecasts, whether intentionally or unintentionally.

Decompose by causal forces (3.2)

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

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

Armstrong, Collopy, and Yokum (2005) test the value of decomposition by causal forces for twelve annual time-series of airline and automobile accidents, airline revenues, computer sales, and cigarette production. Decomposition was hypothesized to provide more accurate forecasts than those from extrapolations of the global series if each of the components could be forecast over a simulation period with less error than could the aggregate, or if the coefficient of variation about the trend line of each of the components was less than that for the global series. Successive updating produced 575 forecasts, some for forecast horizons of one-to-five-years and some for horizons of one-to-10-years. For the nine series that met one or both of the two conditions, forecasting the decomposed series separately reduced the Median Relative Absolute Error (MdRAE) of the combined forecasts by a geometric mean average of 64 percent relative to forecasts from extrapolating the global series. (The original text of that paper states the error reduction was 56 percent, but that is a typographical error.)

Modify trends to incorporate more knowledge (3.3)

Extrapolate conservatively by relying on cumulative knowledge about the trend. In many situations, conservatism calls for a reduction in the magnitude of the trend, which is commonly referred to as damping. Damping keeps the forecasts closer to the estimate of the current situation. However, damping might not be conservative if it were to lead to a substantial departure from a consistent long-term trend arising from well-supported and persistent causal forces. For example, Moore's Law, which states that computer performance doubles roughly every two years, has held up for more than half a century, and there is reason to expect that the causal forces will persist (Mollick 2006). Thus, modifying to incorporate more knowledge can also involve moving a short-term trend toward a long-term trend. Without strong evidence that the causal factors had changed, forecasts derived assuming a weakening of Moore's Law would not be conservative.

Damping is also not conservative for situations in which an important change in causal forces is expected to increase a trend, as might be caused by a substantial reduction in corporate taxes, elimination of a tariff, or introduction of a substantially improved product. The following guidelines help to identify situations where modifying trends is conservative.

Modify trends… if the series is variable or unstable (3.3.1)

Variability and stability can be assessed by statistical measures or judgmentally—or both. Most of the research to date uses statistical measures.

In a review of ten papers, Armstrong (2006a) concludes that damping the trend by using only statistical rules on the variability in the historical data yielded an average error reduction of about 4.6 percent. A reanalysis of the papers using the procedures of this review finds eight of the papers (Gardner and McKenzie, 1985; Makridakis et al., 1982; Makridakis and Hibon, 2000; Gardner, 1990; Schnaars, 1986; Gardner and Anderson, 1997; Miller and Liberatore, 1993; Fildes, Hibon, Makridakis, and Meade, 1998) include relevant evidence on error reduction from damping trends when forecasting by extrapolation. The average error reduction across the eight papers is 12 percent. In all but one of the papers, accuracy was improved by damping.

In his review of research on exponential smoothing, Gardner (2006) concludes that "...it is still difficult to beat the application of a damped trend to every time series" (p. 637). Since the

gains can be achieved easily and without any intervention, the adoption of the damped-trend exponential smoothing method would lead to substantial savings for production and inventory control systems worldwide. Further gains in accuracy can be achieved by modifying trends to incorporate knowledge about the situation and expert judgment in structured ways as the following guidelines describe.

Modify trends… if the historical trend conflicts with causal forces (3.3.2)

If the causal forces acting on a time-series conflict with the observed trend in a time-series, a condition called a contrary series, damp the trend heavily toward the no-change forecast. To identify casual forces, ask a small group of experts (three or more) for their assessment and adopt the majority judgment. Experts typically need only a minute or so to assess the causal forces for a given series, or for a group of related series.

Causal forces may be sufficiently strong as to reverse a long-term trend, such as when a government regulates an industry. In that case, one would expect the iron law of regulation to prevail (Armstrong and Green, 2013) with consequent losses of consumer welfare as Winston (2006) finds.

Research findings to date suggest a simple guideline that works well for contrary series: *ignore trends*. Armstrong and Collopy (1993) apply this contrary-series guideline to forecasts from Holt's exponential smoothing—which takes no account of causal forces. Twenty annual time-series from the M-Competition were rated as contrary. By removing the trend term from Holt's exponential smoothing, the median average percentage error (MdAPE) was reduced by 18 percent for one-year-ahead forecasts, and by 40 percent for six-year-ahead forecasts. Additional testing used contrary series from four other data sets: annual data on Chinese epidemics, unit product sales, economic and demographic variables, and quarterly data on U.S. Navy personnel numbers. On average, the MdAPE for the no-trend forecasts was 17 percent less than Holt's for 943 one-step-ahead forecasts. For 723 long-range forecasts, which were six-ahead for annual and 18-ahead for quarterly data, the error reduction averaged 43 percent over the four data sets, a geometric mean of 31.4 percent across all 10 comparisons.

Modify trends… if the forecast horizon is longer than the historical series (3.3.3)

Uncertainty is higher when the forecast horizon is longer than the length of the historical time-series. If making forecasts in such a situation cannot be avoided, consider (1) damping the

trend toward zero as the forecast horizon increases, and (2) averaging the trend with trends from analogous series. U.S. Fish and Wildlife Service scientists violated this guideline and overlooked the need for damping when they used only five years of historical data to forecast an immediate, strong, and long-term reversal in the trend of polar bear population numbers (Armstrong, Green, and Soon, 2008).

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

Modify trends… if the short- and long-term trend directions are inconsistent (3.3.4)

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

Modify seasonal factors to reflect uncertainty (3.4)

When the situation is uncertain, seasonal adjustment can harm accuracy as was shown long ago by, for example, Groff (1973), and Nelson (1972). Having only a few years of data, large variations in the estimates of seasonal factors from one year to the next, and ignorance about what might cause seasonality are all sources of uncertainty.

One conservative response to uncertainty about seasonality is to damp seasonal factors toward 1.0. That approach has been the most successful one to date. Other approaches to consider are to combine the estimate of a seasonal factor with those for the time period before and the period after; and to combine the seasonal factors estimated for the target series with those estimated for analogous series. The two latter approaches incorporate more information, and might therefore improve upon a damping approach based only on statistical relationships.

Modify seasonal factors… if estimates vary substantially across years (3.4.1)

If estimates of the size of seasonal factors vary substantially from one year to the next, this suggests uncertainty. Variations might be due to shifting dates of major holidays, strikes, natural disasters, irregular marketing actions such as advertising or price reductions, and so on.

To deal with variations in seasonal factor estimates, damp the estimates or use the average of each seasonal factor with those from the time periods immediately before and after.

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

Chen and Boylan (2008) test seasonal factor damping procedures on 218 monthly series of lightbulb sales. They found that two damping procedures on average reduced the error—symmetrical MAPE—of forecasts for all but one of 12 combinations of estimation period—two, three, and four year—and forecast horizon—one, three, six, and nine month. Calculating from Chen and Boylan's Table 6, average error reduction was 3.1 percent.

Modify seasonal factors… if few years of data are available (3.4.2)

Damp seasonal factors strongly—or perhaps avoid using them—unless there are sufficient years of historical data from which to estimate them. Chen and Boylan (2008) find that seasonal factors harmed accuracy when they were estimated from fewer than three years of data.

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

In an analysis of 44 series of retail sales data from a large U.K. department store chain, Bunn and Vassilopoulos (1999) find that forecasts from models that used seasonal factors estimated from analogous series were consistently more accurate than forecasts from models that used seasonal factors estimated from the target series data alone. When analogies were from the

same business class as the target series, the reductions in the Mean Absolute Deviation errors (MADs) compared to forecasts from standard seasonal adjustment were between eight and 25 percent, depending on the model used.

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

Modify seasonal factors… if causal knowledge is weak (3.4.3)

Without prior knowledge on the causes of seasonality in the series to be forecast, seasonal factors are likely to increase forecasting error. To the extent that the causal knowledge is weak, damp the factors toward cumulative knowledge on seasonality. If there is no established causal basis for seasonality, do not use seasonal factors.

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

Armstrong (2001b, p. 428) finds error reductions from combining forecasts from different extrapolation methods in five studies. The error reductions ranged from 4.3 to 24.2 percent, with an average of 16 percent.

Analogous time-series can provide useful information for extrapolation models. The information is relevant for levels—or base rates for cross-sectional data—and for trends. For example, consider that one wishes to forecast sales of the Hyundai Genesis automobile. Rather than relying only on the Genesis sales trend data, use the data for all luxury cars to forecast the trend, and then combine the two forecasts.

Causal methods (4)

Regression analysis is currently the most common approach for developing and estimating causal models. The method is conservative in that it regresses to the mean value of the series in response to unattributed variability in the data. However, regression analysis has characteristics that limit its usefulness for forecasting.

Regression is not sufficiently conservative because it does not reflect uncertainty regarding causal effects arising from omitted variables, predicting the causal variables, changing causal relationships, and inferred causality if variables in the model correlate with important excluded variables over the estimation period. In addition, using statistical significance tests and

sophisticated statistical methods to help select predictor variables is problematic when large databases are used. That is because sophisticated statistical techniques and an abundance of observations tend to seduce forecasters and their clients away from using cumulative knowledge and evidence-based forecasting procedures. In other words, they lead forecasters to ignore the Golden Rule. For a more detailed discussion of problems with using regression analysis for forecasting, see Armstrong (2012a), and Soyer and Hogarth (2012).

Use prior knowledge to specify variables, relationships, and effects (4.1)

Scientific discoveries about causality were made in the absence of sophisticated statistical analyses. For example, John Snow identified the cause of cholera in London in 1854 as a result of "the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data" (Freedman 1991, p. 298).

Only variables that are known to be related to the variable to be forecast should be included in a model. Ideally, variables should be identified from well established theory—e.g., price elasticities for normal goods—obvious relationships— e.g., rainfall and crop production—or experimental evidence. For simple problems, one might use statistical analysis of non-experimental data, but valid causal relationships cannot be discovered in this way for complex problems.

A priori analyses to obtain knowledge, and to specify variables, relationships, and effects, can be time consuming, expensive, and difficult. Finding and understanding the relevant research is necessary. Perhaps unsurprisingly, then, since the middle of the Twentieth Century, forecasters have turned to sophisticated statistical procedures such as stepwise regression and data mining, along with large databases and high-speed computers, in the hope that these would replace the need for a priori analyses. Ziliak and McCloskey (2004) provide evidence for this trend with their analysis of papers that were published in the *American Economic Review* in the 1980s and then in the 1990s. While 32 percent chose variables solely on the basis of statistical significance in the 1980s, 74 percent did so in the 1990s.

There is little reason to believe that statistical analyses will lead to better forecasting models. Consider the case of data mining. Data mining involves searching for relationships in data without a priori analysis. Academic literature on data mining goes back many decades. A Google Scholar Search for "data mining" and "predict or forecast" at the end of December

2014 produced about 175,000 hits. Two of the leading books on data mining have each been cited more than 23,000 times. Despite the efforts in support of data mining, comparative studies that show data mining provides substantive and consistent improvements in forecast accuracy are lacking.

Keogh and Kasetty (2003) conduct a comprehensive search for empirical studies on data mining. They criticize the failure of data mining researchers to test alternative methods. To address the lack of testing, they tested procedures from more than 25 papers on data mining on 50 diverse empirical data sets. In a personal correspondence with the first author of this article, Keogh wrote:

"[Professor X] claimed to be able to do 68% accuracy. I sent them some 'stock' data and asked them to do prediction on it, they got 68% accuracy. However, the 'stock' data I sent them was actually random walk! When I pointed this out, they did not seem to think it important. The same authors have another paper in [the same journal], doing prediction of respiration data. When I pointed out that they were training and testing on the same data and therefore their experiments are worthless, they agreed (but did not withdraw the paper). The bottom line is that although I read every paper on time-series data mining, I have never seen a paper that convinced me that they were doing anything better than random guessing for prediction. Maybe there is such a paper out there, but I doubt it."

More than ten years later, the authors of this article asked Keogh for an update. He responded on 15 January 2015, "I have never seen a paper that convinces me that the data mining (big data) community are making a contribution to forecasting (although I have seen papers that make that claim)."

Statistical analyses of non-experimental data are unlikely ever to successfully replace a priori analyses of experimental data.

Armstrong (1970) tests the value of a priori analysis in his study on forecasting international camera sales. A fully specified model was developed from prior knowledge about causal relationships before analyzing data. Data from 1960 to 1965 for 17 countries were then used to estimate regression model coefficients. The final model coefficients were calculated as an average of the a priori estimates and regression coefficients, a process later referred to as a poor man's Bayesian regression analysis. To test the predictive value of the approach, the model was used to forecast backwards in time (backcast) 1954's camera sales. Compared to forecasts from

a benchmark regression model with statistically estimated coefficients, forecasts from the model with coefficient estimates that included the a priori knowledge reduced MAPE by 23 percent. Another test estimated models using 1960 to 1965 data for 19 countries. The models were then used to predict market size in 11 holdout countries. The models that used *a priori* knowledge in estimating coefficients reduced the MAPE of forecasts by 40 percent.

Economists and other social scientists concerned with specifying relationships use elasticities to summarize prior knowledge. Elasticities are unit-free and easy to understand. They represent the percentage change that occurs in the variable to be forecast in response to a one-percent change in the causal variable. For example, a price elasticity of demand of -1.5 would mean that if the price were increased by 1 percent, all else being equal, one would expect unit sales to go down by 1.5 percent. Forecasters can examine prior research in order to estimate elasticities and their plausible lower and upper bounds for the situation they are concerned with. For example, in forecasting sales, one can find income, price, and advertising elasticities for various product types in published meta-analyses. If little prior research exists, obtain estimates by surveying domain experts.

Modify effect estimates to reflect uncertainty (4.2)

Causal variable coefficients should be modified in the direction of having no effect when uncertainty about the effect that variables have on the dependent variable is high and when the forecaster is uncertain about how much the causal variables will change. Modification of that kind is referred to as damping, or shrinkage. In general, the greater the uncertainty, the greater should be the damping.

Another strategy for addressing uncertainty over relationships is to adjust the weights of the causal variables so that they are more equal with one another, in other words to adjust the variable coefficients towards equality. Equalizing requires expressing the variables as differences from their mean divided by their standard deviation—i.e., as standardized variables—estimating the model coefficients using regression analysis, and then adjusting the estimated coefficients toward equality. When uncertainty about relative effect sizes is high, consider assigning equal-weights to the standardized variables, which is the most extreme case of equalizing. Dana and Dawes (2004) analyze the relative predictive performance of regression and equal-weights models for five real non-experimental social science datasets and a large number of synthetic datasets. The regression weights models failed to yield forecasts

that were more accurate than those from equal-weights models, except for sample sizes larger than one hundred observations per predictor and situations in which prediction error was likely to be very small—i.e., adjusted-$R^2 > .9$. The optimal approach most likely lies in between these two methods, statistically optimal and equal, and so averaging the forecasts from an equal-weights model and a regression model is a sensible strategy.

As Graefe (this issue) summarizes, much evidence since the 1970s shows that equal-weights models often provide more accurate ex ante forecasts than do regression models. Graefe's article also provides evidence from U.S. presidential election forecasting. Equal-weights variants of nine established regression models yielded forecasts that were more accurate for six of the nine models. On average, the equal-weights models' forecasts reduced the MAE compared to the original regression models' forecasts by five percent.

Use all important variables (4.3)

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

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

To develop an index model, use prior knowledge to identify all relevant variables and their expected directional influence on whatever is being forecast—e.g., a candidate's performance in a job. Ideally one should develop an index model using knowledge gained by reviewing experimental studies. In fields where experimental studies are scarce, survey independent experts who, among them, have diverse knowledge. Calculate an index score by determining the values of variables for a situation of interest and then adding the values. Consider using

different weights for the variables only if there is strong prior evidence that the variables have differential effects. The index score is then used to calculate the forecast. For selection problems, the option with the highest score is favored. For numerical forecasts, use a simple linear regression model to estimate the relationship between the index score and the variable to be predicted—e.g., box-office sales of a new movie.

Consider the problem of predicting judges' decisions in court cases—a situation for which the causal variables are determined by the relevant law. Kort (1957) uses the index method to test the predictability of U.S. Supreme Court decisions on right-to-counsel cases. Kort selected right-to-counsel cases because experts considered the Court's decisions on these cases to be unpredictable. Kort's review of the law led to the identification of 26 key variables, for example the youth and the literacy of the offender. He assigned the variables importance values—weights—based on an analysis of 14 cases decided between 1932 and 1947. Kort then tested the resulting model by forecasting the decisions made in 14 out-of-sample cases decided between 1947 and 1956. The model's index scores accurately forecast 12 decisions. Two decisions were too close to call based on the index scores.

The index method has also been used to forecast U.S. presidential elections, a situation with knowledge about a large number of causal variables. An index model based on 59 biographical variables correctly predicted the winners in 28 of 30 U.S. presidential elections up through 2012 (Armstrong and Graefe, 2011). For the four elections from 1996, the many-variable biographical model provided forecasts that reduced MAE by 37 percent compared to the average econometric model; all of which included few causal variables.

Another index model was based on surveys of how voters expected U.S. presidential candidates to handle up to 47 important issues. The model correctly predicted the election winner in 10 of the 11 elections up to 2012 (Graefe and Armstrong, 2013). For the three elections from 2000, issues-index model forecast errors (MAEs) were 50 percent smaller than the average econometric model forecast error.

Graefe (this issue) creates an index model by adding the standardized values of all 29 variables that were used by nine established models for forecasting the results of U.S. presidential election. Across the 10 elections from 1976 to 2012, the errors of the forecasts from that index model were 48 percent smaller than the errors of the typical individual

regression model forecasts, and were 29 percent smaller than the errors of the forecasts from the most accurate individual model.

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

Combine forecasts from dissimilar models (4.4)

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

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 MAE that was 30 percent less than that of the typical model (Graefe, Armstrong, Jones Jr., and Cuzán, 2014).

Combine forecasts from diverse evidence-based methods (5)

Combining forecasts from evidence-based methods is conservative in that more knowledge is used, and the effects of biases and mistakes such as data errors, computational errors, and

poor model specification are likely to offset one another. Consequently, combining forecasts reduces the likelihood of large errors. Equally weighting component forecasts is conservative in the absence of strong evidence of large differences in out-of-sample forecast accuracy from different methods.

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

A meta-analysis by Armstrong (2001b, p. 428) finds 11 studies on the effect of averaging forecasts from different methods. On average, the errors of the combined forecasts were 11.5 percent lower than the average error of the component forecasts. More recent research on U.S. presidential election forecasting (Graefe, Armstrong, Jones, and Cuzán, 2014) finds much larger gains when forecasts are combined from different evidence-based methods that draw upon different data. Averaging forecasts within and across four established election-forecasting methods (polls, prediction markets, expert judgments, and regression models) yielded forecasts that were more accurate than those from each of the component methods. Across six elections, the average error reduction compared to the typical component method forecast error was 47 percent.

Many scholars have proposed methods for how to best weight the component forecasts. However, Clemen's (1989) review of over 200 published papers from the fields of forecasting, psychology, statistics, and management science concluded that simple averages—i.e., using equal-weights—usually provides the most accurate forecasts.

Graefe, Küchenhoff, Stierle, and Riedl (2014) find that simple averages provide forecasts that are more accurate than those from Bayesian combining methods in four of five studies on economic forecasting, with an average error reduction of five percent. Their study also provides new evidence from U.S. presidential election forecasting, where the error of the simple average forecasts were 25 percent less than the error of the Bayesian Model Averaging forecasts. A study that tested the range of theoretically possible combinations finds that easily understood and implemented heuristics, such as take-the-average, will, in most situations, perform as well as the rather complex Bayesian approach (Goodwin, this issue).

Avoid unstructured judgmental adjustments to forecasts (6)

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

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

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

Most forecasting practitioners expect that judgmental adjustments will lead to error reductions of between five and 10 percent (Fildes and Goodwin, 2007). Yet little evidence supports that belief. For example, Franses and Legerstee (2010) analyze the relative accuracy of forecasts from models and forecasts that experts had subsequently adjusted for 194 combinations of one-step-ahead forecasts in 35 countries and across seven pharmaceutical product categories. On average, the adjusted forecasts were less accurate than the original model forecasts in 57 percent of the 194 country-category combinations.

Judgmental adjustments that are the product of structured procedures are less harmful. In an experiment by Goodwin (2000), 48 subjects reviewed one-period ahead statistical sales forecasts. When no specific instructions were provided, the subjects adjusted 85 percent of the statistical forecasts; the revised forecasts had a median absolute percentage error (MdAPE) of

10 percent. In comparison, when subjects were asked to justify any adjustments by picking a reason from a pre-specified list, they adjusted only 35 percent of the forecasts. The MdAPE was 3.6 percent and thus 64 percent less than the error of the unstructured adjustment. In both cases, however, the judgmental adjustments yielded forecasts that were 2.8 percent less accurate than the original statistical forecasts.

Judgmental adjustments should only be considered when the conditions for successful adjustment are met and when bias can be avoided (Goodwin and Fildes, 1999; Fildes, Goodwin, Lawrence, and Nikolopoulos, 2009). In particular, accuracy-enhancing judgmental adjustments may be possible when experts have good knowledge of important influences not included in the forecasting model such as special events and changes in causal forces (Fildes and Goodwin, 2007). Estimates of the effects should be made in ignorance of the model forecasts, but with knowledge of what method and information the model is based upon (Armstrong and Collopy, 1998; Armstrong, Adya, and Collopy, 2001). The experts' estimates should be derived in a structured way (Armstrong and Collopy, 1998), and the rationale and process documented and disclosed (Goodwin, 2000). In practice, documentation of the reasons for adjustments is uncommon (Fildes and Goodwin, 2007). The final forecasts should be composed from the model forecasts and the experts' adjustments. Judgmental adjustment under these conditions is conservative in that more knowledge and information is used in the forecasting process. Sanders and Ritzman (2001) found that subjective adjustments helped in six of the eight studies in which the adjustments were made by those with domain knowledge, but in only one of the seven studies that involved judges who lacked domain knowledge.

Discussion

Checklists are useful when dealing with complex problems. Unaided judgment is inadequate for analyzing the multifarious aspects of complex problems. Checklists are of enormous value as a tool to help decision-makers working in complex fields. Think of skilled workers involved with, for example, manufacturing and healthcare (Gawande, 2010).

In their review of 15 studies on the use of checklists in healthcare, Hales and Provonost (2006) find substantial improvements in outcomes in all studies. For example, an experiment on avoiding infection in intensive care units of 103 Michigan hospitals required physicians to follow five rules when inserting catheters: (1) wash hands, (2) clean the patient's skin, (3) use

full-barrier precautions when inserting central venous catheters, (4) avoid the femoral site, and (5) remove unnecessary catheters. Adhering to this simple checklist reduced the median infection rate from 2.7 per 1,000 patients to zero after three months. Benefits persisted 16 to 18 months after the checklist was introduced, and infection rates decreased by 66 percent.

Another study reports on the application of a 19-item checklist to surgical procedures on thousands of patients in eight hospitals in cities around the world. Following the introduction of the checklist, death rates declined by almost half, from 1.5 to 0.8 percent, and complications declined by over one-third, from 11 to seven percent (Haynes, Weiser, Berry, Lipsitz, Breizat, and Dellinger, 2009).

Given the effects of mistakes on human welfare, making decisions about complex problems without the aid of a checklist when one is available is foolish. In fact, organizations and regulators often require the use of checklists and penalize those who fail to follow them. The completion of an aviation checklist by memory, for example, is considered a violation of proper procedures.

Checklists should follow evidence. For example, evidence is identified for 21 of the 28 guidelines in the Golden Rule checklist. The other seven guidelines are logical consequences of the Golden Rule's unifying theory of conservatism in forecasting. Checklists based on faulty evidence or faulty logic might cause harm by encouraging users to do the wrong thing and to do so more consistently. Checklists that omit critical items risk doing more harm than good in the hands of trusting users.

Even a comprehensive evidence-based checklist might be misapplied. To reduce the effects of biases, omissions, and misinterpretations, ask two or more people to apply the checklist to the problem independently. Select people who are likely to be unbiased and ask them to sign a statement declaring that they have no biases pertaining to the problem at hand.

Computer-aided checklists are especially effective. Boorman (2001) finds that they decreased errors by an additional 46 percent as compared to paper checklists. With that in mind, a computer-aided Golden Rule checklist is available at no cost from goldenruleofforecasting.com.

The Golden Rule has face validity in that forecasting experts tend to agree with the guidelines of the Golden Rule Checklist. In a survey of forecasting experts conducted while this article was being written, most respondents stated that they typically follow or would

consider following all but three of the guidelines. The guidelines that most experts disagreed with were 1.2.1/1.2.2—which were originally formulated as one guideline: "specify multiple hypotheses or conceal the purpose of the forecast"—and 2.6—"use structured analogies". The survey questionnaire and responses are available at goldenruleofforecasting.com.

Table 2 summarizes the evidence on conservatism by type of forecasting method. There are at least 12 papers providing evidence for each method, 105 papers in total for all methods including combining, but excluding the guideline on unstructured judgmental adjustments (6). Conservatism is found to improve or not harm forecast accuracy in 102 or 97 percent of the 105 papers. Rejecting conservative procedures increased error for the type of method by between 25 percent, for extrapolation methods, and roughly 45 percent for both problem formulation and causal methods. In other words, no matter what type of forecasting method is appropriate for the forecasting problem, formulating the problem and implementing forecasting methods in accordance with the relevant conservative guidelines will avoid substantial error.

Table 2: Evidence on accuracy of forecasts from conservative procedures by method type

	----- Number of Comparisons -----			
Method type	**Total papers**	**Conservative better or similar**	**Effect size**	**Error increase vs conservative (%)**
Problem formulation	25	25	12	45
Judgmental	36	34	16	36
Extrapolative	17	16	16	25
Causal	12	12	11	44
Combined	15	15	14	18
All method types	**105**	**102**	**69**	**33**
Weighted average*				**32**

*Weighted by total papers

Table 3 summarizes the evidence to date on the Checklist guidelines. All the evidence is consistent with the guidelines provided in the Checklist, and the gains in accuracy are

large on average. Details on how these improvements were assessed are provided in a spreadsheet available from goldenruleofforecasting.com.

Table 3: Evidence on the 28 Golden Rule Guidelines

Evidence available on	21
Effect size reported for	20
More than one paper with effect size comparison	15
Range of error reductions	-17 – 71%
Average error reduction per guideline	31%

There are, however, gaps in the evidence. For example, no evidence was found for seven of the guidelines, and five guidelines were supported by single comparisons only. Research on those guidelines would likely improve knowledge on how to most effectively implement conservatism in forecasting.

Tracking down relevant studies is difficult, so there are likely to be more than the 109 papers with experimental comparisons identified in this article. Surely, then, new or improved ways of being conservative can be found and improvements can be made in how and when to apply the guidelines.

Current forecasting practice

Pop management books on forecasting appear over the years, and usually claim that forecasting is predestined to fail. The books are often popular, but are the claims that forecasting is impossible true?

No, they are not.

Substantial advances have been made in the development and validation of forecasting procedures over the past century. That is evident, for example, in the astonishing improvements summarized in Table 3. Improvements in forecasting knowledge have, however, had little effect on practice in some areas. For example, in his review of forecasting for population, economics, energy, transportation, and technology, Ascher (1978) concludes that forecast accuracy had not improved over time. Similar findings have been obtained in agriculture (Allen 1994); population (Booth 2006, and Keilman 2008); sales (McCarthy, Davis, Golicic, and Mentzer, 2006); and public transportation (Flyvbjerg, Skamris, Holm, and Buhl, 2005).

In other areas, forecasting practice has improved. Weather, sports, and election forecasting are examples. Why does progress occur in some areas and not others?

The answer appears to be that practitioners in many areas fail to use evidence-based forecasting procedures. That neglect might be due to ignorance of proper forecasting procedures, or to the desire to satisfy a client with a forecast that supports a pre-determined decision.

Bias can be introduced by forecasters as well as by clients. For example, once a forecasting method is established, those who benefit from the status quo will fight against change. This occurred when Billy Beane of the Oakland Athletics baseball team adopted evidence-based forecasting methods for selecting and playing baseball players. The baseball scouts had been making forecasts about player performance using their unaided judgment, and they were incensed by Beane's changes. Given the won-lost records, it soon became obvious that Oakland won more games after the team had implemented evidence-based selection. The change is described in Michael Lewis's book, *Moneyball*, and depicted in the movie of the same name. Most sports teams have now learned that they can either adopt evidence-based forecasting procedures for selecting players, or lose more games (Armstrong, 2012b).

Statisticians may have a biased influence on forecasting research and practice in that their skills lead them to prefer complex statistical methods and large databases. That bias would tend to lead them to depart from cumulative knowledge about forecasting methods and domain knowledge about the problem. The authors of this paper have between them about 75 years of experience with forecasting, and they have done many literature reviews, yet they have not found evidence that complex statistical procedures can produce consistent and reliable improvements in the forecast accuracy relative to conservative forecasts from simple validated procedures (Green and Armstrong, this issue).

Forecasters are more motivated to adopt evidence-based methods when they work in a field in which there is competition, in which the forecasting is repetitive rather than one-off, and in which forecast errors are salient to forecast users. Such fields include sports betting, engineering, agriculture, and weather forecasting. Weather forecasters, for example, are well calibrated in their short-term forecasts (Murphy and Winkler, 1984). In another example, independently prepared forecasts of U.S. presidential election vote shares are unbiased and extremely accurate (Graefe et al., 2014).

How to use the Golden Rule Checklist to improve forecasting practice

The Golden Rule Checklist provides evidence-based standards for forecasting procedures. Using the Checklist requires little training—intelligent people with no background in forecasting can use it. Clients can require that forecasters use the checklist in order to fulfill their contract. Clients can also rate the forecasting procedures used by forecasters against the checklist. As an additional safeguard against bias, clients can ask independent raters to rate the forecasting procedures used by forecasters against the checklist. Taking that extra step helps to guard against violations of the Golden Rule by the client, as well as by the forecaster.

If the client is unable to assess whether the forecaster followed the guidelines in the Checklist, the client should reject the forecasts on the basis that the forecaster provided inadequate information on the forecasting process. If guidelines were violated, clients should insist that the forecaster corrects the violations and resubmits forecasts.

The accuracy of forecasts should be judged relative to those from the next best method or other evidence-based methods—with errors measured in ways that are relevant to decision makers—and not by reference to a graphical display. The latter can easily be used to suggest that the forecasts and outcomes are somewhat similar.

Software providers could help their clients avoid violations of the Golden Rule by implementing the Checklist guidelines as defaults in forecasting software. For example, it would be a simple and inexpensive matter to include the contrary-series rule (3.3.2), and to avoid using seasonal factors if there are fewer than three years of data (3.4.2).

The checklist can be applied quickly and at little expense. With about two hours of preparation, analysts who understand the forecasting procedure should be able to guard against forecasts that are unconservative. The goal of the Checklist is to ensure that there are no violations. Remember that even a single violation can have a substantial effect on accuracy. On average, the *violation of a typical guideline* increases the forecast error by 44 percent.

When bad outcomes occur in medicine, doctors are often sued if they failed to follow proper evidence-based procedures. In engineering, aviation, and mining a failure to follow proper procedures can lead to lawsuits even when damages have not occurred. The interests of both clients and forecasters would be better served if clients insisted that forecasters use the

evidence-based Golden Rule Checklist, and that they sign a document to certify that they did so.

Conclusions

The first paragraph of this paper asked how a decision maker should evaluate a forecast. This article proposes following the Golden Rule. The Golden Rule provides a unifying theory of forecasting: Be conservative by adhering to cumulative knowledge about the situation and about forecasting methods. The theory is easy to understand and provides the basis for a checklist that forecasters and decision-makers can use to improve the accuracy of forecasts and to reject forecasts that are likely to be biased and dangerously inaccurate.

The Golden Rule Checklist provides easily understood guidance on how to make forecasts for any situation. The 28 guidelines in the Checklist are simple, using the definition of simplicity provided by Green and Armstrong (this issue).

Use of the Golden Rule Guidelines improves accuracy substantially and consistently no matter what is being forecast, what type of forecasting method is used, how long the forecast horizon, how much data are available, how good the data are, or what criteria are used for accuracy. The Golden Rule is especially useful for situations in which decision makers are likely to be intimidated by forecasting experts.

The error reduction from following a single guideline—based on experimental comparisons from 70 papers—ranged from four to 64 percent and averaged 31 percent. In other words, violating a single guideline typically increased forecast error by 44 percent. Imagine the effect of violating more than one guideline.

The Golden Rule makes scientific forecasting comprehensible and accessible to all: Analysts, clients, critics, and lawyers should use the checklist to ensure that there are no violations of the Golden Rule.

References

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

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*, *83*(June/July), 18–24.

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

Armstrong, J. S. (2001a). 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. (2001b). 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. (2001c). Evaluating forecasting methods. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 443–472). New York: Springer.

Armstrong, J. S. (2001d). *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*, 3–8.

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

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

Armstrong, J. S. (2012b). Predicting job performance: The moneyball factor**.** *Foresight: The International Journal of Applied Forecasting*, *25*, 31–34.

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., Du, R., Green, K. C., & Graefe, A. (2014). Predictive validity of evidence-based advertising principles. Working Paper. Available at https://marketing.wharton.upenn.edu/files/?whdmsaction=public:main.file&fileID=6794.

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

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

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

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

Boorman, D. (2001). Today's electronic checklisys reduce likelihood of crew errors and help prevent mishaps. *International Civil Aviation Organization Journal*, *1*, 17–36.

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

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

Carson, R. T., Cenesizoglu, T., & Parker, R. (2011). Forecasting (aggregate) demand for US commercial air travel. *International Journal of Forecasting, 27*, 923–941.

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. (1992). 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.

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

Dorn, H. F. (1950). Pitfalls in population forecasts and projections. *Journal of the American Statistical Association*, *45*, 311–334.

Erikson, R. S., & Wlezien, C. (2012). *The timeline of presidential elections: How campaigns do (and do not) matter*. Chicago: University of Chicago Press.

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

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*(1), 3–23.

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

Fildes, R., Hibon, M, Makridakis, S., & Meade, N. (1998). Generalizing about univariate forecasting methods: Further empirical evidence. *International Journal of Forecasting*, *14*, 339–358.

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.

Fildes, R., & Petropoulos, F. (2015). An evaluation of simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, [this issue], xxx–yyy.

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.

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2005). How (in)accurate are demand forecasts in public works projects?: The case of transportation. *Journal of the American Planning Association, 71*, 131–146.

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

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

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

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

Gardner, E. S., Jr. (1990). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*, 1–28.

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

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

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

Gardner, E. S. Jr., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, *31*, 1237–1246.

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

Goodwin, P. (2015). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. *Journal of Business Research*, [this issue], xxx–yyy.

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

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

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

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

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

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, [this issue], xxx–yyy.

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

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

Graefe, A., Küchenhoff, H., Stierle, V., & Riedl, B. (2014). Limitations of ensemble bayesian model averaging for forecasting social science problems. *International Journal of Forecasting* (Forthcoming), Available at http://ssrn.com/abstract=2266307.

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

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

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. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, [this issue], xxx–yyy.

Green, K. C., & Armstrong, J. S. (2014). Forecasting global climate change. In A. Moran (Ed.), *Climate change: The facts* (pp. 170–186), Melbourne: IPA.

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

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

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

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

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

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

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

Keilman, N. (2008). European demographic forecasts have not become more accurate over the past 25 years. *Population and Development Review, 34*(1), 137–153.

Keogh, E. & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*. *7*(4), 349–371.

Kim, M. & Hunter, J. E. (1993). Relationships among attitudes, behavioral intentions, and behavior: A meta-analysis of past research, Part 2. *Communication Research*, *20*(3), 331–364.

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

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

Kort, F. (1957), Predicting Supreme Court decisions mathematically: A quantitative analysis of "right to counsel" cases. *The Americam Political Science Review*, *51*, 1–12.

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

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

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

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. L. (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition. *Journal of Forecasting, 1*(2), 111–153.

Makridakis, S., & Hibon, M. (2000). The M-3 competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

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

McNown, R., Rogers, A., & Little, J. (1995). Simplicity and complexity in extrapolation population forecasting models. *Mathematical Population Studies, 5*(3), 235–257.

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).

Miller, T., & Liberatore, M. (1993). Seasonal exponential smoothing with damped trends: An application for production planning. *International Journal of Forecasting*, *9*, 509–515.

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

Morwitz, V. G. (2001). Methods for forecasting from intentions data. In J. S. Armstrong (Ed.), *Principles of Forecasting*, Boston: Kluwer Academic Publishers.

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

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

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukosa, V., & Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of Business Research*, [this issue], xxx–yyy.

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

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

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

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

Ryan, P., & Sessions, J. (2013). Sessions, Ryan Call For Halt On Taxpayer Funding For Risky High-Speed Rail Project. *U.S. Senate Budget Committee.* Available at 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. (2014). *Big data driven supply chain management: A framework for implementing analystics and turning analystics into intelligence*. New Jersey: Pearson Education.

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

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

Schnaars, S. P. (1986). A comparison of extrapolation models on yearly sales forecasts. *International Journal of Forecasting*, *2*, 71–85.

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

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

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

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

Tessier, T. H., & Armstrong, J. S. (2015). Decomposition of time-series by level and change. *Journal of Business Research*, [this issue], xxx–yyy.

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.

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

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

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

Wright, M., & MacRae, M. (2007). Bias and variability in purchase intention scales. *Journal of the Academy of Marketing Science, 35*(4), 617–624.

Wright, M., & Stern, P. (2015). Forecasting new product trial with analogous series. *Journal of Business Research*, [this issue], xxx–yyy.

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

Zellner, A. (2001). Keep it sophisticatedly simple. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.) *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple* (pp. 242–262). Cambridge: Cambridge University Press.

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