Forecasting methods and principles: Evidence-based checklists

J. Scott Armstrong & Kesten C. Green

To cite this article: J. Scott Armstrong & Kesten C. Green (2018) Forecasting methods and principles: Evidence-based checklists, Journal of Global Scholars of Marketing Science, 28:2, 103-159, DOI: 10.1080/21639159.2018.1441735

To link to this article: https://doi.org/10.1080/21639159.2018.1441735

Published online: 14 Mar 2018.

Submit your article to this journal

View related articles

View Crossmark data
Forecasting methods and principles: Evidence-based checklists

J. Scott Armstrong⁹ᵇ and Kesten C. Greenᶜ

⁹The Wharton School, University of Pennsylvania, Philadelphia, PA, USA; bEhrenberg-Bass Institute, University of South Australia Business School, University of South Australia, Adelaide, Australia; cSchool of Commerce and Ehrenberg-Bass Institute, University of South Australia Business School, University of South Australia, Adelaide, Australia

ABSTRACT
Problem: How to help practitioners, academics, and decision makers use experimental research findings to substantially reduce forecast errors for all types of forecasting problems.

Methods: Findings from our review of forecasting experiments were used to identify methods and principles that lead to accurate forecasts. Cited authors were contacted to verify that summaries of their research were correct. Checklists to help forecasters and their clients undertake and commission studies that adhere to principles and use valid methods were developed. Leading researchers were asked to identify errors of omission or commission in the analyses and summaries of research findings.

Findings: Forecast accuracy can be improved by using one of 15 relatively simple evidence-based forecasting methods. One of those methods, knowledge models, provides substantial improvements in accuracy when causal knowledge is good. On the other hand, data models – developed using multiple regression, data mining, neural nets, and “big data analytics” – are unsuited for forecasting.

Originality: Three new checklists for choosing validated methods, developing knowledge models, and assessing uncertainty are presented. A fourth checklist, based on the Golden Rule of Forecasting, was improved.

Usefulness: Combining forecasts within individual methods and across different methods can reduce forecast errors by as much as 50%. Forecasts errors from currently used methods can be reduced by increasing their compliance with the principles of conservatism (Golden Rule of Forecasting) and simplicity (Occam’s Razor). Clients and other interested parties can use the checklists to determine whether forecasts were derived using evidence-based procedures and can, therefore, be trusted for making decisions. Scientists can use the checklists to devise tests of the predictive validity of their findings.

KEYWORDS
combining forecasts; data models; decomposition; equalizing; expectations; extrapolation; knowledge models; intentions; Occam’s razor; prediction intervals; predictive validity; regression analysis; uncertainty

预测方法和原则: 循证清单

问题: 如何帮助从业人员、学者和决策者使用实验研究成果，从而大幅降低各类预测问题的预测误差。
方法：我们利用预测实验回顾的成果，确定准确预测的方法和原则。联系被引作者，核实他们的研究综述是否正确。我们制定了清单，以帮助预测者和他们的客户实践、委托研究、这些研究遵循原则并使用有效方法。我们要求主要研究人员，通过分析和总结研究成果，找出遗漏误差或委托误差。

发现：这些15个循证预测方法相对简单，使用其中之一，就可提高预测准确性。其中一种方法，即知识模型，在基本知识良好时，可大大提高预测准确性。另一方面，使用多元回归、数据挖掘、神经网络和“大数据分析”开发的数据模型不适合进行预测。

独创性：我们介绍了三种新清单，用于选择验证方法、开发知识模型和评估不确定性。基于预测黄金法则的第四个清单得到了改进。

实用性：组合预测法，使用个别方法以及交叉不同方法做出的预测，可减少高达50%的预测误差。若我们目前所使用的方法遵循稳健性原则（预测的黄金法则）和简单性原则（奥卡姆剃刀定律），我们可以降低它们的预测误差。客户和其他相关方可使用清单来确定预测是否是利用循证程序推导出的，若是，这些预测值得信赖。因此，我们可以做出决定。科学家可使用清单设计实验，测试他们成果的预测效果。

Introduction

Forecasts are important for decision-making in businesses and other organizations, and for governments. A survey of practitioners, educators, and decision-makers found that they rated “accuracy” as the most important of 13 criteria for judging forecasts (Yokum & Armstrong, 1995). Researchers were especially concerned with accuracy.

Consistent with those findings, improving forecast accuracy is the primary concern of this paper. Since the 1930s, researchers have responded to the need for accurate forecasts by conducting experiments testing multiple reasonable methods. The findings from those groundbreaking experiments have greatly improved forecasting knowledge. In the late 1990s, 39 forecasting researchers from a variety of disciplines summarized scientific knowledge on forecasting. They were assisted by 123 expert reviewers (Armstrong, 2001a). The findings were used to develop 139 principles (condition–action statements) for forecasting in various situations. In 2015, two papers further condensed forecasting knowledge as two overarching principles: conservatism and simplicity (Armstrong, Green, & Graefe, 2015; and Green & Armstrong, 2015, respectively).

While the advances in forecasting knowledge allow for substantial improvements in forecast accuracy, that knowledge is largely ignored in academic journal articles and, we expect, also by practitioners. At the time that the original 139 forecasting principles were published in 2001, a review of 17 forecasting textbooks found that the typical book mentioned only 19% of the principles (Cox & Loomis, 2001). Moreover, forecasting software packages, which could help to ensure that the principles are used, were found to ignore about half of the forecasting principles (Tashman & Hoover, 2001).
Checklists to improve forecasting

The use of evidence-based checklists avoids the need for memorizing and simplifies complex tasks. In fields such as medicine, aeronautics, and engineering, a failure to follow an appropriate checklist can be grounds for a lawsuit.

The use of checklists is supported by much research (e.g. Hales & Pronovost, 2006). One experiment assessed the effects of using a 19-item checklist for a hospital procedure. The study compared thousands of patient outcomes in hospitals in eight cities around the world before and after the checklist was used. Use of the checklist reduced deaths from 1.5 to .8% in the month after the medical procedures (Haynes et al., 2009). Importantly, checklists improve decision-making even when the knowledge incorporated in them is well known to practitioners, and is known to be important (Hales & Pronovost, 2006). To ensure that they include the latest evidence, checklists should be revised routinely.

Convincing people to use checklists is easy. When engineers and medical doctors are told they must use the checklist as a condition of their employment, and when use of the checklist is monitored, they use the checklists. When we have paid people modest sums to complete tasks using checklists, almost all of those who accepted the task did so effectively. For example, to assess the persuasiveness of print advertisements, raters hired through Amazon’s Mechanical Turk used a 195-item checklist to evaluate advertisements’ conformance to persuasion principles. The inter-rater reliability was high (Armstrong, Du, Green, & Graefe, 2016).

Research methods

We reviewed prior experimental research on which forecasting methods and principles lead to improved forecast accuracy. To do so, we first identified relevant research by:

1. searching the Internet, mostly using Google Scholar;
2. contacting leading researchers for suggestions of important experimental findings;
3. checking key papers referred to in experimental studies and meta-analyses;
4. putting our working paper online with requests for evidence that we might have overlooked;
5. providing links to all papers in an OpenAccess version of this paper in order to allow readers to check our interpretations of the original findings.

Given the enormous number of papers with promising titles, we screened papers by assessing whether the “Abstract” or “Conclusions” sections provided evidence on the comparative value of alternative methods, and full disclosure. Only a small percentage of the papers with promising titles met those criteria.

Only studies that examine many out-of-sample (ex ante) forecasts are considered as evidence in this paper. For cross-sectional data, the “jack-knife” procedure allows for many forecasts by using all but one data point to estimate the model, making a forecast for the excluded observation, then replacing that observation and excluding another, and so on until forecasts have been made for all data points. Successive updating can be used to increase the number of out-of-sample forecasts for time-series data. For example, to test the predictive validity of alternative models for forecasting the next 100 years of global mean temperatures, annual forecasts were made for horizons from one to 100 years-ahead starting in 1851. The
forecasts were updated as if in 1852, then 1853, and so on, thus providing errors for 157 one-year-ahead forecasts … and 58 one-hundred-year-ahead forecasts (Green, Armstrong, & Soon, 2009).

We attempted to contact the authors of all papers that we cited regarding substantive findings. We did so on the basis of evidence that findings cited in papers in leading scientific journals are often described incorrectly (Wright & Armstrong, 2008). We asked the authors if our summary of their findings was correct and whether our description could be improved. We also asked them to suggest relevant papers that we had overlooked – especially papers describing experiments with findings that conflicted with our conclusions. That practice was shown to contribute to a substantially more comprehensive search for evidence than was achieved by computer searches (Armstrong & Pagell, 2003). In the case of six papers, we could not agree with the authors on the interpretation of findings.

We discarded our citations of those papers, as they were not essential to this paper. Of the 90 papers with substantive findings that were not our own, we were able to contact the authors of 73 and received substantive, and often helpful, replies from 69.

Our review led to the development of five checklists. They provide evidence-based guidance on forecasting methods, knowledge models, the Golden Rule of Forecasting, simplicity, and uncertainty.

**Valid forecasting methods: Checklist and evidence**

The predictive validity of a forecasting method is assessed by comparing the accuracy of forecasts from the method with the accuracy of forecasts from currently used methods, or from simple benchmark methods such as the naïve no-trend model, or from other evidence-based methods. Such testing of multiple reasonable hypotheses is a requirement of the scientific method as described by Chamberlin (1890/1965).

For categorical forecasts – such as whether $a$, $b$, or $c$ will happen, or which of them would be better – accuracy is typically measured as a variation of percent correct. For quantitative forecasts, accuracy is assessed by differences between *ex ante* forecasts and data on what actually transpired. The benchmark error measure for evaluating forecasting methods is the relative absolute error, or “RAE.” It has been shown to be more reliable than the root mean square error (Armstrong & Collopy, 1992). Tests of a new method – a development of the RAE – called the unscaled mean bounded relative absolute error (UMBRAE) – suggest that it is superior to the RAE and other proposed alternatives (Chen, Twycross, & Garibaldi, 2017). We suggest using both the RAE and UMBRAE until additional testing has been done to provide a definitive conclusion on which is the better measure.

Figure 1 lists 15 individual evidence-based forecasting methods. They are consistent with forecasting principles and have been shown to provide out-of-sample forecasts with superior accuracy. The Figure also identifies the knowledge needed to use each method. Combining within and across methods is recommended (Checklist items 16 and 17.)

For most forecasting problems, several of the methods will be usable, and should be used, as we describe below. An electronic version of the Figure 1 checklist is provided at ForecastingPrinciples.com in the top menu bar under “Methods Checklist.”
Because we are concerned with methods that have been shown to improve forecast accuracy relative to methods that are commonly used in practice, not all methods that have been used for forecasting are included in the Figure 1 checklist. For example, multiple regression analysis is apparently one of the most widely used methods for developing forecasting models. Given the evidence summarized in this paper, however, we recommend against the use of multiple regression analysis and other data modeling approaches.

Clients should ask forecasters what methods they propose using and why. If they mention a method that is not listed in Figure 1, they should be asked to produce evidence that their method will provide forecasts with smaller errors.

**Figure 1.** Forecasting methods application checklist.

<table>
<thead>
<tr>
<th>Method</th>
<th>Knowledge needed</th>
<th>Usable method</th>
<th>Variations within components</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Prediction markets</td>
<td>Survey/ market design</td>
<td>Domain, Problem</td>
<td>□</td>
</tr>
<tr>
<td>2. Multiplicative decomposition</td>
<td>Domain, Structural relationships</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>3. Intentsional surveys</td>
<td>Survey design</td>
<td>Own plans/behavior</td>
<td>□</td>
</tr>
<tr>
<td>4. Expectations surveys</td>
<td>Survey design</td>
<td>Others’ behavior</td>
<td>□</td>
</tr>
<tr>
<td>5. Expert surveys (Delphi, etc.)</td>
<td>Survey design</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>6. Simulated interaction</td>
<td>Survey/experimental design</td>
<td>Normal human responses</td>
<td>□</td>
</tr>
<tr>
<td>7. Structured analogies</td>
<td>Survey design</td>
<td>Analogous events</td>
<td>□</td>
</tr>
<tr>
<td>8. Experimentation</td>
<td>Experimental design</td>
<td>Normal human responses</td>
<td>□</td>
</tr>
<tr>
<td>9. Expert systems</td>
<td>Survey design</td>
<td>Domain</td>
<td>□</td>
</tr>
</tbody>
</table>

**Judgmental methods (Judgmental inputs sometimes required)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Knowledge needed</th>
<th>Usable method</th>
<th>Variations within components</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Extrapolation</td>
<td>Time-series methods, Data</td>
<td>a/o</td>
<td>□</td>
</tr>
<tr>
<td>11. Rule-based forecasting</td>
<td>Causality, Time-series methods</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>12. Judgemental bootstrapping</td>
<td>Survey/Experimental design</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>13. Segmentation</td>
<td>Causality, Data</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>14. Simple regression</td>
<td>Causality, Data</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>15. Knowledge models</td>
<td>Cumulative causal knowledge</td>
<td>Domain</td>
<td>□</td>
</tr>
<tr>
<td>16. Combining forecasts from a single method</td>
<td>SUM of VARIATIONS</td>
<td></td>
<td>□</td>
</tr>
<tr>
<td>17. Combining forecasts from several methods</td>
<td>COUNT of METHODS</td>
<td></td>
<td>□</td>
</tr>
</tbody>
</table>

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

*Experts who are consulted by the forecaster about their domain knowledge should be aware of relevant findings from experiments. Failing that, the forecaster is responsible for obtaining that knowledge.

**Judgmental methods**

Expertise based on experience in similar situations can be useful for forecasting. Experience can lead to simple “rules of thumb,” or heuristics, that provide quick forecasts for rapid decision-making. For example, the emergency landing of US Airways Flight 1549 – the “Miracle on the Hudson” – was a success because the pilot used the gaze heuristic to forecast that landing on the Hudson River was a viable option, whereas returning to La Guardia Airport was not (Hafenbrädl, Waeger, Marewski, & Gigerenzer, 2016). Extensive research conducted by Gerd Gigerenzer and the ABC group of the Max Planck Institute for Human Development in Berlin has found that simple heuristics are superior to more complex and information intensive methods for many practical problems.
For situations in which there are two or more important causal factors and where experts do not receive frequent well-summarized feedback on the accuracy of their predictions, however, expertise and experience are, on their own, of no apparent value. Such situations are common in business and government decision-making. Even leading experts’ unaided judgmental forecasts often turn out to be disastrously wrong, sometimes to the delight of the media (e.g. see Cerf & Navasky, 1998; Perry, 2017).

Research on the accuracy of experts’ unaided judgmental forecasts about complex situations dates from the early 1900s. An early review of the research led to the Seer-Sucker Theory: “No matter how much evidence exists that seers do not exist, suckers will pay for the existence of seers” (Armstrong, 1980). The Seer-Sucker Theory has held up well over the years; in particular, a 20-year study comparing the accuracy of many forecasts from experts with those of forecasts from novices and from naïve rules provided support (Tetlock, 2005).

While unaided expert judgments should be avoided, domain experts can play a vital role in forecasting when their judgments are incorporated using evidence-based methods. The next section describes nine structured methods for forecasting using expert judgment.

1. Prediction markets

Prediction markets – also known as betting markets, information markets, and futures markets – have been used for forecasting since the 16th century (Rhode & Strumpf, 2004). Monetary rewards attract people who believe they have knowledge or information that enables them to make accurate predictions about the situation they are betting on.

Prediction markets are especially useful when knowledge is dispersed and many participants are motivated to trade repeatedly. Markets rapidly revise forecasts when new information becomes available. Forecasters using prediction markets need to be familiar with designing prediction markets and surveys.

The accuracy of forecasts from prediction markets was tested in eight published comparisons in the field of business forecasting (Graefe, 2011). The results were mixed. For example, prediction markets’ out-of-sample forecast errors were 28% smaller than those from no-change models in one comparison. On the other hand, averaging people’s judgments outperformed market forecasts in two of three comparisons. In another comparison, forecasts from the Iowa Electronic Market (IEM) prediction market across the 100 days before each U.S. presidential election from 2004 to 2016 were, on average, less accurate than forecasts from the RealClearPolitics poll average, a survey of experts, and citizen forecasts (Graefe, 2017a). The IEM prediction market limits the bets to no more than $500, which likely reduces the number and motivation of participants. Comparative accuracy tests based on 44 elections in eight countries other than the U.S., however, found that forecasts from betting markets were more accurate than forecasts by experts, econometric models, and polls (Graefe, 2017b).

2. Multiplicative decomposition

Decomposition has long been a key element of forecasting. A Google search for “decomposition” and either “forecast” or “predict” found over 45 million results in December 2017.

Multiplicative decomposition involves dividing a forecasting problem into parts, forecasting each part separately, and multiplying the forecasts of the parts to forecast the whole. For example, to forecast sales for a brand, a firm might separately forecast total market sales and market share, and then multiply those components. Decomposition is expected to be most
effective at reducing forecast errors when suitable forecasting methods, data availability, and directional effects of causal factors vary among the parts.

To assess the effect of decomposition on forecast accuracy, subjects in an experiment were presented with five problems from an almanac, such as “How many packs (rolls) of Polaroid color films do you think were used in the United States in 1970?” Some subjects were asked to make estimates of the total figure, while others were asked to estimate each of the decomposed elements (Armstrong, Denniston, & Gordon, 1975). Across that study and two similar studies, forecast error was reduced by an average of 42% (MacGregor, 2001).

Another study used graphical software to display the 68 monthly series from the M-Competition (Makridakis et al., 1982) in ways that were designed to help users identify and forecast seasonality and trend independently using their judgment. The study found that three postgraduate students with knowledge of time series analysis and the software produced forecasts for 1 to 12 months into the future that had errors that were 7% less than those from the leading M-Competition method of deseasonalized single exponential smoothing. The error reduction from software-assisted judgmental decomposition by 35 novices forecasting five time-series each was 5% (Edmundson, 1990, Table 2).

3. Intentions surveys

Intentions surveys ask people how they plan to behave in specified situations. They can be used, for example, to predict how people would respond to major changes in the design of a product. One meta-analysis included 47 comparisons with over 10,000 subjects, and another provided a meta-analysis of 10 meta-analyses involving over 83,000 subjects. Both found a strong relationship between people’s intentions and their future behavior (Kim & Hunter, 1993; Sheeran, 2002).

Intentions surveys are especially useful when historical data are not available. They are most likely to provide useful forecasts for short forecast time-horizons, and for important decisions (Morwitz, 2001; Morwitz, Steckel, & Gupta, 2007).

To assess people’s intentions, the forecaster should prepare brief unbiased descriptions of the situation (Armstrong & Overton, 1971). Intentions should be expressed as probabilities such as 0 = “No chance, or almost no chance (1 in 100)”, to 10 = “Certain, or practically certain (99 in 100).” Responses can be used to calculate a forecast of how people will behave, such as “3.2% of the population will buy the product in the next three months” (Morwitz, 2001).

The way a question is asked can have a large effect on responses. Two ways to reduce response error are to: (1) pretest the questions to ensure that the respondents understand them in the way the forecaster intends and (2) use alternative ways to state a question, then average responses across questions. For more advice, see Bradburn, Sudman, and Wansink (2004).

4. Expectations surveys
Expectations surveys ask people how they expect they or others will behave. Expectations differ from intentions because people realize that the situation can change. For example, if you were asked whether you intend to purchase a vehicle over the next year, you might say that you have no intention of doing so. However, you realize that it is possible that your vehicle will develop a major problem. As a consequence, you might expect that there is a chance that you will purchase a new car. As with intentions surveys, expectations surveys should use probability scales, follow evidence-based procedures for survey design, use representative samples, obtain high response rates, and correct for non-response bias by extrapolating across waves.

Following the U.S. government’s 1932 prohibition of prediction markets for political elections, expectations surveys – which poll a representative sample of potential voters on how others would vote – were introduced (Hayes, 1936). Those “citizen expectations” surveys correctly predicted the popular vote winners of the U.S. Presidential elections in 89% of the 217 surveys from 1932 to 2012. Furthermore, citizens’ expectations provided more accurate out-of-sample forecasts of the national vote share than polls, prediction markets, models, and experts across the seven U.S. Presidential elections from 1988 to 2012 (Graefe, 2014), and again in 2016. Over the 100 days before the 2016 election, the error of citizens’ expectations forecasts of the popular vote in seven U.S. Presidential elections from 1992 through 2016 averaged 1.2 percentage points. In comparison, the error of a typical poll aggregator was, at 2.6 percentage points, more than twice as high. (Graefe, Armstrong, Jones, & Cuzán, 2017).

5. Expert surveys
Use written questions and instructions for self-completion surveys to ensure that each expert is questioned in the same way. Apply the same procedures for developing questions as those described for expectations surveys above.

Forecasters should obtain forecasts from at least 5 experts, and up to 20 for important forecasts (Hogarth, 1978). That advice was followed in forecasting the popular vote for U.S. Presidential elections from 2004 to 2016, when surveys of about 15 experts led to an average error of 1.6 percentage points, compared to 1.7 percentage points for combined polls (Graefe et al., 2017; and personal correspondence with Graefe). Additional advice on the design of expert surveys is provided in Armstrong (1985, pp.108–116).

Delphi is an extension of the expert survey approach whereby the survey is conducted over two or more rounds. After each round, anonymous summaries of the experts’ forecasts and reasons are provided to the experts. The process is repeated until forecasts change little between rounds – usually two or three rounds are sufficient. Use the median or mode of the experts’ final-round forecasts as the Delphi forecast. Delphi is expected to be most useful when the different experts each have different information relevant to the problem (Jones, Armstrong, & Cuzán, 2007).

Forecasts from Delphi were more accurate than forecasts from traditional meetings in five studies, similarly accurate in two, and less accurate in one. Delphi forecasts were more accurate than forecasts from traditional surveys of expert opinion for 12 of 16 studies, with two ties and two cases in which Delphi was less accurate. Among those 24 comparisons, Delphi improved accuracy in 71% and harmed it in 12% (Rowe & Wright, 2001).

Delphi is attractive to managers because judgments from dispersed experts can be obtained without the expense of arranging meetings. It has an advantage over prediction
markets in that the participants provide reasons for their forecasts (Green, Armstrong, & Graefe, 2007). Software for the procedure is freely available at ForecastingPrinciples.com.

6. Simulated interaction

Simulated interaction uses role-playing to forecast decisions by two or more parties with conflicting interests. Situations that have been used for testing the method include an attempt to secure an exclusive distribution arrangement with a major supplier, a union–management dispute over pay and conditions, and artists demanding that the government provide them with financial support.

The forecaster provides each role-player with a description of one the main protagonists' roles, and a brief description of the situation including a list of possible decisions. The role-players are asked to engage in realistic interactions with one another, staying in their roles until a decision is reached. The simulations typically last less than an hour.

Relative to unaided expert judgment – the most common method – simulated interaction reduced forecast errors by 57% on average for eight conflict situations, including those described above and an attempted hostile takeover of a corporation, and a military standoff between two countries over access to water (Green, 2005). The method seems to work best when naïve role players do not know each other, have no prior opinions about the situation, and no agenda beyond that indicated by their role.

The alternative approach of “putting oneself in the other person's shoes” has been proposed. U.S. Secretary of Defense Robert McNamara suggested that if he had done this during the Vietnam War, he would have made better decisions. A test of the “role-thinking” approach, however, found no improvement in forecast accuracy relative to that of unaided judgment. It is too difficult to think through the interactions in a complex situation – active role-playing between parties is necessary to provide sufficient realism (Green & Armstrong, 2011).

7. Structured analogies

The structured analogies method involves asking 10 or so experts to suggest situations that were similar to the one for which a forecast is required, the target situation. The experts are given a description of the target situation and are asked to identify analogous situations, rate their similarity to the target, and match the outcomes of their analogies with possible outcomes of the target situation. An administrator takes the target situation outcome implied by each expert's top-rated analogy and calculates the modal outcome as the forecast. The method should not be confused with the common use of analogies to justify a decision that is preferred by the forecaster or client.

Structured analogies forecasts were 41% more accurate than unaided judgment forecasts in forecasting decisions in the eight real conflicts used in research on the simulated interaction method described above (Green & Armstrong, 2007a). Structured analogies were also used to forecast the effects of incentives to promote laptop purchases by university students, and a program offering certification on Internet safety to parents of high school students. The error of those structured analogies forecasts was 8% lower than the error of forecasts from unaided judgment (Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015). A procedure akin to structured analogies was used to forecast box office revenue for 19 unreleased movies, in which raters identified analogous movies from a database and rated them for similarity. The revenue forecasts from the analogies were adjusted for advertising expenditure and whether the movie was a sequel. Errors from the structured analogies
forecasts were less than half those of forecasts from simple and complex regression models (Lovallo, Clarke, & Camerer, 2012). Across the 10 comparative tests from the 3 studies described above, the error reductions from using structured analogies averaged about 40%.

8. Experimentation

Experimentation is widely used and is the most valid and reliable method for determining cause-and-effect relationships. Knowledge of the direction of effects and estimates of the strength of effects can then be used to make forecasts. Experiments can be conducted in laboratories. An analysis of organizational behavior experiments found that laboratory experiments yielded similar findings to field experiments (Locke, 1986).

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

9. Expert systems

Expert systems are developed by asking experts to describe the steps they take while they make forecasts, then describing that process using software. The resulting expert system should be complete, simple, and clearly described.

A review of 15 comparisons found that expert system forecasts were more accurate than forecasts from unaided judgments (Collopy, Adya, & Armstrong, 2001). Two of the studies – on gas, and on mail order catalog sales – found that the expert systems’ forecast errors were 10 and 5% smaller, respectively, than those of unaided judgment. While the evidence available on predictive validity is scant, the method appears promising.

Quantitative methods

Quantitative methods require numerical data on or related to the forecasting problem. Quantitative methods can also draw upon judgmental methods, such as decomposition, in order to make the best use of knowledge and data.

This section describes six evidence-based quantitative forecasting methods. Other than the first of the methods (extrapolation), the methods rely heavily on causal knowledge to forecast the effects of changes in causal variables. Such forecasts can be used for policy-making, and for developing contingency plans. Forecasting what will happen when the causal variables are out of the decision-makers’ control, however, requires that the causal variables are accurately forecast.

10. Extrapolation

While extrapolation methods can be used for any problem requiring forecasts of a time series, they are especially useful when little is known about the factors affecting the forecast variable, causal variables are not expected to change much, or causal variables cannot be forecast with much accuracy.

Exponential smoothing, which dates back to Brown (1959, 1962), is easy to understand. It is a sensible approach because it uses all historical data in a moving average that puts more weight on the most recent data. For a review of exponential smoothing, see Gardner (2006).
One should not assume that a trend will continue at the same rate, even in the short term. It could increase or decrease in response to changes in the causal forces that drive the trend. The greater the uncertainty about the situation, the greater is the need to damp the trend toward zero – the no change forecast. A review of 10 experimental comparisons found that, on average, damping the trend toward zero reduced forecast errors by almost 5% and reduced the risk of large errors compared to forecasts that assumed a constant trend (Armstrong, 2006). Gardner's software for damped-trend extrapolation can be found at ForecastingPrinciples.com. When there is a long-term trend and the causal factors are expected to continue – such as with the real prices of resources (Simon, 1996) – damping toward the long-term trend is appropriate.

When extrapolating for time periods less than a year, estimate the effects of seasonal influences and remove them from the data. Forecast the seasonally adjusted series, then “seasonalize” the forecasts. In forecasts for 68 monthly economic series over 18-month horizons from the M-Competition, seasonal adjustment reduced forecast errors by 23% (Makridakis et al. 1984, Table 14).

Forecasters should damp statistical estimates of seasonal influences. Such estimates are uncertain and standard seasonal adjustment procedures tend to “overfit” the data. Miller and Williams (2003, 2004) provide procedures for damping seasonal factors. When they damped the seasonal adjustments for the 1,428 monthly time-series from the M3-Competition, the accuracy of the forecasts improved for 59–65% of the time series, depending on the horizon. The broad findings were replicated by Boylan, Goodwin, Mohammadipour, and Syntetos (2015). Software for the Miller–Williams procedures and the M3-Competition data are freely available at ForecastingPrinciples.com.

Damping by averaging seasonal factors across analogous series also improves forecast accuracy. In one study, combining seasonal factors from related products, such as snow blowers and snow shovels, reduced the average forecast error by about 20% (Bunn & Vassilopoulos, 1999). In another study, pooling monthly seasonal factors for crime rates for six city precincts reduced the error of exponential smoothing forecasts by about 7% compared to using seasonal factors that were estimated individually for each precinct (Gorr, Oligschlager, & Thompson, 2003, Figure 4).

Multiplicative decomposition can be used to incorporate causal knowledge into extrapolation forecasts. For example, when forecasting time-series data, it often happens that the series is affected by causal forces – characterized as growth, decay, opposing, regressing, supporting, or unknown. In such a case, one can decompose the time series by causal forces that have different directional effects, extrapolate each component, and then recombine. Doing so is likely to improve accuracy under two conditions: (1) domain knowledge can be used to structure the problem so that causal forces differ for two or more of the component series and (2) it is possible to obtain relatively accurate forecasts for each component. For example, to forecast motor vehicle deaths, one study forecast the number of miles driven, a series that would be expected to grow, and the death rate per million passenger miles, a series that would be expected to decrease due to better roads and safer cars. The two extrapolation forecasts were then multiplied to get total deaths. When tested on five time series that clearly met the two conditions, decomposition by causal forces reduced out-of-sample forecast errors by two-thirds. For the four series that partially met the conditions, decomposition by causal forces reduced error by one-half. There was no gain or loss in forecast accuracy when the conditions did not apply (Armstrong, Collopy, & Yokum, 2005).
Additive decomposition can also be considered for extrapolation problems. One approach that is useful when the most recent data are uncertain or liable to subsequent revision is to forecast the starting level and trend separately, and then add them—a procedure called “nowcasting.” Three comparative studies found that, on average, nowcasting reduced errors for short-range forecasts by 37% (Tessier & Armstrong, 2015).

11. Rule-based forecasting

Rule-based forecasting (RBF) uses knowledge about evidence-based extrapolation along with causal knowledge to forecast time-series data. To use RBF, first identify which of 28 “features” best characterize the series to be forecast. Features include forecast horizons, the amount of data available, and the existence of outliers. Then use the 99 RBF rules to weight the alternative extrapolation models and combine the models’ forecasts (Armstrong, Adya, & Collopy, 2001).

For one-year-ahead ex ante forecasts of 90 annual series from the M-Competition (available on ForecastingPrinciples.com), the median absolute percentage error of RBF forecasts was 13% smaller than that of equally weighted combined forecasts. For six-year-ahead ex ante forecasts, the RBF forecast errors were 42% smaller, likely due to the increasing importance of causal effects over longer horizons. RBF forecasts were also more accurate than equally weighted combinations of forecasts in situations involving strong trends, low uncertainty, stability, and good domain expertise. RBF forecasts had little or no accuracy advantage over unweighted combinations of forecasts for other situations (Collopy & Armstrong, 1992). Testing by Vokurka, Flores, & Pearce (1996) provided supporting evidence for the relative accuracy of RBF forecasts.

One of the 99 RBF rules, the “contrary series rule” is especially important, as well as simple and inexpensive to apply. It states that one should not extrapolate a trend if the direction of a time series expected by domain experts is contrary to the recent trend of the time series. The use of that rule alone yielded improvements in extrapolating time-series data from five data sets. In particular, for longer term (six-years ahead) forecasts, the error reductions exceeded 40% (Armstrong & Collopy, 1993).

12. Judgmental bootstrapping

This method was developed in the early 1900s to provide forecasts of the size of the upcoming corn harvest in the U.S. In the 1940s, the method was used successfully for personnel selection (Meehl, 1954) and has been supported by subsequent research (e.g. Dawes & Corrigan, 1974; Grove, Zaid, Lebow, Snitz, & Nelson, 2000). The method uses regression analysis to estimate coefficients for the variables that experts use to make judgmental forecasts. The dependent variable is not the outcome, but rather the experts’ predictions of the outcome given the values of the causal variables. Among researchers in forecasting, the method has, in recent decades, been called “judgmental bootstrapping.” In effect, it uses a quantitative model of the experts’ use of causal information for forecasting to improve upon the experts’ forecast accuracy.

In comparative studies to date, the bootstrap model’s forecasts were more accurate than those of the experts whose judgments they were based on. The gain in accuracy arises from the quantitative model’s more consistent application of the expert’s mental model. In addition, the model does not become distracted by irrelevant information and variables, nor does it become tired or irritable.
The first step for developing a judgmental bootstrap model is to ask experts to identify causal variables based on their domain knowledge. Then ask them to make predictions using data on the variables. For example, they could be asked to forecast the likelihood of success of doctoral candidates (Dawes, 1971).

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

Despite the discovery of the method and evidence on its usefulness, its early use was confined to agricultural predictions. Social scientists rediscovered the method in the 1960s, and tested its predictive validity. A review of those studies found that judgmental bootstrapping forecasts were more accurate than those from unaided judgments in eight of 11 comparisons, with two tests finding no difference and one finding a small loss in accuracy (Armstrong, 2001b). The one failure occurred when the experts relied on an irrelevant variable that was not excluded from the bootstrap model. The typical error reduction was about 6% relative to unaided judgment. Many universities taught the methods to their students, but we are aware of only one that adopted the method.

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

13. Segmentation

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

The Port of New York Authority used the method in 1955 to forecast air travel demand ten years hence. Their analysts divided airline travelers into segments of 130 business traveler types and 160 personal traveler types. The personal travelers were segmented by age, then by occupation, income, and education; and the business travelers were segmented by occupation, then industry, and finally income. Data on each segment were obtained from the census and from a survey on travel behavior. To derive the forecast, the official projected air travel population for 1965 was allocated among the segments, and the number of travelers and trip frequency were extrapolated using 1935 as the starting year with zero travelers. The resulting forecast of 90 million trips was only 3% different from the actual 1965 figure (Armstrong, 1985).
To use segmentation, identify important causal variables that can be used to define the segments, and their priorities. Then determine cut-points – e.g. different age categories of people – for each variable. Use more cut-points when there are nonlinearities in the relationships and fewer cut points when the samples of data are smaller. Next, forecast the population of each segment and the behavior of the population within each segment using the typical behavior. Finally, combine the population and behavior forecasts for each segment and sum across segments. The method is most likely to be useful when much data are available.

Segmentation is suitable for situations in which variables are interrelated, the effects of variables are nonlinear, and prior causal knowledge is good. These conditions occurred, to a moderate extent, in a study where data from 2,717 gas stations were used to estimate a segmentation model for forecasting weekly gasoline sales volumes. Data were available on 9 binary variables and 10 other variables including type of area, traffic volumes, length of street frontage, presence of a canopy, and whether the station was open 24 hours a day. The method was tested using a holdout sample of 3,000 stations. The segmentation model forecast errors (mean absolute percentage errors) were 29% smaller than the errors of a multiple regression model estimated using the same variables and data (Armstrong & Andress, 1970).

A review of the literature on segmentation is provided by Armstrong (1985, Chapter 9). While the evidence on predictive validity is not substantial, the method is sensible, as it is based on decomposition. Interest in segmentation fell away after the 1970s, but we expect that it would be more useful now than ever before, given the availability of large databases.

14. Simple regression

Simple regression analysis can be used to forecast the effect of changes in a single causal variable. The method is conservative in that it reduces the effect size estimate toward the mean – via the calculation of a constant term – in response to variations in the relationship found in the estimation data. For a forecasting model estimated using simple regression to be useful, one must be able to control or accurately forecast the causal variable.

The traditional form of a simple regression model is $y = a + bx$, where “$y$” is the variable to be forecast (dependent variable), “$a$” is the constant, “$b$” is the effect size, and “$x$” is the causal variable. The method is appropriate for forecasting problems that involve good prior knowledge about a strong causal relationship, along with valid and reliable data on the dependent and causal variables.

Transform the data so that the simple regression model provides a realistic representation of the causal relationship. For example, calculating logarithms of the causal and dependent variables before estimating the model will result in an effect size estimate in the form of an elasticity. Elasticities are the percentage change in the variable to be forecast that would result from a 1% change in the causal variable. A price elasticity of demand of -1.2 for beef, for example, means that one would expect a price increase of 10% to result in a 12% decrease in the quantity demanded, all else being equal. Other transformations to consider include expressing the variables in per capita terms, and adjusting the data for the effect of currency inflation and seasonality.
The least squares method of estimating regression model coefficients has the effect of giving extreme data values an excessive influence on the estimate of the effect size. To avoid that, adjust or remove outliers from the estimation data. One way to do so – known as “winsorizing” – is to set the outlier to the value of the most extreme observation in which you have confidence (Tukey, 1962). Forecasters should specify the rules for determining outliers before doing any analysis in order to avoid the temptation to make adjustments to support a preferred hypothesis. Another sensible approach is to estimate the regression model by minimizing the absolute error (e.g. Dielman, 1986, 1989).

14.1. Multiple regression

What if more than one causal variable is important? Multiple regression analysis (MRA) might seem to be an obvious solution, but its use with non-experimental data leads to multicollinearity and interactions among causal variables. In addition, data on the variables are typically subject to measurement errors and validity concerns that make assessing the relative weight of each variable problematic. That complexity puts MRA at a considerable disadvantage to simple regression as a method for estimating causal relationships: MRA fails Occam’s razor.

To our knowledge, MRA was adopted without any testing of its predictive validity. The first comparative test that we are aware of involved making 10-year ahead forecasts of the populations of 100 counties in North Carolina. A multiple regression model with six causal variables was used to make the forecasts. For comparison, six simple regressions were estimated, one for each variable; their forecasts were then averaged for each county. The mean absolute percentage error of forecasts from the MRA model was 64% higher than that of the combined simple-regression model forecasts (Namboodiri & Lalu, 1971).

Another test obtained forecasts for 20 data-sets using MRA models with from 3 to 19 causal variables. The data-sets included problems such as predicting professors’ salaries and high school dropout rates. MRA was compared with an equal weights model using the same variables, and also with the simple “take-the-best” (causal variable) approach based on the forecaster’s information. The MRA produced 1% fewer correct forecasts than were obtained from equal weights models and 3% fewer than from the take-the-best approach (Czerlinski et al., 1999, Tables 5-4).

MRA forecasts of the popular vote for U.S. Presidential elections were available from eight leading political forecasters. Their accuracy was compared with those from a simple regression using the “best” variable (typically the “economy”). Forecasts were made for each of the last 100 days of the 10 U.S. Presidential election years from 1972 to 2008; a total of 1,000 forecasts. The MRA forecasts were less accurate than the simple regression model forecasts with a mean absolute error of 3.8% compared to 3.6% (Graefe & Armstrong, 2012).

14.2. Data models

Beginning in the 1960s, advances in technology made it feasible for analysts to use tests of statistical significance to select multiple “predictor variables” and estimate relationships. We refer to the resulting models as “data models.” The trend started in the mid-1900s with stepwise regression. It spawned procedures with names such as big data, analytics, data mining, machine learning, and neural nets. One claim is that objectivity is increased by
letting the data speak for themselves. As we show below, in practice these techniques have the opposite effect.

Einhorn (1972, p. 367) was among the first to warn against data models. He concluded, “Access to powerful new computers has encouraged routine use of highly complex analytic techniques, often in the absence of any theory, hypotheses, or model to guide the researcher’s expectations of results.” He likened the practice to alchemy. For a further discussion of the deficiencies of regression analysis in practice, see Armstrong (2012b).

The only scientific way to identify relationships in complex situations is to conduct experiments to identify the effects of proposed causal variables under different conditions. Data models ignore cumulative scientific knowledge, and rely only on the data.

Despite the widespread understanding that correlation does not imply causation, data models are based on statistically significant correlations: not on causal relationships but on “predictor variables.” About 32% of the 182 regression papers published in the American Economic Review in the 1980s relied on statistical significance for choosing predictor variables (Ziliak & McCloskey, 2004). The situation was worse in the 1990s, as 74% of 137 such papers did so.

Statistical significance testing is detrimental to advances in science (Armstrong, 2007a, 2007b). A theoretical analysis titled “Why most published research findings are false” demonstrated how using statistical significance testing along with testing for a preferred hypothesis leads to the publication of incorrect research findings (Ioannidis, 2005). Data models can be, and are, used to support any desired conclusions through such dubious practices as proposing hypotheses after analyzing the data, trying out variables in order to find ones that support a preferred hypothesis, discarding observations that do not support the desired hypothesis, selecting unreasonable null hypotheses, using large sample sizes to ensure statistical significance, and ignoring findings by other researchers that do not support the desired hypothesis. These procedures are common tactics in advocacy research. Armstrong and Green (2018) summarize evidence on the extent to which such questionable procedures are used in scientific journals.

Our searches have been unable to find any experimental comparisons showing that MRA or other data modeling techniques have out-of-sample predictive validity equal to that of the simple evidence-based methods identified in Figure 1. To the contrary, the evidence that we have found shows that data models are unsuited to forecasting.

A comprehensive analysis of the accuracy of data mining found that forecasts from data-mining models had consistently lower out-of-sample predictive validity than simple alternative models. In one test, the authors of the study asked a data-mining expert to make predictions using a set of data. The expert did so, and identified many statistically significant relationships in the data. Unbeknownst to the data miner, the numbers were random (Keogh & Kasetty, 2003). In personal correspondence with us, Keogh stated,

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.

15. Knowledge models

Some forecasting problems are characterized by knowledge of many important causal variables. Consider, for example, predicting which players will do well in sports, who would
be an effective company executive, which countries will have the highest economic growth, or which applicants for immigration are most likely to pose a security risk. Knowledge models are suitable for such problems.

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

A similar approach, called “experience tables,” was used in the early 1900s for deciding which prisoners should be given parole (Burgess, 1936). Another version, called “configural analysis” was used in the mid-1900s and was found to have predictive validity (e.g. see Babst et al., 1968). Yet another version was developed more recently under the term “index method” where there was considerable testing as we describe below.

We propose the name “knowledge model” because the term is more descriptive than the previous terms. Figure 2 provides a checklist for developing a knowledge model.

a. **Identify all important causal variables using domain knowledge and findings from experiments** – Follow the scientific method by using prior knowledge to identify causal variables. With knowledge models, causal variables can be as simple as binary; for example, “is taller than opponent” for an election forecasting model. In some situations, causal variables are obvious from logical relationships. In cases where they are not, consider surveying three to five domain experts. When the validity of a proposed causal variable is uncertain, consult findings from experiments, especially meta-analyses of experiments, in order to determine whether there is sufficient support for the use of the variable. To test the importance of relying on experimental evidence for specifying effects, the direction of the effect of each of 56 persuasion principles from Armstrong (2010) was obtained from non-experimental data as well as from experimental data. The findings from different experiments were in the same direction for each principle, but for only two-thirds of the principles in the non-experimental data (Armstrong & Patnaik, 2009).

b. **Discard a causal variable if it cannot be controlled, or accurately forecast** – If a causal variable cannot be forecast or controlled, including it in a model is likely to harm the accuracy of forecasts from the model.

c. **Determine the directions of causal variables’ effects on the variable to be forecast** – The directional effects of some variables are obvious from logic or common knowledge about the domain. If the direction is not obvious, refer to experimental studies. For example, opinions about the effects of gun regulations on crime vary and opposing opinions among voters and politicians have led U.S. cities and states to change their

---

**Figure 2.** Knowledge model development checklist.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>Identify all important causal variables using domain knowledge and findings from experiments</td>
</tr>
<tr>
<td>b.</td>
<td>Discard a causal variable if it cannot be controlled, or accurately forecast</td>
</tr>
<tr>
<td>c.</td>
<td>Determine the directions of causal variables’ effects on the variable to be forecast</td>
</tr>
<tr>
<td>d.</td>
<td>Determine the relative magnitudes of causal variables’ effects on the variable to be forecast if possible</td>
</tr>
<tr>
<td>e.</td>
<td>Specify model as dependent variable score equals the sum of weighted causal variables</td>
</tr>
<tr>
<td>f.</td>
<td>Estimate relationship between scores and dependent variable values by regression analysis if feasible</td>
</tr>
</tbody>
</table>
laws to either restrict or make gun ownership easier. These natural experiments provide a method to scientifically determine which opinion is correct, as was done by Lott (2010, 2016). If there is neither obviousness nor experimental evidence in its favor, discard the variable.

d. Determine the relative magnitudes of causal variables’ effects on the variable to be forecast if possible – Consider whether there is sufficient evidence that changes in some causal variables have stronger influences on the dependent variable than others. Consult experimental evidence and consider surveying domain experts to determine differential weights. Vary weights from unity only if there is strong evidence of differences in effect sizes among the causal variables. Avoid changing the a priori weights to improve in-sample fit.

e. Specify model as dependent variable score equals the sum of weighted causal variable values – Knowledge models simply calculate a score by summing the products of the signed causal variable weights and the variable values. The score is a forecast: a higher score means the outcome is more likely.

f. Estimate relationship between the score and the dependent variable values by regression analysis if feasible – Where sufficient historical data are available on the dependent variable, one can estimate the relationship between the knowledge model scores and a continuous dependent variable using simple regression analysis. Quantitative forecasts can then be obtained by applying the regression-estimated parameters – constant and score coefficient – to the knowledge model score for a particular situation.

While we believe that Benjamin Franklin was correct when he suggested considering differential weights, they should be used only when they are supported by strong evidence. For example, how much do experts know about the causal relationships, and how much experimental data is available on the relationships? For problems where domain knowledge and data are insufficient for confident estimates of differential weights, use equal weights. The first empirical demonstration of the power of equal weights was by Schmidt (1971). That was followed by Einhorn and Hogarth (1975) and Dana and Dawes (2004) who showed some of the conditions under which equal weights models provide more accurate forecasts than regression weights.

Lichtman’s “Keys to the White House” model used 13 equally weighted variables selected by an expert to forecast the popular vote in U.S. Presidential elections. The model accurately predicted which candidate won the popular vote for all elections from 1984 to date, except 2016 (Armstrong & Cuzán, 2006). Another equal-weights election forecasting model included all of the 27 variables that had been used in nine independent econometric (multiple regression) models. The ex ante average forecast error was 29% lower than the average error of the most accurate of the 10 original regression models (Graefe, 2015). Graefe and Armstrong (2013) reviewed empirical forecasting studies in psychology, biology, economics, elections, health, and personnel selection. Knowledge models provided more accurate forecasts than did regression models for ten of the 13 studies.

Even when there is a strong case for differential weights, consider adjusting the weights toward equality. Equalizing was tested in election forecasting using eight independent econometric election forecasting models estimated from data that was standardized and positively correlated with the dependent variable. Where equalizing coefficients by 100%
amounts to using equal weights, equalizing by between 10 and 60% reduced the absolute errors of the forecasts for all of the models. (Graefe, Armstrong, Jones, & Cuzán, 2014).

A study assessed the predictions from a knowledge model of the relative effectiveness of the advertising in 96 pairs of advertisements used differential weights influenced by experimental evidence. With 195 potentially relevant variables, regression was not feasible. Guessing would result in 50% correct predictions of which of each pair was more effective. Judgmental predictions by novices were correct for 54% of the pairs; those with experience in advertising made 55% correct predictions. Copy testing (e.g. showing ads to subjects and asking them to assess their likelihood of purchase) yielded 59% correct predictions. In contrast, the knowledge model forecasts were correct for 75% of the pairs of advertisements – an error reduction of 37% compared to copy testing (Armstrong, Du, Green & Graefe 2016). In an extension of the study, the model was tested using a variation of equal weights. At 32%, the resulting error reduction was broadly similar to that from the original model (Green, Armstrong, Du & Graefe 2016).

16. and 17. Combining forecasts

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

The basic rules for combining within and across methods are: (1) obtain forecasts that are the products of diverse experts, data, procedures, and variations of all valid evidence-based methods; (2) for each component method, combine forecasts from the variations by calculating equally weighted averages; (3) combine the combined forecasts from the component methods by calculating an equally weighted average across the methods used. The rules for equal weighting should only be relaxed if there is strong evidence of differences in forecast accuracy, in which case, the weights should be specified before making the forecasts.

For important problems, we suggest obtaining forecasts from at least two variations of each component method, and from three different component methods. That is, combine across combined forecasts in order to improve reliability and validity. For more details on combining forecasts, see Graefe, Armstrong, Jones, & Cuzán (2014) and Graefe (2015).

The combining procedures described guarantee that the resulting forecast will not be the worst forecast, and that it will perform at least as well as the typical component forecast. In addition, the absolute error of the combined forecast will be smaller than the average of the component forecast errors when the components’ range includes (brackets) the true value. Combined forecasts can be, and often are, more accurate than the most accurate component forecast. Because bracketing is always possible, combining should always be used when two or more forecasts from evidence-based methods can be obtained.

Combining is not intuitive. In a series of experiments with highly qualified MBA students, a majority of participants thought that averaging estimates would deliver only average performance (Larrick & Soll, 2006). In another experiment, a paid panel of U.S. adults were given data on five experts’ recent forecast errors in predicting attendance at film screenings. When asked to nominate which experts forecasts they would combine for forecasting attendance at future screenings, only 5% of the 203 participants chose to use forecasts from all five experts. The rest chose to combine only the forecasts of the experts whose previous errors had been smallest (Mannes et al., 2014).
With the same intuition, when New York City officials received two different forecasts for an impending snowstorm in January 2015, they acted on the forecast that they believed would be the best. As it turned out, it was the worst.

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

Combining forecasts from variations of a single method or from independent forecasters using the same method helps to compensate for mistakes, errors in the data, and small sample sizes in any of the component forecasts. In other words, combining within a single method is likely to be most useful for improving the reliability of forecasts.

One review identified 30 studies that compared combinations of forecasts mostly from a single method. The unweighted arithmetic mean error of the combined forecasts was 12.5% smaller than the average error of the typical forecast, with a range from 3 to 24% (Armstrong, 2001c).

Another study compared the accuracy of the forecasts from eight independent multiple regression models for forecasting the popular vote in U.S. Presidential elections with the accuracy of an average of their forecasts. The combined forecasts reduced error compared to the typical individual model’s forecast by 36% across the 15 elections in the study (Graefe et al., 2014).

Different forecasting methods are likely to have different biases because they utilize different assumptions, knowledge, and data. As a consequence, forecasts from diverse methods are more likely than those from a single method to bracket the actual outcome. Moreover, by including more information about the situation, combining forecasts across multiple methods is also likely to increase reliability. For example, one study examined the effect of combining time-series extrapolations and intentions forecasts on accuracy. The study found that combining forecasts from the two methods reduced errors by one-third compared to extrapolation forecasts alone (Armstrong et al., 2000).

Consider also the case of combining the forecasts of economists who ascribe to different economic theories. In one study, combinations of 12-month ahead real GNP growth forecasts from two economists with similar theories reduced the mean square errors by 11% on average, whereas combinations of forecasts from two economists with dissimilar theories reduced errors by 23%. Combinations from pairs of economists who used similar forecasting techniques reduced errors by 2%, while combinations from pairs who used dissimilar techniques yielded a 21% error reduction (Batchelor & Dua, 1995, Table 2). The error-reduction advantage for diversity in combinations was much larger for five of the six other comparisons in the study, in which economists with similar/dissimilar theories/techniques forecast the GNP deflator, corporate profit growth, and the unemployment rate.

The PollyVote.com election-forecasting project provided data for testing the accuracy of combining forecasts across four to six different methods for predicting the popular vote in the seven U.S. Presidential elections from 1992 to 2016. The individual method forecasts (e.g. from election polls) were first combined. Combined forecasts from several methods were then combined. Over the 100 days prior to the elections, the mean absolute error of
the PollyVote forecast was, at 1.1 percentage points, smaller than the average errors of each of the component combinations which ranged from 1.2 to 2.6 percentage points with a median of 1.8 (Graefe et al., 2017).

Combining across methods provided an error reduction of roughly 40% relative to the typical single method combination. Taken together with the previously mentioned error reduction of 12.5% for combining within a method (Armstrong, 2001c), a crude estimate of the expected error reduction from combining within methods then across methods is that it would be more than one-half.

**Forecasting principles: Golden Rule and Occam’s razor**

We turn our attention now from methods to principles. The forecasting methods listed in the Figure 1 checklist are consistent with forecasting principles, so following the Forecasting Methods Application Checklist can help to ensure that the principles are adhered to. More importantly, however, forecasters who persist in using methods other than those listed in Figure 1 can greatly improve the accuracy of their forecasts if they take steps to comply with two overarching forecasting principles: the Golden Rule, and Occam’s razor.

The Golden Rule and Simple Forecasting checklists described below provide guidance on how to comply with the two principles. They differ from a previously published checklist of principles – the Forecasting Audit checklist, available at ForecastingPrinciples.com – which is intended for forecasting academics and practitioners. For example, we used the Forecasting Audit checklist to assess the forecasting procedures used to produce the U.N. Intergovernmental Panel on Climate Change projections of global mean temperatures (Green & Armstrong, 2007b).

In contrast, the checklists that we present in this section are intended to empower all interested parties to conduct audits of forecasting procedures. The two principles checklists apply to all types of forecasting problems, and to all forecasting methods.

**Golden Rule**

The Golden Rule is to be conservative. More specifically, to be conservative by adhering to cumulative knowledge about the situation and about forecasting methods. (Armstrong et al., 2015).

The Golden Rule of Forecasting is also an ethical principle, as it implies “forecast unto others as you would have them forecast unto you.” The Rule is a useful reference when objectivity must be demonstrated, as is the case in legal or public policy disputes (Green et al., 2015).

Figure 3 is a revised version of Armstrong, Green, and Graefe’s (2015, p. 1718, Table 1). It includes 28 guidelines logically deduced from the Golden Rule of Forecasting.

There are two key changes from the previously published version. The first is that Guideline 5 now includes the injunction to “combine forecasts from diverse methods.” The change is based on the evidence presented in the previous section.

The second change is that Guideline 6 originally suggested caution in using judgmental adjustments, but is now a prohibition: one should “avoid adjusting forecasts.” The primary reason is that the use of diverse methods leads to increased use of information about the
situation, and hence a lower likelihood of bias arising due to the omission of key information. Moreover, adjustments are liable to introduce intentional bias. For example, a survey of nine divisions within a British multinational firm found that 64% of the 45 respondents agreed that “forecasts are frequently politically modified” (Fildes & Hastings, 1994). In another study, 29 Israeli political surveys were classified according to the independence of the pollster from low to high, as “in-house” – such as a poll run by a political party – “commissioned,” or “self-supporting.” The independent polls provided forecasts that were more accurate than the in-house pollsters. For example, 71% of the independent polls had relatively high accuracy, whereas 60% of the most least in dependent polls had relatively low accuracy (Shamir, 1986, Table 4).

Figure 3. Golden Rule of forecasting checklist: Version 2.
Meehl (1954) concluded that forecasters should not make subjective adjustments to forecasts made by quantitative methods. Since then, research in psychology has continued to support Meehl’s findings (see Grove et al., 2000). Research on adjusting forecasts from statistical models found that adjustments often increase errors (e.g. Belvedere & Goodwin, 2017; Fildes et al., 2009) or have mixed results (e.g. Franses, 2014; Lin et al., 2014).

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

We have been unable to find any evidence that adjustments would reduce forecast errors relative to the errors of forecasts derived in ways that were consistent with the guidance presented in this paper. In particular, following Guideline 1.1.2 – to decompose the forecasting problem to make best use of knowledge, information, and judgment – and the revised Guideline 5 – to combine forecasts from diverse methods – helps to ensure that all relevant knowledge and information are included in the forecast, leaving no valid reason for adjusting forecasts.

Evidence was found on the effects of 18 of the Golden Rule guidelines. On average, the violation of a typical guideline increased error by 40%, as detailed in Figure 3. Errors can be expected to accumulate as more guidelines are violated. Although, we have no systematic information on the extent that the Golden Rule is followed in practice, we expect that forecasting studies published in scientific journals typically violate most of the Golden Rule Guidelines. For example, our audit concluded that the U.N.’s International Panel on Climate Change ignored the Golden Rule in deriving their projections of dangerous man-made global warming.

Any stakeholder can use the Golden Rule of Forecasting Checklist. Experts and non-experts can complete the Golden Rule of Forecasting Checklist in less than an hour. Stakeholders do not need to be forecasting experts to use the checklist because the onus is on forecasters to fully and clearly disclose their methods (Guideline 1.3.) To help improve the reliability of the checklist ratings, stakeholders could ask at least three people, each working independently, to rate the forecasting procedures and then average the ratings.

Simple forecasting

The “simplicity principle” (Occam’s razor) is the scientific principle that the simplest explanation of evidence is the best. The principle was proposed by Aristotle and later named after fourteenth-century scholar William of Ockham. The principle also applies to scientific forecasting: forecasters should use methods that are no more complex than necessary to develop the simplest model that is consistent with knowledge about the situation.

Do forecasters ascribe to Occam’s razor? Apparently not: in 1978, when 21 of the world’s leading experts in econometric forecasting were asked whether more complex econometric methods produced more accurate forecasts than simple methods, 72% replied that they did. In that survey, “complexity” was defined as an index reflecting the methods used to develop the forecasting model: (1) the use of coefficients other than 0 or 1; (2) the number
of variables; (3) the functional relationship; (4) the number of equations; and (5) the use of simultaneous equations (Armstrong, 1978).

Starting in the 1950s, researchers developed complex statistical models to extrapolate time-series data. They derived models using mathematics, and reported on the ability of the models to fit data. The models were popular and widely used by academics and practitioners, but their predictive validity was not tested against alternative methods.

In the late 1970s, researchers were invited to enter their models in a competition to extrapolate 111 unidentified business and economic time series of monthly, quarterly, and annual data up to six years ahead. The accuracies of the forecasts from the different methods were assessed against those of the relevant no-change benchmark model forecasts. The simple naïve models performed well, with only minor differences in accuracy compared with forecasts from the more complex models. The findings were published with commentary by 14 leading statisticians (Makridakis & Hibon, 1979). Makridakis went on to conduct extensions of the competitions – which were referred to as the M-competitions (Makridakis & Hibon, 2000; Makridakis et al., 1993) – that led to the conclusion that simple methods provide extrapolation forecasts that are competitive with those from complex methods.

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

We proposed a new operational definition of simplicity in forecasting, one that could be assessed by any stakeholder. It consisted of a four-item checklist to rate simplicity in forecasting as the ease of understanding by a potential forecast user. The checklist was created before any analysis was done and it was not changed as a result of testing. Figure 4 provides an abridged version of the checklist provided on ForecastingPrinciples.com (Green & Armstrong, 2015).

Green & Armstrong’s (2015) search identified 32 published papers that allowed for a comparison of the accuracy of forecasts from simple methods with those from complex methods. Four of those papers tested judgmental methods, 17 tested extrapolative methods, 8 tested causal methods, and 3 tested forecast combining methods. The findings were consistent across the methods. On average, across each comparison, the more complex methods produced ex ante forecast errors that were 27% larger than those from simpler evidence-based methods. The finding was surprising because the papers had apparently proposed the more complex methods with the expectation that they would provide more accurate forecasts. To our knowledge, complex methods have never been shown to provide forecasts for complex situations that are as accurate as those from simple evidence-based methods. The late Arnold Zellner, founder of the Journal of Econometrics, reached the same conclusion.3

Are the descriptions of the following aspects of the forecasting process sufficiently uncomplicated as to be easily understood by decision makers?

1. method
2. representation of cumulative knowledge
3. relationships in models
4. relationships among models, forecasts, and decisions

Simple Forecasting Average (out of 10)

Simplicity rating (0–10)

[ ]

Figure 4. “Simple forecasting” checklist: Occam’s razor.
Assessing forecast uncertainty

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

This section first describes error measures for estimating prediction intervals. Currently, the estimates of prediction intervals are typically much too narrow. We suggest doubling the width of statistically estimated 95% confidence intervals to approximate the likely 95% prediction intervals. But use of the guidelines below, in Figure 5, would be better.

Error measures

Earlier, we discussed error measures suitable for evaluating forecasting methods by comparing the accuracy of their forecasts with those from alternative methods. Here, for the purpose of estimating prediction intervals that are useful for managerial decisions we suggest the mean absolute deviation (MAD) of forecasts from actual values. The MAD is easy to calculate, and is easily understood by decision-makers. On the other hand, the commonly used root mean square error measure should be avoided as it cannot be related to benefits.

For forecasting problems that are expected to involve asymmetric errors – i.e. negative errors are larger than positive errors, or vice versa – calculate the logarithms of the forecast and actual values and calculate the errors using the logged values. Use those errors to estimate prediction intervals, and then convert the bounds of the intervals back to actual values (Armstrong & Collopy, 2001, p. 281).

Loss functions can also be asymmetric. For example, the losses due to a forecast that is too low by 50 units may differ from the losses if a forecast is too high by 50 units. Asymmetric losses are, however, a problem for the planner, not the forecaster: the planner must assess the damages resulting from positive versus negative errors.

Methods to forecast uncertainty

Figure 5 presents a checklist of methods to forecast uncertainty. The checklist includes four valid methods to use, and two commonly used but invalid methods to avoid.

1. Use empirical prediction intervals or likelihoods estimated from out-of-sample tests
2. Decompose errors by source in order to estimate the uncertainty of each
3. Use structured judgment to estimate prediction intervals or likelihoods
4. Combine alternative valid estimates of uncertainty
5. Avoid using statistical fit with historical data to assess uncertainty
6. Avoid using tests of statistical significance to assess uncertainty

Figure 5. Methods to forecast uncertainty checklist.
1. **Use empirical prediction intervals or likelihoods estimated from out-of-sample tests**

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

   In order to provide forecast users with useful information on forecast uncertainty, there is no alternative to estimating empirical prediction intervals based on out-of-sample forecast errors. To that end, simulate the actual forecasting procedure as closely as possible and use the distribution of the errors of the resulting forecasts to assess uncertainty. Tashman (2000) provides guidance on out-of-sample testing. For more on estimating prediction intervals, see Chatfield (2001).

   When analyzing time-series forecast errors, use successive updating to increase the number of predictions. If sufficient validation data are not available, consider using data from analogous situations.

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

   Most forecasting problems are subject to several sources of forecast error. To help ensure that all possible errors are accounted for, consider decomposing errors by source of error to estimate each, then combine the estimates. For example, when polling to predict the outcomes of political elections, survey researchers report only the error expected due to random variation based on the size of the sample. Response and non-response bias errors are ignored. As a consequence, the 95% confidence intervals reported for polls are about half as large as they should be (Buchanan, 1986). In other words, decision-makers should double political polls’ confidence intervals to obtain more realistic estimates of the prediction intervals.

   When uncertainty is high – such as with surveying citizens to forecast their behavior in response to changes in government regulations – response error is likely to be particularly high due to survey respondents’ lack of self-knowledge about how they make decisions (see Nisbett & Wilson, 1977). Non-response can also be a large source of error because people who are most affected by the topic of the survey are more likely to respond. While the latter error can be reduced to some extent by the “extrapolation-across-waves” method (Armstrong & Overton, 1977), forecasters still need to consider that source of error when assessing uncertainty.

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

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

   One common judgmental approach to assessing uncertainty is to ask experts to express their confidence in their own judgmental forecasts in the form of 95% prediction intervals. One concern with that approach is that experts are typically overconfident about the accuracy of their forecasts. For example, an analysis of judgmental confidence intervals for economic forecasts from 22 economists over 11 years found that outcomes were within the range of their 95% confidence intervals only 57% of the time (McNees, 1992). Another study tracked members of a 10-year panel who provided 13,300 estimates of expected stock
market returns by company; the actual returns were within the executives’ 80% confidence intervals only 36% of the time (Ben-David et al., 2013).

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

Finally, to improve the calibration of forecasters’ estimates of uncertainty in the future, ensure that they receive timely, accurate, frequent, and well-summarized information on what actually happened and reasons why their forecasts were right or wrong. For example, weather forecasters use such procedures, and their forecasts are well calibrated for a few days ahead: When they say that there is a 40% chance of rain, rain falls 40% of the time on average (Murphy & Winkler, 1984).

4. Combine alternative valid estimates of uncertainty
The logic behind combining uncertainty estimates is the same as that for combining forecasts. Thus, the estimates of uncertainty based on combined estimates can never be worse than the typical estimate, and the combined estimate will always be better than the typical estimate as long as bracketing occurs.

5. Avoid using statistical fit with historical data to assess uncertainty
In a study using data consisting of 31 observations on 30 variables, stepwise regression was used with a rule that only variables with a \( t \)-statistic greater than 2.0 would be included in the model. The data were from a book of random numbers. Despite that, the stepwise method delivered an eight-variable regression model with good statistical fit – an \( R^2 \) of .85 adjusted for degrees of freedom (Armstrong, 1970).

Measures of statistical fit do not provide useful information about out-of-sample predictive validity (Armstrong, 2001d). Experiments testing analysts’ interpretation of standard statistical fit information on regression models found that 72% grossly underestimated the uncertainty of forecasts associated with changes to the model’s causal (policy) variable (Soyer & Hogarth, 2012). Further discussion on why forecasters should avoid such measures as adjusted \( R^2 \) is provided in Armstrong (2001d).

6. Avoid using tests of statistical significance to assess uncertainty
Statistical significance tests do not provide estimates of forecast uncertainty. Attempts to use them in that way will likely lead to confusion and poor decision-making. Experimental studies over more than half a century support that conclusion (e.g. see Armstrong & Green, 2018; Hubbard, 2016; Ziliak & McCloskey, 2008).

One experiment presented leading researchers with a treatment difference between two drugs, as well as a “\( p \)-value” for the difference, and asked them which of the drugs they would recommend to a potential patient. When the difference in the effects of treatments was large but reported to be \( p > .05 \), nearly half responded that they would advise that there was no difference between the two drugs. By contrast, when the difference between the treatment effects was small but reported to be statistically significant (\( p < .05 \)), 87% of the respondents replied that they would advise taking the drug (McShane & Gal, 2015). Many of those teaching statistics also failed to draw logical conclusions as was shown in another experiment by McShane and Gal (2017).
Errors in interpretation of findings due to the provision of statistical significance information have led to poor decisions. Hauer (2004) described the harm caused by decisions related to automobile traffic safety, such as the “right-turn-on-red” policy. Ziliak and McCloskey (2008) provide other examples.

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

**Discussion**

The accumulation of scientific knowledge about forecasting over the past century enables improvements in forecast accuracy. Regrettably, that knowledge is often ignored, and forecasting practice appears to be in decline. There are two related reasons: first, advocacy research has tended to replace objective forecasting; and second, an unsupported faith in data models has resulted in forecasters ignoring cumulative knowledge about causal relationships and validated methods (Armstrong & Green, 2018).

What recourse do clients and citizens have when they make decisions on the basis of forecasts that turn out to be inaccurate? The answer has traditionally been that there is none, because it has been impossible to distinguish between forecasts that were wrong due to random or unpredictable changes in the situation and those that were wrong due to the forecaster’s failure to follow evidence-based procedures. This paper follows in the footsteps of medicine, engineering, and aviation by providing checklists that can be used to hold forecasters – including scientists who make forecasts, and public policy-makers – responsible if they fail to follow evidence-based forecasting procedures.

Forecasters can use the checklists to improve the accuracy of their forecasts and, by communicating that they have followed the checklists, protect themselves from claims against them. Forecasters who follow the checklists might also – as do medical practitioners – obtain protection against damage claims by arranging insurance on the understanding that they follow the forecasting procedures required by the checklists.

Science requires that the predictive validities of hypotheses and new findings are tested. Milton Friedman (1953) viewed out-of-sample predictive validity testing of competing hypotheses as an essential element of economics as a social science. The checklists in this paper can help scientists to design such tests.

**Conclusions**

Forecasting practice can be improved such that the accuracy of forecasts upon which decision-makers in business and public policy depend is greatly increased. The best way to achieve that objective is to require forecasters to comply with evidence-based checklists.

Figure 1, the Forecasting Methods Application Checklist, lists 15 individual evidence-based forecasting methods. The use of those methods substantially improves the ex ante accuracy of forecasts relative to forecasts from commonly used methods, including
experts’ unaided judgments. Error reductions range from approximately 5% – for damped-trend extrapolation and decomposition by seasonality – to over 50% – for simulated interaction and knowledge models.

Rather than hoping to identify the one best model, forecasters should employ diverse models from different evidence-based methods, and combine the forecasts from them. Doing so reduces the bias than can arise from using a single method and improves reliability by incorporating more knowledge and information. Combining also avoids the risk of making the worst forecast and guarantees that the combined forecast will be more accurate than the typical forecast if, as is likely, any of the forecasts in the combination bracket the outcome.

Data models are not suitable for forecasting. In particular, multiple regression approaches violate evidence-based forecasting principles and provide forecasts that are substantially less accurate that those from the methods listed in the Figure 1 checklist. Data models can be and are being used to support clients’ and funders’ prior beliefs and preferences to the further detriment of forecast accuracy.

The Golden Rule and Simple Forecasting principles checklists (Figures 3 and 4) can help forecasters to implement the evidence-based methods listed in Figure 1, and can help forecasters to improve currently used methods in situations where it is not feasible to replace them with evidence-based methods. Following the Golden Rule of Forecasting can help forecasters reduce forecast errors by over half, while ignoring Occam’s Razor is likely to increase errors by around 27%. The principles checklists can also help clients, sponsors, and users to assess whether proper procedures were followed.

Procedures currently used to assess forecast uncertainty mislead analysts, clients, and users into excessive confidence. As a rule of thumb, they are half as uncertain as they should be. Prediction intervals should only be estimated using out-of-sample testing, as described in the Figure 5 checklist.

Clients and other funders who are interested in accurate forecasts should require forecasters to follow the five evidence-based checklists provided in this paper, and should audit the forecasters’ procedures to ensure that they did so. Clients and other forecast stakeholders can use the checklists to assess the worth of forecasts by determining whether they were the product of scientific forecasting procedures.

引言

预测对于企业和其他组织以及政府的决策制定至关重要。从对从业人员、教育工作者和决策者的调查中发现，在13个评判预测的标准中，他们将“准确率”评为最重要的一个标准(Yokum and Armstrong, 1995)。研究人员特别关心准确性。与此发现相一致，提高预测准确率是本文的主要关注点。

尽管预测知识的进步使得预测准确率有了实质性的提高，但是在学术期刊文章中，却大多忽略了这一知识，我们由此预见，从业者也忽视了这一点。在2001年发布最初的139个预测原则的时候，从17本预测教科书的回顾中发现，这些典型书籍仅提及到19%的预测原则(Cox and Loomis, 2001)。此外，预测软件包，本可有助于确保原则使用率，却忽略了近一半预测原则的使用(Tashman and Hoover, 2001)。

预测改善清单

使用循证清单就不必记忆和简化复杂任务。在医学、航空和工程等领域，若遵循了不当的的清单就可能会引发诉讼。

许多研究支持使用清单（例如，Hales and Pronovost, 2006）。曾有一项实验评估了一份有19个项目的清单在医疗程序中的应用效果。该研究在全球选取八个国家的医院，比较了这些医院中使用清单前后数千例患者的治疗效果。在接受医疗程序后一个月中，死亡率从1.5%减低到了0.8% (Haynes et al. 2009)。重要的是，即使清单中的知识为从业者所熟知，使用清单仍可改善决策（Hales and Pronovost, 2006）。为了确保清单中包含最新的证据，应该定期修改清单。

说服人们使用清单很容易。告知工程师和医生必须使用清单是他们的就业条件，并且监控清单的使用时，他们会就会使用清单。当我们支付给人们适量的酬金并要求他们使用清单来完成任务时，几乎所有接受这项任务的人都很有效率地完成了任务。例如，为评估印刷广告的说服力，通过亚马逊劳务众包平台雇佣的评估人员使用了一个有195项项目的清单，以评估广告是否符合说服原则。测试者间信度很高（Armstrong, Du, Green, and Graefe, 2016）。

研究方法

我们回顾了之前以提高预测准确性而进行的预测方法和原则的实验研究。为此，我们首先通过以下途径确定相关研究：

1) 网上搜索，主要使用谷歌学术进行搜索；
2) 联系领军的研究人员，寻求他们对重要实验成果的建议；
3) 核查在实验研究和元分析中提到的关键论文；
4) 将我们的工作文件发布于网上，请求提供我们可能忽略的证据；
5) 在本文的开放存取版本中提供所有论文的链接，以便读者查看我们对这些原始研究成果的解释。

鉴于大量论文的题目与我们的研究内容相关，所以我们筛选论文的标准是：评估“摘要”和“结论”部分是否提供了替代方法比较值和充分披露的证据。只有少数这样的论文符合这些标准。

只有进行了许多样本外预测的研究，才可作为本文的证据。横截面数据，其“刀切法”程序可进行许多预测，方法是使用除一个数据点以外的所有数据点来估计模型，对排除的观测值进行预测，然后替换该观测值并排除其他值，如此反复，直到对所有数据点都进行了预测。连续更新的使用，能够增加时间序列数据的样本外预测数量。例如，替代模型可预测未来100年全球平均气温，为测试其预测效果，从1851年开始，对全球平均气温的年度预测从1年提前到了100年。从1851年，到1852年，1853年，……这些预测不断更新，从而提前1年为157个预测提供误差……直到提前100年为58个预测提供误差 (Green, Armstrong, and Soon, 2009)。
本文所引用的所有论文，若涉及到实质性发现，我们都尝试着联系其作者。我们这样做是因为有证据表明，一些主要的科学期刊论文中引用的研究成果经常被错误地描述。（Wright and Armstrong, 2008）。我们向作者询问：我们对其研究成果的总结是否正确，我们的描述是否还可以完善。还请求他们向我们推荐一些论文，这些论文可能在我们所忽略的——特别是那些描述实验的论文，那些实验取得了研究成果而且成果与我们的结论相冲突。事实证明，与计算机搜索相比，这种做法有助于更为全面地搜索证据（Armstrong and Pagell, 2003）。有六篇论文，我们不同意作者对文中实验成果的阐释。是否引用这些论文不影响本文的写作目的，所以我们放弃了对这些论文的引用。

90篇论文中具有实质性发现，但这些发现不归属于我们，我们能够联系到其中73名论文的作者，并且从69名作者那里收到实质性且有帮助的回复。我们在本文的参考文献部分对这些论文进行了编码，包括我们与作者联系的结果。

我们的核查完善了五个清单。它们提供预测方法循证指南、知识模型循证指南、预测的黄金法则循证指南、简单性原则循证指南和不确定性循证指南。

有效预测方法: 清单和证据

评估一种预测方法的预测效果，是比较该方法与另一种方法的预测准确率，比较该方法与目前所使用方法的预测准确率，或比较该方法与简单的方法（例如朴素无趋势模型）的预测准确率，亦或比较该方法与其他循证方法的预测准确率。正如张伯伦所述，这样检验多种合理的假设是科学方法的要求（1890）。

对于分类预测（例如a, b, c是否会发生，或者哪个更好）来说，准确率是正确率的一种变体，通常是由正确率来衡量。对于定量预测来说，评价准确率要看样本外预测和实际发生预测的数据，并比较两者之间的差异。评估预测方法的基准误差度量单位是相对误差绝对值（即“RAE”），它已被证明比均方根误差更可靠（Armstrong and Collopy, 1992）。还有一种新方法被称为非标定平均有界相对误差绝对值（UMBRAE），即相对误差绝对值的演化，对这种新方法的测试表明，它优于相对误差绝对值以及其他所推荐的度量单位（Chen, Twycross, and Garibaldi, 2017）。我们建议使用相对误差绝对值和非标定平均有界相对误差绝对值，除非不得不进行额外测试来得出一个确切的结论，即哪一个度量单位更好。

附表1列出了15个独立的循证预测方法。它们符合预测原则，并已被证实能进行高峰准确率的样本预测。该附表还标识了使用每种方法时所需要的知识。建议方法内部和方法之间进行结合（清单第16，17项）。

对于大多数预测问题，我们可使用多种方法而且应该使用，如下所述。读者可从ForecastingPrinciples.com“方法清单”下的顶部菜单栏中查看电子版的“附件1”清单。

常用于实践的方法中，我们所关注的方法确实可以提高预测准确率，所以我们不讨论所有用于预测的方法。例如，在开发预测模型中，我们最广泛使用的方法之一显然是多元回归分析。鉴于文本总结的证据，我们建议不要使用多元回归分析和其他数据建模方法。

客户应询问预测者使用的方法和其原因。如果他们提到的方法在附表1中没有列出，则应该要求他们提供证据，证明与附表中列出的方法相比，使用他们的方法做出的预测误差更小。

基于相似情景经验的专业知识对于预测很有帮助。经验可以演变成简单的方法，或者是启发式方法，为快速决策提供快速预测。例如，美国航空公司1549航班的紧急降落——“哈德逊岛上的奇迹”——是成功的，因为飞行员使用凝视启发法预测哈德逊河上的降落是个可行的选择。

本文所引用的所有论文，若涉及到实质性发现，我们都尝试着联系其作者。我们这样做是因为有证据表明，一些主要的科学期刊论文中引用的研究成果经常被错误地描述。（Wright and Armstrong, 2008）。我们向作者询问：我们对其研究成果的总结是否正确，我们的描述是否还可以完善。还请求他们向我们推荐一些论文，这些论文可能在我们所忽略的——特别是那些描述实验的论文，那些实验取得了研究成果而且成果与我们的结论相冲突。事实证明，与计算机搜索相比，这种做法有助于更为全面地搜索证据（Armstrong and Pagell, 2003）。有六篇论文，我们不同意作者对文中实验成果的阐释。是否引用这些论文不影响本文的写作目的，所以我们放弃了对这些论文的引用。

90篇论文中具有实质性发现，但这些发现不归属于我们，我们能够联系到其中73名论文的作者，并且从69名作者那里收到实质性且有帮助的回复。我们在本文的参考文献部分对这些论文进行了编码，包括我们与作者联系的结果。

我们的核查完善了五个清单。它们提供预测方法循证指南、知识模型循证指南、预测的黄金法则循证指南、简单性原则循证指南和不确定性循证指南。
马列夫斯基和吉仁泽（Gerd Gigerenzer）和马克思普朗克柏林人类发展研究所的ABC小组进行的大量研究发现，对于许多实际问题，简单的启发式优于更复杂和信息密集的方法。对于存在两个或两个以上重要因果因素的情况，专家并没有经常对他们的预测准确性进行总结性的反馈，但专业知识和经验本身并没有明显的价值。这种情况在商业和政府决策中很常见。即使是领先专家的独立判断预测也常常是灾难性的错误，有时甚至正中媒体下怀（e.g., 见Cerf and Navasky, 1998; Perry, 2017）。

研究专家对复杂情况的独立判断预测的准确性可以追溯到20世纪初。对这项研究的早期回顾导致了先知理论：“无论有多少证据存在先见也不存在，迷信于先知存在的人将会为此付出代价”（Armstrong, 1980）。先知理论多年来一直盛行，特别是一项为期20年的研究比较了专家和新手预测的准确性和单纯原则的准确性，并为先知理论提供了支持（Tetlock, 2005）。

尽管应该避免独立的专家判断，但专题专家可以在预测何时使用基于证据的方法进行判断时起到至关重要的作用。下一节将介绍九种使用专家判断进行预测的结构化方法。

### 1. 预测市场

预测市场——也被称为博彩市场、信息市场和期货市场——自16世纪以来就被用于预测（Rhode and Strumpf, 2004）。货币奖励吸引了相信自己有知识和信息对投注情况进行准确预测的人。

预测市场在知识分散及参与者反复交易的情况下特别有用，当出现新信息时，市场可以迅速修饰预测。使用预测市场的预测者需要熟悉设计预测市场和调查。
预测市场预测的准确性在商业预测领域进行了八次公开的比较测试（Graefe, 2011）。结果是混合的。例如，在一个比较中，预测市场的样本外预测误差比没有变化的样本预测误差小28%。另一方面，三项比较中的两次，人均判断的平均水平超过了市场预测。

2. 乘法分解

乘法分解一直是预测的关键因素。谷歌搜索“分解”、“预测”或“预测”，在2017年12月有超过4500万的结果。

乘法分解包括将预测问题分为若干部分，分别预测各部分，并将各部分的预测相乘以预测整体。例如，为了预测一个品牌的销售量，公司可以单独预测总的市场销售额和市场份额，然后再乘以这些成分。当合适的预测方法、数据可用性和因果因素的方向效应在各个部分之间变化时，预测分解对于减少预测误差是最有效的。


另一项研究使用图形软件以市场竞争模型（Makridakis et al. 1982）的方式显示68个月度序列，这些方式旨在帮助用户使用他们的判断来独立识别和预测季节性和趋势。研究发现，三位具有时间序列分析和软件知识的研究生做出未来一至十二个月的预测，其误差比领先的市场竞争模型消除季节变动后的单指数平滑法的误差要低7%。由35个新手预测五个时间序列的软件辅助判断分解的误差减少了5%（表二，Edmundson, 1990）。

3. 意向调查

意向调查问人们如何计划在特定的情况下行事。例如，它们可以用来预测人们如何对产品设计的重大变化。一项荟萃分析包括与超过10,000名受试者进行的47项比较，另一项提供了一项涉及超过83,000名受试者的10项元分析的荟萃分析。他们都发现人们的意图和他们未来的行为之间有着密切的关系（Kim and Hunter, 1993; Sheeran, 2002）。

为了评估人们的意图，预测者应该对形势做一个简短而不偏不倚的描述（Armstrong and Overton, 1971）。意图应该被表达为0 = “没有机会，或者几乎没有机会（百分之一）”的概率，到10 = “肯定的或几乎确定的（百分之九十九）”。人们的回应可以用来计算人们将如何表现，如“3.2%的人口将在未来三个月内购买产品”（Morwitz, 2001）。

提问的方式会对答复产生很大的影响。减少回复错误的两种方法是：（1）预先测试问题以确保被调查者按照预测者的意图来理解他们;（2）用其他方式来表达问题，然后再寻求出折中的回复。有关更多建议，请参阅Bradburn, Sudman和Wansink（2004）。

包括在回答调查问卷的同时给予金钱上的奖励可以减少不答复的错误（Armstrong and Yokum, 1994）。预测者应在随访中重新发送问卷给无回复者。这样做可以使人们通过分离序列的方法来估计不答复而造成的影响（Armstrong and
J. S. ARMSTRONG AND K. C. GREEN

Overton, 1977)。Dillman, Smyth和Christian描述了选择样本和获得高回复率的其他基于证据的程序（2004）。

4. 期望调查
期望调查询问人们对他人表现的期望。期望与意图不同，因为人们意识到情况会改变。例如，你被问及是否打算在明年买车，你可能会说你并无打算。但是，你意识到自己的车可能会出现严重问题。因此，你可能会期望有机会买辆新车。同意向调查一样，期望调查应该使用概率尺度，遵循基于证据的调查设计程序，使用代表性样本，获得较高的答复率，并通过外推法来纠正不答复偏差。

美国政府1932年政治选举的预测市场被禁止之后，引入了期望调查（即对潜在选民有代表性的样本进行民意调查）（Hayes, 1936）。这些“公民预期”调查正确地预测了1932年到2012年217次调查中89%的美国总统选举的普选票获胜者。此外，从1988年至2012年(Graefe, 2014)的七次美国总统选举中公民期望提供了比民意调查、预测市场、模型和专家更准确的样本外抽样预测，2016年的结果依然如此。自1992年至2016年每到选举前的100天，公民期望预测7次美国总统普选票结果的误差平均为1.2个百分点。相比之下，一个典型的民意调查结果的误差是2.6个百分点，其两倍多(Graefe, Armstrong, Jones, and Cuzán, 2017)。

5. 专家调查
使用书面的问题和自填方式调查以确保每位专家都回答同样的问题。采用与上述期望调查所述相同的程序来开发问题。
德尔菲法是专家调查方法的延伸，调查进行两轮或多轮。在每一轮之后，专家的预测的总结及理由会提供给下一轮的专家。重复这个过程，直到轮次之间的预测几乎一致——通常两轮或三轮就足够了。德尔菲预测使用专家最后一轮预测的中位数或模式。如果不同的专家每个人都针对同一问题不同方面的信息（Jones, Armstrong and Cuzán, 2007），那么德尔菲法就是最有用的。
在五项研究中，来自德尔菲法的预测比传统会议的预测更准确；在两项研究中，二者精度相同；在一项研究中，德尔菲法的精度低于传统的会议预测。在16项研究中的12项里，德尔菲法的预测比传统的专家意见调查而得出的预测更准确，其他两项精度相同，还有两项德尔菲法的精度低。在这24个比较中，依靠德尔菲法，71%提高了精度，12%降低了精度（Rowe and Wright, 2001）。
德尔菲法对管理人员很有吸引力，因为可以从分散专家那里得到判断而无需花钱组织会议。它比预测市场具有优势，因为参与者提供了预测的原因（Green, Armstrong and Graefe, 2007）。ForecastingPrinciples.com免费提供该程序的软件。

6. 模拟交互
模拟交互使用角色扮演来预测两个或多个利益冲突的各方的决策。已经用于测试该方法的情形包括尝试与主要供应商保持独家分销安排、工资管理方面的薪酬和条件争议以及艺术家要求政府提供财政支持。
预测者为每个角色扮演者提供一个主要角色的描述，并简要描述情况，包括一系列可能的决定。角色扮演者被要求彼此进行现实的交流，直到做出决定。模拟通常持续不到一个小时。

相对于最常见的独立的专家判断来说，模拟交互方式将八种冲突情况（包括上述情况和企图恶意收购一家公司，还有两国之间为了争取一个水域的通道而进行的军事对抗）的平均预测误差降低了57%（Green, 2005）。如果缺少经验的角色扮演者不了解对方，对情况没有事先的看法，也没有超出他们的角色所示的安排，那么这种方法似乎是最好的。

另一种替代方法是“换位思考”。美国国防部长罗伯特•麦克纳马拉认为，如果他在越战期间这样做了，他会做出更好的决定1。然而，对“角色思维”方法的测试发现，与独立判断相比，预测的准确性没有改善。在一个复杂的情况下考虑相互作用是非常困难的——为了提供足够的现实性，各方之间积极的角色扮演是必要的（Green and Armstrong, 2011）。

7. 构建类比

结构化的类比方法包括要求十多位专家提出类似于需要预测的情况和目标情况。专家给出目标情况的描述，并被要求识别类似的情况，评估他们与目标的相似程度，并将他们类比的结果与目标情况的可能结果相匹配。管理者根据每位专家的最高评分类比所含有的目标情况结果，并根据预测计算模态结果。不应该将这种方法与通常的类推法相混淆，并以此为预测者或客户所偏好的决定正名。

在上述模拟交互方法研究（Green and Armstrong, 2007a）的八个实际冲突预测决策中，结构类比预测比单独的判断预测精度高41%。结构类比也用于预测激励措施促进大学生购买笔记本电脑的效果，以及为高中学生父母提供互联网安全认证的计划。这些结构类比预测的误差比独立判断的预测误差（Nikolopoulos, Petropoulos, Bougioukos and Khammash, 2015）要低8%。使用类似于结构化类比的程序来预测19部未发行电影的票房收入，其中评估者从数据库中识别类似的电影并对它们进行相似性评估。类比的收入预测因广告支出和电影是否续集而进行了调整。从结构化类比预测的误差还不到简单和复杂回归模型预测的一半（Lovallo, Clarke and Camerer, 2012）。在上述三项研究的十个对比测试中，使用结构化类比的误差平均约为40%。

8. 实验

实验被广泛使用，是确定因果关系的最有效和可靠的方法。可以使用关于效应方向和效应强度估计的知识来进行预测。实验可以在实验室进行。对组织行为实验的分析发现，实验室实验与现场实验相似的结果（Locke, 1986）。

或者，预测者可以分析自然实验来确定因果关系并进行预测。例如，工业的管制和放松管制就对消费者福利的影响提供了天然的实验。温斯顿（1993）发现，在有实验性数据的九个市场中，有八个市场的监管受到了损害，而在第九市场中却没有任何净收益。

9. 专家系统

专家系统是通过询问专家来描述他们在做预测时采取的步骤，然后用软件来描述这个过程。由此产生的专家系统应该是完整的、简单的、描述清楚的。

对15个比较的回顾发现专家系统预测比独立判断的预测更精确（Collopy, Adya and Armstrong, 2001）。有关天然气和邮购目录销售的两项研究发现，专家系统的预测误差分别比独立判断的误差小10%和5%。尽管有关预测有效性的证据不足，但该方法似乎很有前景。
定量法

定量法需要预测问题的或与之相关的数据。定量法也可以借鉴一些判断类方法，如分解法，以便充分利用知识和数据。这些模型也明确使用因果关系。

这一部分描述了六种以证据为基础的定量预测方法。与第一种方法（外推法）不同，这些方法极大地依靠因果知识，以此来预测因变量中变化的效果。这些预测可以应用于决策制定，也可用于安排紧急计划。然而，在因变量不受决策者控制时，预测未来会发生什么要求因变量的预测十分准确。

10. 外推法

外推法适用于任何需要时间序列的预测问题，尤其适用于以下情况：对影响预测变量的因素了解较少时、因变量不会发生巨大改变时、或者因变量不能准确预测时。

指数平滑法容易理解，这种方法始于布朗（1959和1962）。这是一个明智的做法，因为它使用了移动平均数中的所有历史数据，这让最近的数据更加重要。关于指数平滑法的回顾，参见Gardner（2006）。

即使在短期内，人们也不应该认为一种趋势会以相同的速率继续发展。它会随着推动趋势发展的因果力量的变化而上升或降低。这种情况的不确定性越大，就越需要去抑制趋势趋向于0——即无变化预测。通过对10项实验的比较回顾，我们发现，一般情况下，抑制趋势趋向于0会使预测误差减少约5%，与当前趋势相比，会降低较大误差出现的风险（Armstrong, 2006）。Gardner的阻碍趋势外推软件可以在ForcastingPrinciples.com网站上找到。如果出现长期趋势，因果因素还会继续发展——如资源的实际价格（Simon, 1996）——这时，阻碍长期趋势是必要的。

若外推的时间不足一年，应当估测季节性影响的效应，并将其从数据中移除。预测随季节调整的序列，然后“按季节”预测。在预测市场竞争模式中的68个月经济序列中，有18个月范围内的季节调整使误差减少了23%（Makridakis, Andersen, Carbone, et al. 1984，表14）。

预测者应抑制季节性影响下的数据估计。这种预测带有不确定性，而标准季节性调整程序倾向于“过度迎合”数据。Miller和Williams (2003, 2004) 提供了一些阻碍季节性因素的程序。当他们从三类市场竞争模型中抑制了1428个月的时间序列的季节性调整时，时间序列的预测准确率将会提升59%到65%，这取决于范围的大小。Boylan, Goodwin, Mohammadiopour和Syntetos（2015）研究出了更广泛的发现成果。Miller-Williams的程序和三类市场竞争模型数据的软件可以免费获取。

抑制类比序列中的季节性因素平均的方式也会提高预测的准确性。在某项研究中，把季节性因素从相关的产品中加以结合，例如吹雪机和雪铲，这样会使平均预测误差减少约20%（Bunn and Vassilopoulos, 1999）。在另一项研究中，与每个城区单独估计的季节性因素相比，把六个城区的影响犯罪率的每月季节性因素汇总，这样指数平滑的预测误差就会下降约7%（Gorr, Oligschlager, and Thompson, 2003，见图4）。

乘法分解可用于将因果知识外推预测的结合。例如，在预测时间序列数据时，经常会发现这样一个现象，即序列受到因果力的影响，这些因果力以增长、衰减、反对、回归、支持或未知等表现。在这种情况下，可以通过具有不同方向效应的因果力来分解时间序列，推断每个分量，然后再重新组合。在以下两个条件下这样做可能提高准确性：（1）可以使用领域知识来构造问题，使得两个或多个组成系列的因果力不同；（2）可以获得每个组成部分的相对准确的预测。例如，为了预测机动车辆的死亡率，一项研究预测了驾驶里程数，这一序列数据预计会增长；同时也预测每百万乘客里程的死亡率，这一序列的数据因道路更好、汽车更安全而下降。这两
个外推预测相乘得到死亡总数。在测试五个时间序列中，有两个条件明显符合，那么由因果分解引起的样本外预测误差就减少了3/2。对于部分满足条件的四个序列，由因果分解引起的误差减少了一半。当条件不适用时，预测准确性没有增加或降低（Armstrong, Collopy and Yokum, 2005）。

合成分解也可以考虑外推问题。当最近的数据不确定或者需要随时更改时，分别预测起始水平和趋势，然后再相加的方法是十分有用的。这种方法叫“临近预报”。三个比较研究发现，通常情况下，“临近预报”可以减少37%的短期误差（Tessier and Armstrong, 2015）。

11. 以规则为基础的预测

以规则为基础的预测（RBF）运用以证据为基础的外推知识和因果知识来预测时间序列的数据。在使用RBF进行预测时，首先要从28个“特征”中鉴别出哪一个最能反应出要预测的序列的特点。这些特征包括预测范围、可使用的数据数量、异常值的存在等。然后利用99条RBF规则衡量可供选择的外推模式，然后把这些模式预测结合起来（Armstrong, Adya and Collopy, 2001）。

对于市场竞争类型下的90个年度序列的事前一年预测（可在ForecastingPrinciples.com网站上查询），RBF预测的绝对百分比误差比同等条件下的加权综合预测值小13%。对于六年的事前预测，RBF的预测误差小42%，这主要是因为长期范围的因果效应重要性提高了。在涉及强趋势、低不确定性、稳定性和优势专长方面，RBF的预测要比同等的加权预测组合更准确。在其他情况下，RBF对非加权的预测几乎没有优势（Collopy and Armstrong, 1992）。Vokurka, Flores and Pearce (1996) 的实验为RBF预测的相关预测提供证据。

99条RBF规则之一，“反向系列规则”是十分重要的，实施起来也相对简单便宜。这个规则指出，如果领域专家期待的时间序列与当前的时间序列相反，人们需要外推。单一使用该规矩要从五个数据集中外推时间序列。特别需要注意的是，对于长期的预测（6年的事前预测），误差减少率超40%（Armstrong and Collopy, 1993）。

12. 判断拔靴法

该方法在20世纪初研发，以预测美国即将成熟的玉米产量。在20世纪40年代，这种方法成功地运用于人员选择中（Meehl, 1954），同时也得到了后续研究的支持（例：Dawes and Corrigan, 1974; Grove, Zaid, Lebow, Snitz, and Nelson, 2000）。该方法利用回归分析法估测专家们进行判断预测时使用的变量系数。这一独立的变量不是结果，而是专家根据因变量值估计出的结果。近几年，参与预测的所有研究人员把这种方法称作“判断拔靴法”。实际上，在进行预测时，它使用的是专家们使用的因果信息模型，以此来提高专家预测的准确率。

在目前进行的比较研究中，拔靴模型预测的精度比以专家判断为基础的预测精度更高。准确性提高的原因在于这种定量模型与专家心理模型的应用具有更高的一致性。除此之外，这一模型不会受到无关信息和变量的干扰，同时也不会变得疲倦或烦躁。

建构拔靴模型的第一步是让专家根据自己的领域知识鉴别因变量。然后他们利用变量数据进行预测。例如，他们会要求预测博士生成功录取的可能性。

判断性拔靴模型可以根据专家依靠因变量做出的假设数据的预测来估计。这样做可以让预测者判断每个假设项是否合理，各假设项之间相互独立。该实验设计的使用克服了多元回归的许多缺陷。它也可以确保在实际数据不可用时进行预测。一旦发展起来，拔靴模型可以在较低成本的条件下对不同情况进行预测。例如，预测具有不同特征的新产品。
尽管人们发现了该方法有用的证据，但是这一方法在早期仅在农业预测中适用。社会科学家在20世纪60年代重新发现了这一方法，并对其预测效度进行了测试。这些研究发现，在11个比较试验中，有8个试验的判断性拔靴预测的结果比独立性判断的结果更加准确，有2个试验没有基本差异，一个试验的准确率略低（Armstrong, 2001a）专家依靠没有排除拔靴模型的无关变量进行的预测失败。与独立判断有关的特殊误差降低约6%。

很多大学将这种方法传授给学生，尽管最早进行的验证测试表明，该方法在预测博士项目成功与否方面更加准确、花费更低，但是我们意识到只有一个学校使用了该方法（Dawes, 1971）。

2002年，奥克兰田径棒球队使用了判断性拔靴法。专家们尝试着阻碍该方法的应用，这些专家通常运用自己的判断做出选择——选管理者、所有者和观察员。但是新的总经理坚持这样做，并且队员表现良好。其他职业运动队陆续采用了该方法，提高了比赛输赢率和效益（Armstrong, 2012a）。

13. 分割
预测中的分割涉及到问题的重建，以充分利用部分或子群体的知识和数据。而这这些问题表现不同。在预测每一部分时，要利用适当的方法，然后依据部分的预测得到对整体的预测。在20世纪60年代肯尼迪和尼克松的美国总统大选结果预测中，分割法吸引了很多人关注。

1995年，纽约港当局使用该方法预测未来十年的航空旅行需要。分析师把需要航空旅行的游客分成了130个商务游和160个个人游。个人游首先按照年龄划分，然后按照职业、收入和教育状况划分。商务游首先按照职业划分，然后按照从事的产业划分，最后按收入划分。每个分组的数据来源于人口普查和对旅游行为的调查。为了得到预测结果，官方把预测到的1965年航空旅行人口的数据分配各部分，以1935年无人旅游的数据为起点，外推出旅行人口和频率。对9000万次旅行的预测结果中，只有3%的数值与1965年的实际数值不同。

为了使用分割法，要鉴别出用于确定分割组成部分及其优先性的重要因变量。然后确定每个变量的切割点，如不同年龄的人。当存在非线性关系时，要使用更多的切割点，样品数据较少时，使用较少的切割点。然后，通过使用典型行为预测每个部分的人口及其行为。最后，把每个部分的人口及其行为的预测结合起来，并进行总和。该方法适用于有大量数据可用时。

分割法适用于以下几种情况：变量相互交织时、变量的影响力是非线性时、已知的因果知识较好时。这些情况发生于一个较为合理的研究中，即2717个加油站的数据用来作为预测汽油的周销量的分割模型。数据可以在九个二进制变量和其他十个变量中找到，这些变量包括区域类型、交通状况、是否有遮篷以及车站是否24小时开放。该方法使用了3000个车站样本进行测试。分割法模型预测误差（平均绝对百分比误差）比同样数据和变量下使用多元回归法造成的误差低约29%（Armstrong and Andress, 1970）。

有关分割法的文献综述由Armstrong提供（1985，第九章）。虽然预测准确性没有实质证据，但是该方法是合理的。因为这一方法以分解法为基础。在20实际70年代后，人们对分割法的兴趣降低，但是鉴于现在可以获得大数据，我们希望这种方法在现在要比过去可用性更高。

14. 一元回归
一元回归分析可以用来预测单一因变量的变化影响。这一方法相对保守，因为它通过计算一个常数项，减少效应大小估计值，来呼应估计数据中发现的关系的
变化。由于使用一元回归预测模型的估计有效用，人们必须控制或者准确预测因变量。

一元回归的传统形式是 $y = a + bx$，其中“$y$”是要预测的量（独立变量），“$a$”是常数，“$b$”是效应大小，“$x$”是因变量。这一方法适用于预测有较强的因果关系的较好的知识，同时还有独立的因变量中的有效和可依靠的变量。基本的假设是预测者必须有能力准确控制或预测因变量。

转换数据使得一元回归模型为因果关系提供了一个真实的代表。例如，在预测模式之前计算出因变量和独立变量的对数会以弹性的形式估计出效应大小。弹性指的是预测变量的百分比变化，它来源于因变量中一个百分比的变化。例如，牛肉价格的需求弹性为—1.2，这意味着价格上涨10%，此时所需要的牛肉量减少12%，这样才能达到平衡。其他需要考虑的转换内容包括以人均为单位表示的变量，和根据通货膨胀和季节变化对货币效应数据的调整。

窗口顶端
估计回归模型系数的最小二乘法会给极端数据值带来过度影响。为避免这种情况发生，应从估算数据中调整或删除异常值。其中一种方法就是所谓的“缩尾调整”（Tukey, 1962）。在做任何分析之前，预测者应该规定确定异常值的规则，以避免诱惑去调整支持优选的假设。另一个明智的方法是通过最小化绝对误差来估计回归模型（例如Dielman, 1986; Dielman, 1989）。

14.1. 多元回归
如果多个因变量均非常重要，该采取何种分析方式？多元回归分析（MRA）也许是一个最可取的解决方案，但是MRA分析方法采用的是非实验数据，这会导致多元共线性，因变量之间也会产生交互作用。此外，数据变量通常受测量误差和有效性的影响，而难以评定每个变量的相对权重。正是因为MRA的这种复杂性，在评定因果关系时，相比于一元回归来说，MRA处于相当大的劣势，即MRA不适用于“奥卡姆剃刀定律”。

据我们所知，MRA并没有经过预测有效性测试即被用于预测。我们所了解的第一个对比试验是1972年到2008年十次总统选举的选民票数。其精度被拿来与利用“最佳”变量（通常是“经济”）的一元回归模型相比较。我们对1972年到2008年十次总统选举（共1000项预测）最后100天的数据进行了预测，结果表明，MRA预测模型的平均绝对百分比误差为3.8%，而一元回归模型的误差为3.6%，与其相比，MRA的精度欠佳（Graefe and Armstrong, 2012）。
数据模型

从1960年代开始，技术的进步使得分析师能够利用统计显著性检验这一方式来选择多个“预测变量”并对其进行估计。我们将生成的模型称为“数据模型”。这种趋势始于20世纪中期，采用逐步回归的分析方法。它催生了如“大数据”、“解析”、“数据挖掘”、“神经网络”等程序。其中一个说法是，让数据说话能够增强其客观性。然而，如下文所述，实践中，这些技术往往产生相反的效用。

Einhorn（第367页，1972）是最早对数据模型提出质疑的人物之一。他总结道：“功能强大的电脑的应用使我们能够经常使用高度复杂的分析技术进行预测，但是这种方式通常没有任何理论、假设，或模型来指导研究者对于结果的预期”。他认为这一实践如炼金术一般不切实际。为进一步讨论回归分析在实践中应用的缺陷，请参见Armstrong（2012b）。

确定复杂情况下变量之间关系的唯一科学方法，是通过实验方式来确定所列举的因变量在不同条件下的效用。数据模型仅依靠数据进行分析而忽略了人们所积累的科学知识。

尽管大家普遍认为具有相关性并不意味着具有因果关系，数据模型却是基于统计上显著的相关性来建立的。20世纪80年代发表在《美国经济评论》的182篇有关回归分析的论文当中，大约有32%是依赖统计显著性来选择预测变量的（Ziliak and McCloskey, 2004）。20世纪90年代，情况更加糟糕，137篇文章当中有74%是依赖此方法选择预测变量。

统计显著性测试不利于科学进步（Armstrong, 2007a, 2007b）。一篇名为《缘何大多数出版的研究结果是谬误》的理论分析则展示了统计显著性测试以及首选假设测试是如何得出错误的研究结果，及此研究做得被予以出版的（Ioannidis, 2005）。利用数据模型，通过模糊的实验可以得出任何预期结论。如在分析完数据之后提出假设，不断尝试使用不同变量以找出能够支持首选假设的变量，舍弃不支持所需假设的观察结果，选择不合理的虚假设，使用大量样本用以确保统计的显著性，以及忽视其他研究人员所得出的不支持所需假设的结果等。这些程序是宣传研究中常采用的策略。Armstrong和Green（2018）汇总了例证，阐述了科学期刊会在何种程度上采用这种值得商榷的程序。

我们的调查无法找到任何能够表明MRA或其他数据建模技术与表1所确定的实证方法具有同样的样本外预测效度的实验对比结果。相反，我们所发现的证据表明，数据模型不适合被用来预测。

针对数据挖掘模型精确度的一个综合分析表明，数据挖掘模型得出所预测的样本外预测效度一直低于简单替代模型。在其中一个测试中，该研究的作者请求一个数据挖掘专家使用一组数据进行预测，专家照做了并且从数据中获得了许多具有统计显著性关系的数据挖掘者不知道，这些数字是随机的（Keogh and Kasetty, 2003）。在与我们的私人信件当中，Keogh说道：“虽然我读过每一篇关于时间序列数据挖掘的论文，我还从来没有见过一篇能让我相信数据挖掘预测法能够比随机猜测更有效。也许会有这样一篇论文，但我对此表示怀疑。”

知识模型

要对一些问题进行预测，我们需要了解许多重要的因变量。例如，预测哪位运动员将会在运动方面大展身手、谁将是一位行事高效的公司高管、哪些国家的经济增速最快、或者哪些移民申请者最有可能带来安全风险。知识模型对于分析这类问题最为合适。

本杰明·富兰克林在一封写给他的朋友约瑟夫·普里斯特利的信中提出过一种知识模型，在此之前普里斯特利曾写信给富兰克林，提到他正试图做出一个“棘手的决
富兰克林的方法是，列出每种选择的利弊，确定其主观权重，然后将列表相加来确定对其有利的并且得分最高的变量。富兰克林将他的方法称为“审慎代数”。

有一种叫做“经验表”的类似方法在20世纪初被用于决定哪些囚犯可以被予以假释（Burgess, 1936）。另一个方法被称为“结构分析”，从20世纪中期开始被投入使用。这种方法具有有效性（如，见Babst, Gottfredson and Ballard, 1968）。还有一个方法是最近基于“指数法”而开发的预测方法，对此我们在下文有相当多有关这种方法的实验。

因为“知识模型”这一术语比之前的术语更具描述性，因此我们将其命名为“知识模型”。表2列出了开发知识模型所需的清单。

a. 从专业知识和实验结论中确定所有重要的因变量。可以利用先验知识、遵循科学的方法来确定因变量。利用知识模型，因变量可以与二进制一样简单，例如，“比对手高”就可以成为选举预测模型的一个因变量。因变量在某些情况下从逻辑关系角度很容易辨别，万一遇到不易辨别的情况，可以考虑咨询三五个领域专家。当无法确定所列举的因变量的有效性时，可以查阅实验结果——尤其是实验的元分析结果，进而确定是否有足够的证据来支持此变量的使用。以下这个例子描述了依靠实验证据的重要性：Armstrong所提出的56个疏导原则（2010）所产生的效应的方向的例证，有的取自于非实验数据，有的则取自实验数据。为每个原则执行不同的实验，得出的实验结果的效应方向均相同；但对于非实验数据来说，只有三分之二的原则所对应的结果的效应方向相同（Armstrong and Patnik, 2009）。

b. 舍弃不能予以控制或不能提供精确预测的因变量——如果一个因变量不能被预测或控制，将其纳入到模型当中只会有损预测的精确性。

c. 确定将被用于预测的因变量所产生的效应为正还是为负。有些变量的方向效应可以从逻辑或者相关领域的常识中清楚地看出来。如果方向性不明显，那就参考实验证据。例如，人们对“枪支管理对犯罪的影响”这一问题持有不同的态度，选民和政客的反对意见导致美国各个州县改变其法律，或限制枪支的使用，抑或简化枪支持有权。这些自然科学实验为我们提供了一种科学地判断观点正误的方法，就像Lott, 2010年和2016年所做的那样。如果变量不具有显著性，也不能提供实验例证作为支持，那就舍弃这一变量。

d. 条件允许的情况下，确定因变量对将被用于预测的变量所产生的效应的相对值。应该考虑是否有可能的证据表明，一些因变量的变化相较于其它变量，对应变量的影响更大。可以参考实验证据，并考虑咨询相关领域的专家来确定不同的权重。在因变量所产生的效应差别明显的情况下，需要从统一中区分不同的权重。避免为了提高样本内契合度而改变先验变量权重。

e. 指定模型体现为应变量的分值，其等于因变量加权之和。知识模型简单地通过加总所确定的带权重的因变量值和变量值而得出相应的分值。得出的分值可以用于预测，分值越高，得出预期的结果的几率越高。

| a. 从专业知识和实验结论中确定所有重要的因变量 | □ |
| b. 舍弃不能予以控制或不能提供精确预测的因变量 | □ |
| c. 确定将被用于预测的因变量所产生的效应为正还是为负 | □ |
| d. 条件允许的情况下，确定因变量对将被用于预测的变量所产生的效应的相对值 | □ |
| e. 指定模型体现为应变量的分值，其等于因变量加权之和 | □ |
| f. 若可行，可采用回归分析方法估计分值和应变量数值之间的关系 | □ |

表2. 知识模型开发清单。
f. 若可行，可采用回归分析方法估计分值与应变量数值之间的关系。我们有足够的应变量的历史数据，因此可以估计知识模型的得分与使用简单回归分析的连续应变量之间的关系。在特定情况下，可以将回归估计参数（参数恒定且与分数共同作用）应用到知识模型得分之上，进而得出定量预测结果。

我们认为本杰明·富兰克林建议在预测时应考虑变量的不同权重这一想法没错，但只有在有强有力证据支持的情况下可以采用这种方法。例如，我们应该考虑的是，有多少专家了解这种因果关系，又有多少有关于这种关系的可用实验数据？在缺乏相关领域知识、数据不足以估算不同权重的情况下，则使用相同的权重。

第一个进行实证检验得出等权重有效性的人物是Schmidt（1971）。随后Einhorn和Hogarth（1975）及Dana和Dawes（2004）也证实，在某些情况下，等权重模型比回归权重所提供的预测结果更准确。


Graefe与Armstrong（2013）回顾了心理学、生物学、经济学、选举、健康、和人员选择有关的实证检验研究。结果表明，13项研究当中，有10项研究的知识模型所提供的预测结果比回归模型更加准确。

即使有充足的理由要求使用不同的权重，也要考虑调整权重，使其更趋平衡。我们在选举预测当中，使用8项独立的计量经济学选举预测模型来检验权重的平衡性，模型中使用的数据都具标准化，且与应变量正向相关。均衡系数为100%意味着使用同等权重，对于所有模型来说，当平衡性处于10%到60%之间时，预测的绝对误差都有所降低（Graefe, Armstrong and Green, 2014）。

一项研究通过实证检验，得出变量不同的权重，利用知识模型对96对广告的相对有效性进行预测，进而评估预测结果的准确性。在面对195个潜在的相关变量的时候，回归方式不再可行。在预测哪对广告更有效的时候，猜测法的正确率是50%。新手的判断预测法对54%对广告的预测结果是准确的，有广告行业经验的人预测结果的正确率是55%。广告文案测试（如，将广告展示给受试者，并要求他们评估购买此产品的可能性）预测的正确率为59%。相比之下，知识模型的预测有75%是准确的，比广告文案测试的错误率低37%（Armstrong, Du, Green and Graefe, 2016）。在扩展的研究当中，使用权重更均衡的变量对模型进行测试，得出的误差降低率为32%，与前述测试结果基本相同（Green, Armstrong, Du and Graefe, 2016）。

### 16.&17. 组合预测

表1所列的最后两个方法用于处理组合预测。这两种方法被认为是改善事前预测精确度的重要方法。

方法内和跨方法组合的基本规则是：(1) 从所有有效的循证方法的变量，如不同专家、数据、程序、和实施的成果中获取预测结果；(2) 在每一项组成方法内部，通过计算等权重的平均值获得变量，把各个变量的预测结果相结合；(3) 计算每一项组成方法的等权重平均值，将每一项组成方法的预测结果再次组合。如果存在有力证据表明预测精度存在差异，可以放松对等权重的规则的限制。在此情况下，在预测前就应该明确变量的权重。

对于重要的问题，我们的建议是，要想获得预测结果，应该采用三种不同的组成方法，每一项组成方法都应该至少包含两种变量，即组合预测结果之间应再次
结合，以提高预测结果的可靠性和有效性。有关组合预测结果的更多详情，可以参见Graefe, Armstrong, Jones和Cuzan (2014) 以及(Graefe, 2015)。

以上描述的组合预测程序能保证预测结果的准确性不会是最差的，至少与典型的成份预测效果等同。此外，当组成部分的范围包括真值时，组合预测的绝对误差要小于成份预测的误差平均值。通常情况下，组合预测会比最准确的成份预测结果还要精确。因为多重预测永远是可能的，应始终使用组合方式进行预测。因此，若从循证方法中可以获得两个或更多的预测结果，这时就应该使用组合预测方法。

组合预测方法并非直观。在与高素质的MBA学生进行的一系列的实验当中，大多数参与者认为平均估计只能取得中等的预测成果 (Larrick and Soll, 2006)。在另一项实验中，有偿招募了一组美国成年人，给出五个专家最近做出的关于电影放映观众人数的错误预测数据。当被问道要指定哪些专家的预测数据对未来电影放映的观影人数进行组合预测时，203个参与者中只有5%选择使用所有五位专家的预测。其余的人都选择了之前预测的误差最小的专家的预测结果进行组合（Mannes, Soll and Larrick, 2014）。

出于同样的直觉，当纽约市官员收到了关于2015年一月份即将到来的暴风雪的两份不同预测时，他们按照他们认为最好的预测采取行动。结果却发现，这份预测是最糟糕的。

在组合预测当中，还有很多研究需要去做。尤其是，我们需要更加了解（1）如何组合预测来达到最高的预测精度，（2）是否有些方法会比其他方法更能提高预测精度，如果有的话，是在什么样的条件下，以及（3）增加方法对于组合预测精度的边际效应，以及方法变化增加对于组合预测精度的边际效应。

单一方地变化而来的组合预测或是来自独立预测者使用同一方法得出的组合预测都能够帮助我们弥补各组成部分的过失，数据误差和样本容量小的问题。换言之，单一方法内的组合很可能是提高预测可靠性最有效的手段。然而单一方地预测不太可能比得上采用不同方法的预测，因为任何一种特定方法都会使预测产生向着一个方向偏移的误差。

一份报告进行了30次研究，比较来自同一方法的组合预测，结果发现组合预测的不加权算术平均误差为12.5%，小于典型预测3%到24%的平均误差（Armstrong, 2001c）。

另一项研究比较了八个独立的采用多元回归模型预测美国总统大选的预测精度及其中平均精度。通过该研究的15次选举发现组合预测的误差同典型独立模型预测相比减小了36%（Graefe, Armstrong and Green, 2014）。

不同的预测方法可能会有不同偏差，因为他们使用不同的假设，知识和数据。因此，采用多种方法进行的预测可能会比采用单一方法进行的预测更加接近实际结果。此外，因为包含了关于此情景的更多信息，采用多种方法进行的组合预测更有可能提高预测的可靠性。例如，一项研究检验了时序外推法和意图预测法组合预测的精度。研究发现两种方法组合预测与仅用外推法进行预测相比，误差减小了三分之一（Armstrong, Morwitz and Kumar, 2000）。

也要考虑到持有不同经济学理论的经济学家的预测组合在一起的这种情况。在一项研究中，提前12个月对GNP的实际增长进行预测，把持有相似经济学理论的经济学家的预测组合在一起，标准差平均能够减小11%，而把持有不同经济学理论的经济学家的预测组合在一起，误差能够减小23%。使用相似预测方法的经济学家的预测组合在一起，误差能够减小22%，而使用不同预测方法的经济学家的预测组合在一起误差能够减小21%（见Batchelor and Dua, 1995 表2）。组合多样化减小的误差远大于该研究六项比较中的其他五项，在其他五项比较中，经济学家采用相似/不同理论/方法预测国民生产总值平减指数，企业利润增长和失业率。
J. S. ARMSTRONG AND K. C. GREEN

The PollyVote.com网站进行的选举预测为测试采用四到六种不同方法对1992至2016年间的七次美国总统大选的普选结果进行组合预测的预测精度提供了数据。单一方法预测（例如选举民意调查）是第一个被纳入组合的。来自几种方法的组合预测随后被加人到组合中来。在选举进行前100天，该网站预测的平均绝对误差为1.1%，小于各个组成组合预测的平均误差，这些预测的平均误差在1.2%至2.6%之间，中位数为1.8%（Graefe, Armstrong, Jonesa and Cuzán, 2017）。

同典型单一方法组合预测相比，跨方法组合预测的能够减小40%的误差。前文提到过的一种方法内的组合预测误差减小了12.5%，两者放在一起，粗略估计方法内组合再到跨方法组合预测将会减小超过50%的误差。

预测原则：黄金原则和奥卡姆剃刀定律

现在我们把注意力从方法转移到原则上。附件1清单中所列出的预测方法都符合预测原则，因此根据预测方法应用清单可以帮助我们确保遵循了这些原则。然而，更重要的是，如果能够遵循这两条首要的预测原则：黄金法则和奥卡姆剃刀定律的话，那些坚持不使用附件1中给出的那些预测方法进行预测的预测者能够极大地提高他们的预测精度。

黄金法则和下文中的简易预测清单为如何遵循这两条原则提供了指引。他们不同于以前出版的原则清单——预测审计清单，可以在ForecastingPrinciples.com网站上找到——这家网站是为预测学者和从业人员打造的。例如，我们用预测审计清单来评价国际气候变化委员会在预测全球平均温度时所采用的的预测程序（Green and Armstrong, 2007b）。

相反，我们在这部分所给出的清单是想赋予所有相关方参与预测程序审计的权利。两条原则清单适用于各种类型的预测问题，适用于所有预测方法。

黄金原则

黄金原则就是要谨慎。更确切地说，谨慎就是遵循现有的关于情况和关于预测方法的知识积累（Armstrong, Green and Graefe, 2015）。

预测的黄金法则也是一种伦理原则，因为它意味着“己所不欲，勿施于预测。”在像是法律或是公共政策争端这样需要证实客观性的时候，这一原则是非常有用的参考（Green, Armstrong and Graefe, 2015）。

表3是Armstrong, Green, and Graefe图表1的修订版本（2015, p. 1718）。它包含28条由预测黄金法则逻辑推断出来的指南。同之前出版的版本相比，这一版本中有两处关键变化。第一处就是指南5中现在包括指令“组合使用不同方法的预测”，这一变化是基于前文中提供的证据。

第二个变化是指南6最初提醒谨慎使用判断性调整，但是现在变成了一种禁令：预测者应“避免调整预测”。之前的原因是不同方法的使用导致与情景相关的知识使用量上升，因此因遗漏关键信息而导致误差增大的可能性下降。此外，调整有可能会增加故意偏差。例如，在一项有关一家英国跨国公司内九个部门的调查发现，45名受访者中有64%的人都同意“预测通常会进行政治上的调整”（Fildes and Hastings, 1994）。在另一项研究中，29个以色列政治研究以民意调查发起者的独立性为依据从低到高进行了排列，分为“内部”——例如由某个政党发起的民意调查——“受委托的”，或是“自给的”。“独立民意调查的预测比由内部民意调查发起者进行的预测更加精确。例如，71%的独立民意调查精度更高，而60%的非独立民意调查精度相对较低（表4，Shamir, 1986）。
表 3. 预测的黄金法则清单：第二版。

<table>
<thead>
<tr>
<th>指南</th>
<th>( N )</th>
<th>超差减少</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 问题描述</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>1.1 选择适合预测情景的已验证假说</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>1.2 进行分解以更好地利用知识，信息和判断</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>1.3 避免使用不精确，重复试验和延伸，保证预测完全公开</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2. 判断法</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 避免独立判断</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2.2 使用可选择性评估和判断预测问题</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3 要求判断人写出对预测的原因</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2.4 使用推断性推定法</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>2.5 使用结构化模型</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2.6 组合不同判断人的独立预测</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>3. 外推法</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 使用最长时间序列的有关系数据</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2 使用结构进行分解</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.3 组合趋势变动性知识背景中如果…</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3.1 序列是稳定的或不稳定的</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>3.3.2 历史趋势与原动力冲突</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.3.3 预测范围比历史序列长</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3.4 微差趋势和长期趋势方向不一致</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4.1 跨年度预测存在很大的变化</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3.4.2 每年数据年份不多</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3.4.3 有关季节性的因素知识不足</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5 组合使用不同可替代外推法的预测</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4. 因果法</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1 使用先验知识推测变量，关系和效应</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4.2 调整效果预测来反映不确定性</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4.3 使用重要变量</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4.4 组合使用不同可替代因果法的预测</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5. 组合使用不同方法的预测</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. 避免调整预测</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* 表 3. 预测的黄金法则清单：第二版。

表 4. “简洁预测”清单: 奥卡姆剃刀定律。

我们要考虑到经常通过判断性调整统计预测来解决问题：预测一个产品的销售量受到周期性促销活动的影响（例，见Fildes and Goodwin, 2007）。可以通过把问题分解成子问题来避免调整。即分别预测层次，趋势和促销的效果。Trapero, Pedregal, Fildes, and Kourentzes (2013)给这种方法提供了支持，发现同调整过的预测相比这样预测的平均绝对误差平均减小20%左右。

我们已经不能找到任何证据证明调整会降低预测的误差相对于预测误差产生的方式与本论文中呈现出的指导相同。尤其是根据指南1.1.2——分解预测问题以求最好地利用知识，信息和判断——以及修订版指南5——组合使用不同方法的预测——帮助确保所有相关知识和信息都会被包含在预测当中，不给调整预测留下任何合理的理由。

我们在文献搜寻中发现关于证明这18条指南对于预测精度影响的证据。平均来看，违反一条典型指南会增加40%的误差，在附件3中详细列出。预计违反指导条款数增加会导致误差增加。尽管我们没有关于在实践中黄金法则在多大程度上得到了遵循的系统性信息，我们预计发表在科学杂志上的大多数预测研究都典型的违反了黄金法则中的大多数指南。例如，我们审计发现在联合国国际气候变化委员会就推测危险的人为全球变暖的时候就无视了黄金法则。

任何利益相关者都可以使用预测黄金法则清单。专业和非专业人士都能在不到一个小时的时间里完成这份预测黄金法则清单。利益相关者不需要是使用这份清单的专家，因为责任在于预测者，他们要完整清楚的解释他们的方法（指南1.3）。为了帮助提高这份清单评分的可靠性，利益相关者可以要求至少三人参加评分，每人独立完成，对预测过程进行评分之后算出平均分数。

简易预测

“简单性原则”（奥卡姆剃刀定律）是一种科学原则即用最简单的方法解释证据就是最好的。这一原则由亚里士多德提出，后来以14世纪学者奥卡姆的威廉的名字命名。这一原则也适用于科学预测：预测者应该使用最简单的方法来建立最简单的模型来与情景知识保持一致。

预测者们都遵循奥卡姆剃刀定律了吗？显然没有：1978年，21位世界顶尖经济计量预测专家被问到是否复杂的计量经济方法比简单方法的预测更加精确，72%的人回答确实是。在那个研究中，“复杂性”被视为一种反映被用来进行预测的方法数量
指数：（1）系数的使用，系数不是0或1；（2）变量的数量；（3）功能性关系；（4）方程式数量；以及（5）联立方程的使用（Armstrong, 1978）。

20世纪50年代开始，研究人员认识到了复杂的统计模型来对时序数据进行外推。他们导出数学模型，并且报告这一模型适应数据的能力。这一模型十分流行，被学者和从业人员广泛使用，但是他们的预测效度并没有同其他可替代方法进行测试比较。

20世纪70年代后期，研究人员被邀请使用他们的模型参加一个竞赛，来对111个未确定的商业和经济近六年的月度，季度和年度时序数据进行外推。这些采用不同方法进行的预测以不变的基准模型预测结果为标准进行预测精度测定。简单模型效果很好，而那些复杂模型得出的预测精度差别不大。竞赛结果连同14位顶尖统计学家的评论一起出版（Makridakis and Hibon, 1979）。Makridakis扩展了这一竞赛——被称为M-竞赛（Makridakis et al. 1993, Makridakis and Hibon, 2000）——得出这样的结论：使用简单方法进行的外推预测同使用复杂方法得出的预测一样很有竞争力（Gigerenzer, Todd, et al. 1999）。

研究不同类型问题——例如预测高中辍学率——的一系列测试结果显示简单的启发法和复杂预测方法精度相当，而且经常更高 (Gigerenzer, Todd, et al. 1999)。

我们推出一个新的关于预测简捷性的操作性定义，这样任何利益相关者都可以测定简捷性。它由一个四项判定预测简捷性的清单组成，以此来方便潜在的预测使用者。这一清单在尚未进行任何分析时就被制定出来了，而且它作为测试结果是不可改变的。表4是ForecastingPrinciples.com网站上这一清单的缩减版（Green and Armstrong, 2015）。

我们的研究一共鉴别了32份出版的允许进行简单方法预测和复杂方法预测的精度对比的论文。其中4份论文测试了判断法，17份测试了外推法，8份测试了因果法，3份测试了组合预测法。各个方法的结果都是一致的。平均来看，每组比较，同简单方法相比，方法变复杂，预测前误差就增大27%。结果是令人惊奇的，因为这些论文很显然都在推崇更加复杂的方法，他们期望方法越复杂，预测就越精确。他们所有，在复杂情景下，复杂方法从未比简单的预测法预测更精确。后来《计量经济学杂志》的创始者Arnold Zellner也得出了同样的结论。3

预测的不确定性评估

预测的不确定性影响其实用性。例如，如果下一年手机的需求量预计增加20%，手机生产商可能考虑雇佣更多员工并购买更多机器。然而，如果这一预测有很高的不确定性，也就意味着需求下降也极有可能，那么扩展业务可能就变得不谨慎。

本章首先描述估计预测区间的误差度量。当前，预测区间的估计通常太有限。我们建议将统计估计的95%置信区间的宽度加倍，从而尽可能地接近95%预测区间。但是，使用下面附录5中的指导方针会更好。

误差度量

早些时候，通过将它们预测的准确性和替代方法预测的准确性进行对比，我们讨论了适合于评估预测方法的误差度量。这里，为了估计对管理决策有用的预测区间，我们建议根据实际值预测平均绝对偏差（MAD），因为MAD容易计算并且决策者容易理解。另一方面，应该避免使用常用的均方根误差（RMSE）度量，因为它与效益无关。
当预测涉及到非对称误差的问题，也就是负误差大于正误差或正误差大于负误差的问题时，要计算预测值和实际值的对数，并通过记录值计算误差。通过这些误差估计预测区间，然后将区间的界限转回实际值（Armstrong and Collopy, 2001，第281页）。

损失函数也可以是不对称的。例如，预测太低损失的50个单位与预测过高而损失的50个单位并不相同。然而，非对称误差是规划者的问题而不是预测者的问题：因为规划者必须评估正误差与负误差所造成的损失。

预测不确定性的方法

表5展示了预测不确定性的方法清单。清单包括四个有效的使用方法，以及应避免使用的两个常用但无效的方法。

1. 使用经验预测区间或基于样本测试估计的似然

根据历史数据估算的传统统计置信区间通常太有限。一项研究表明，外推预测的95%置信区间之外的实际值的百分比通常大于50%（Makridakis, Hibon, Lusk and Belhadjali, 1987）。

为了向预测用户提供关于预测不确定性的有用信息，除了基于样本预测误差来估计经验预测区间，别无选择。为此，尽可能模拟实际的预测过程，并使用最终预测误差的分布来评估不确定性。Tashman（2000）提供了关于基于样本测试的指导。有关估计预测区间的更多信息，参见Chatfield（2001）。

在分析时间序列预测误差时，使用逐次更新来增加预测的数量。如果没有足够的有效数据，请考虑使用类似情况的数据。

2. 按源分解误差从而估计每一部分的不确定性

大多数预测问题都受到预测误差几个来源的影响。为了帮助确保所有的误差得到解决，需要考虑通过误差来源分解误差从而估计每个误差，然后合并估计值。例如，在投票预测政治选举的结果时，调查研究人员只会报道基于样本大小随机变化而出现的预期误差。响应和无响应偏移误差都被忽略掉了。因此，民意调查报告的95%置信区间大约是它们应该达到的一半（Buchanan, 1986）。换句话说，决策者应该使政治民意调查的置信区间加倍，以获得对预测区间的更真实的估计值。

当不确定性很高时，比如通过调查公民来预测他们响应政府规定变化的行为，由于调查受访者对他们如何做出决定缺乏自我认知，响应误差可能会特别高（参见Nisbett and Wilson, 1977）。非响应也可能是误差的一个很大来源，因为受调查主题影响大的人更可能做出响应。尽管后面的误差在一定程度上可以通过“波浪外推法”（Armstrong and Overton, 1977）来降低，但预测者在评估不确定性时仍然需要考虑误差的来源。

与调查回答的分析一样，回归模型的诊断统计忽略了不确定性的主要来源，例如关键变量的遗漏，控制或预测因果变量的难度，不能准确预测因果变量以及难以估计相互关联的因果变量的相对重要性。当分析师力求与历史数据密切配合时，这些问题就会被放大，当数据挖掘技术被用来实现密切配合时更是如此。

3. 通过结构化判断估计预测区间和似然

评估不确定性的另一个常见判断方法是让专家以95%的预测区间表达他们对自己的判断性预测的信心。这种方法的一个问题是，专家通常对他们预测的准确性过于自信。例如，分析来自22位经济学家超过11年的经济预测的判断置信区间发现，结果仅有57%的时间处于95%置信区间的范围内（McNees, 1992）。另一项研究追踪了
一个为期10年的小组成员，他们提供了公司预期股票市场回报的13,300个估计值，实际只有36%的时间处于80%置信区间范围内（Ben-David, et al. 2013）。一些结构化的方法可以改进判断性预测的校准。确保判断是从许多专家那里获得的，并获得独立的匿名评估。德尔菲法可以用于这个目的。让专家列出所有不确定性来源，并列出他们的预测可能错误的全部原因。Arkes（2001）证明这种方法是有效的。

最后，为了改进未来预测人员对不确定性估计的校准，确保他们得到及时、准确、频繁以及完整总结的实质性信息和他们预测正误的原因。例如，天气预报员使用这样的程序，他们的预测在几天之前就会得到很好的校准：当他们说有40%的机会下雨时，通常，在40%的时间里会下雨（Murphy and Winkler, 1984）。

4. 结合不确定性估计的多种有效值
组合不确定性估计的逻辑与组合预测的逻辑相同。因此，基于组合估计的不确定性预测总会比典型预测要好，只要有不确定性估计出现，组合估计总是优于典型估计。

5. 避免使用有历史数据的统计拟合估计不确定性
在一个对30个变量运用31个观察值的研究中，逐步回归使用的规则是只有t统计值大于2.0的变量才会被包括在模型中。数据来自于一本书中的随机数字。尽管如此，逐步的方法提供了一个具有良好统计拟合的八变量回归模型--根据自由度调整R²为0.85（Armstrong, 1970）。

统计拟合度量不提供有关样本预测有效性的有用信息（Armstrong, 2001c）。实验测试分析师通过对回归模型的标准统计拟合信息进行分析，发现72%严重低估预测的不确定性与模型的因果（政策）变量相关（Soyer and Hogarth, 2012）。Armstrong（2001d）进一步讨论为什么预测人员应该避免类似于调整R²的措施。

6. 避免使用统计学中的显著性测试估计不确定性
统计学中的显著性测试不会提供预测不确定性的估计，所以试图以那种方式使用它们可能会导致混淆并做出糟糕的决策。超半个世纪的实验研究支持这一结论（例如参见Ziliak and McCloskey, 2008; Hubbard, 2016; Armstrong and Green, 2018）。

有一项实验向重要的研究人员展示了两种药物的治疗差异，以及“p值”的差异，并询问他们会推荐哪种药物给潜在的患者。当治疗效果的差异很大，但p> 0.05时，近一半回答说他们认为两种药物没有区别。然而，当治疗效果差异很小但有统计学意义时（p <0.05），87%的受访者回答说他们建议服用这种药物（McShane and Gal, 2015）。McShane和Gal（2017）的另一个实验中的许多教学统计数据也未能得出逻辑结论。

由于提供有统计学意义的信息而对结果有了错误的解读，从而导致糟糕的决定。Hauer (2004) 讲述了与汽车交通安全有关的决定所造成的危害，如“红灯右转”政策。Ziliak和McCloskey（2008）提供了其他的例子。

据我们所知，没有科学研究表明，统计显著性测试能带来好的预测、决策或科学贡献。Schmidt（1996）提出了这样的挑战：“你能列出一项显著性测试对研究型企业做出（或会做出）的合法贡献吗（也就是说，它对累积科学知识的发展做出的任何贡献）？” Schmidt和Hunter（1997）指出，没有出现过这种情况，他们重复了这个挑战，就像我们在这里所做的一样。
讨论

一个世纪以来，关于预测方面所积累的科学知识使预测的准确性得以提高。遗憾的是，那些知识经常被忽视，而预测似乎正在变得不准。有两个相关的原因：第一，倡导性研究倾向于取代客观性预测；第二，不支持数据模型已经导致预测者忽视对于因果关系和对有效预测方法的知识积累。（Armstrong and Green, 2018）

当客户和公民根据不准确的预测结果做出决策时，他们会有什么样的追索权？答案从来都没有，因为我们不可能区分由于随机或是不可预测的变化而导致的预测失误和由于预测者未能遵循基于证据的程序而产生的预测失误。本文以医学、工程和航空为基础，提供了可用于保存预测人员（包括做出预测的科学家，以及公共政策制定者）的清单（如果他们未能遵循基于证据的预测程序）。

预测者可以使用清单来提高预测准确性，通过通信，他们已经遵循了清单，保护自己免受索赔。遵循清单的预测者可能会像医疗从业者一样，通过安排保险，来防止损害索赔，因为他们会遵循清单所要求的预测程序。

科学要求预测的有效性和新的发现是经过检验的，米尔顿•弗里德曼（1953）认为，出于样本的竞争假设预测有效性检验是经济学作为社会科学的一个基本要素。本文中的清单可以帮助科学家设计这样的测试。

结论

可以改进预测实践，从而大大提高企业和公共政策决策者对预测的准确性。实现这一目标的最佳途径是要求预测者遵循基于证据的清单。

附录1，预测方法应用清单，列出了15种基于证据的预测方法。相对于常用的预测方法（包括专家的非辅助判断），使用这些方法大大提高了预测的准确性。误差降低了从大约5%（对于阻尼趋势外推和季节性分解来说）到50%多（对于模拟交互和知识模型来说）不等。

预测者应该采用不同的基于证据的方法，并结合他们的预测，而不是希望找出最好的模型。这样做减少了使用单一方法产生的偏差，通过结合更多的知识和信息提高了可靠性。这种结合也避免了做出最坏预测的风险，并保证结合预测的预测结果比平常的预测结果更准确，因为结合预测包括了所有的预测结果。

数据模型不适合预测，特别是，多元回归方法违反了基于证据的预测原则，它提供的预测远远不及附录1清单中列出的方法准确。数据模型可以并且正在被用来支持客户和资助者对于进一步损害预测准确性的先验信念和偏好。

黄金法则和简单的预测原则清单（附录3和4）可以帮助预测者实施附录1中列出的基于证据的方法，在用基于证据的方法替代它们行不通的情况下，能够帮助预测者改进目前所使用的方法。遵循黄金法则的预测可以帮助预测者减少一半以上的预测误差，而忽略奥卡姆剃刀定律可能会使误差增加27%左右。原则清单还可以帮助客户、赞助商和用户评估是否遵循了适当的程序。

目前用于评估预测不确定性的程序误导了分析师、客户和用户，使他们变得过于自信。作为一个经验法则，它们比较准确。预测区间只能通过使用出于样本的测试来估计，如附录5清单中所描述的那样。

对准确预测感兴趣的客户和其他资助者应该要求预测者遵循本文提供的5个基于证据的清单，并对预测者的程序进行审核，以确保他们做到了这一点。客户和其他预测利益相关者可以使用清单，通过确定它们是否是科学预测程序的产物来评估预测的价值。
Notes

1. From the 2003 documentary film, “Fog of War”.

Acknowledgments

We thank our reviewers, Hal Arkes, Kay A. Armstrong, Roy Batchelor, David Corkindale, Alfred G. Cuzán, John Dawes, Robert Fildes, Paul Goodwin, Andreas Graefe, Rob Hyndman, Randall Jones, Magne Jorgensen, Spyros Makridakis, Kostas Nikolopoulos, Keith Ord, Don Peters, and Malcolm Wright. Thanks also to those who made useful suggestions: Raymond Hubbard, Frank Schmidt, Phil Stern, and Firoozeh Zarkesh. And to our editors: Harrison Beard, Amy Dai, Simone Liao, Brian Moore, Maya Mudambi, Esther Park, Scheherbano Rafay, and Lynn Selhat. Finally, we thank the authors of the papers that we cited for their substantive findings for their prompt confirmation and useful suggestions on how to best summarize their work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We received no funding for this paper and have no commercial interests in any method. We endeavored to conform with the Criteria for Science Checklist at GuidelinesforScience.com.

ORCID

Kesten C. Green http://orcid.org/0000-0001-9122-885X

References


