
Decision Analysis

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# An Overview of Applications of Proper Scoring Rules

Arthur Carvalho

To cite this article:
Arthur Carvalho (2016) An Overview of Applications of Proper Scoring Rules. Decision Analysis 13(4):223-242. https://doi.org/10.1287/deca.2016.0337

DECISION ANALYSIS
Vol. 13, No. 4, December 2016, pp. 223–242
ISSN 1545-8490 (print) | ISSN 1545-8504 (online)

https://doi.org/10.1287/deca.2016.0337


# An Overview of Applications of Proper Scoring Rules

Arthur Carvalho
Farmer School of Business, Miami University, Oxford, Ohio 45056, arthur.carvalho@miamioh.edu

We present a study on the evolution of publications about applications of proper scoring rules. Specifically, we consider articles reporting the use of proper scoring rules when either measuring the accuracy of forecasts or for inducing honest reporting of private information within a certain context. Our analysis of a data set containing 201 articles published between 1950 and 2015 suggests that there has been a tremendous increase in the number of published articles about proper scoring rules over the years. Moreover, the weather/climate, prediction markets, psychology, and energy domains are the four most popular application areas. After providing some insights on how proper scoring rules are applied in different domains, we analyze the publication outlets where the articles in our data set were published. In this regard, we find that an increasing number of articles are now being published in conference proceedings related to artificial intelligence, as opposed to traditional academic journals. We conclude this review by suggesting that the *wisdom-of-crowds* phenomenon might be a driving force behind the recent popularity of proper scoring rules.

*Keywords*: proper scoring rules; forecasting; forecast evaluation; incentive engineering
*History*: Received on September 10, 2015. Accepted by Editor-in-Chief Rakesh K. Sarin on August 27, 2016, after 2 revisions. Published online in *Articles in Advance* November 11, 2016.

## 1. Introduction

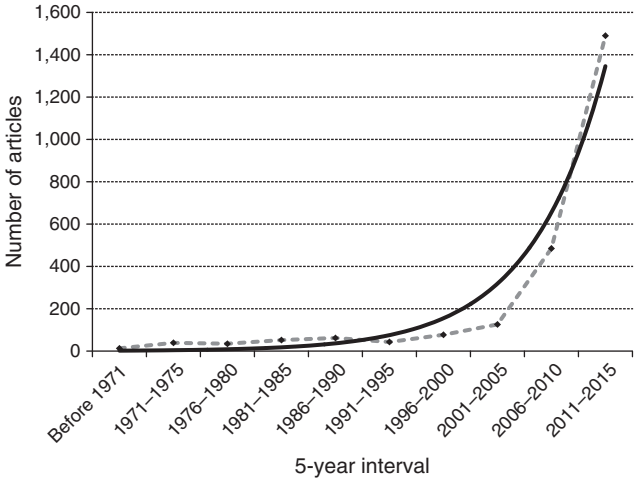
Forecasting is a crucial and ubiquitous activity in current decision-making processes. For example (1) companies rely on predictions about material supply and consumer demand to make their production plans; (2) weather forecasts provide guidelines for long-range and/or seasonal agricultural planning, in a sense that farmers can select crops that are best suited to the anticipated climatic conditions; (3) in supply chain management, forecasts help to ensure that the right product is at the right place at the right time, thus helping retailers to reduce excess inventory and to increase profit margin; and (4) forecasts of economic indicators, such as gross domestic product and unemployment rate, help policy makers to shape economic policies.

Roughly speaking, there are two approaches to forecasting. The *model-based approach* relies on statistical models to forecast the outcomes of future events based upon historical data. Consequently, the model-based approach is appropriate when historical data are available and contain valuable information about the future event of interest. Clearly, it is not always the case that historical data are available/meaningful when forecasting future events. In such cases, an alternative is to perform forecasting using the *expert-driven approach*, where a forecast reflects an expert's belief regarding the occurrence of the future event. Instead of model-based and expert-driven approaches, some authors alternatively call quantitative and qualitative approaches, e.g., see the work by Mahmoud (1984).

The literature related to the elicitation of experts' forecasts and, more broadly, experts' opinions, is typically concerned with (1) how to use the elicited information, (2) how uncertainty is or should be represented, (3) how experts do or should reason with uncertainty, (4) how to evaluate the quality and usefulness of the reported information, and (5) how to induce desirable behavior, such as honest reporting (Cooke 1991). It is the last two questions that we focus on in this paper.

Forecast quality and honest reporting relate to what Winkler and Murphy (1969) described in their seminal paper as the *substantive* and *normative* standards of "goodness." The substantive standard of goodness concerns the degree of perfection of a forecast, i.e., the degree of association between the forecast and the relevant observation. The process of measuring such an association is often referred to as *forecast evaluation*. The normative standard of goodness, on the other hand, relates to whether the forecast an expert reports

**Figure 1** **Number of Articles Containing the Term "Proper Scoring Rules" According to Google Scholar Over Five-Year Intervals**




*Note.* The solid line represents the function $y = e^{0.7205x}$, which is the result of fitting an exponential curve to the data points for the sake of better visualization.

corresponds to the expert's belief regarding the occurrences of the outcomes of interest.

*Scoring rules* are traditional techniques to measure the association between a forecast and an observed outcome. The condition that a scoring rule is *proper* means that an expert maximizes his expected score when he reports a forecast that is equal to his belief. As a consequence, proper scoring rules deal with both the substantive and the normative standards of goodness. Because of such attractive properties, proper scoring rules have received considerable attention over the years, as Figure 1 shows.

In this paper, we provide an overview of the past and current applications of proper scoring rules. The term "application" does not necessarily connote empirical or experimental work. Instead, it means that proper scoring rules are applied either in a normative or in a substantive way, i.e., when either encouraging experts to report their beliefs in a certain domain or for the sake of measuring forecast accuracy. In particular, we collected and analyzed a total of 201 relevant academic articles.

Our first finding is that there has been a tremendous increase in the number of published articles about proper scoring rules over the years. We further identify four major areas of applications of proper scoring rules: the weather/climate, prediction markets, psychology, and energy domains. We also note a shift in the publication patterns in that more and more articles are now being published in conference proceedings, as opposed to academic journals. A closer inspection of this result shows that most of these conferences are related to the artificial intelligence (AI) field.

In addition to connecting similar research conducted by different research communities, we believe our results might be of great value to both practitioners and theoreticians. Practitioners, on the one hand, can learn more about how to apply proper scoring rules in different domains, whereas theoreticians can better guide their work by having different domains in mind. We hope to provide a starting point for anyone interested in proper scoring rules and to motivate further research interest in this area.

The rest of this paper is organized as follows. In Section 2, we review the basic theory behind proper scoring rules. In Section 3, we describe the data collection process, i.e., the process of collecting the articles we use in our analysis. We present the results of our descriptive analysis and derive our conclusions regarding applications of proper scoring rules in Section 4. Finally, we conclude our work in Section 5, where we explain why the wisdom-of-crowds phenomenon might be the driving force behind the recent popularity of proper scoring rules.

## 2. Proper Scoring Rules

In this section, we provide a short summary of the theory behind proper scoring rules. We refer the interested readers to the work by Gneiting and Raftery (2007) for an in-depth coverage of the concepts explained next. Consider an expert who must report a probability assessment concerning an event that consists of a set of exhaustive and mutually exclusive outcomes $\theta_1, \theta_2, \ldots, \theta_n$, for $n \geq 2$. Each expert has a *belief* regarding the occurrence of the outcomes. We denote an expert's belief by the probability vector $\mathbf{p} = (p_1, \ldots, p_n)$, where $p_k$ is the expert's subjective probability regarding the occurrence of outcome $\theta_k$, for $k \in \{1, \ldots, n\}$.

It is often the case that experts are self-interested and, consequently, they are not necessarily honest when reporting their beliefs. For example, experts with a reputation to protect might report forecasts near the most likely group consensus, whereas experts who have a reputation to build might overstate the probabilities of outcomes they believe will be understated in a possible

consensus (Nakazono 2013). Therefore, we distinguish between an expert's belief $\mathbf{p}$, and the expert's *reported forecast* $\mathbf{q} = (q_1, \ldots, q_n)$, henceforth referred to only as *forecast*.

From a decision-making perspective, it is desirable that an expert's forecast equals the expert's belief, thus meeting the normative standard of goodness. Furthermore, the expert's forecast must be accurate, in that it corresponds as much as possible to future observations, thus meeting the substantive standard of goodness. As Winkler and Murphy (1969, p. 753) eloquently wrote, "the former [condition] requires probabilities to correspond to judgments, while the latter [condition] requires the probabilities to correspond to something in reality."

*Proper scoring rules* are traditional techniques that deal with both conditions; i.e., they promote honest reporting of beliefs by risk-neutral experts, and they measure how accurate forecasts are. Formally, a *scoring rule* $R(\mathbf{q}, \theta_x)$ is a function that provides a *score* for the forecast $\mathbf{q}$ upon observing outcome $\theta_x$, for $x \in \{1, \ldots, n\}$. Scores are often coupled with relevant incentives, be they social/psychological, such as praise or visibility, or material rewards through prizes or money. A scoring rule is called *proper* when an expert maximizes his expected score by reporting a forecast $\mathbf{q}$ that corresponds to his belief $\mathbf{p}$. Hence, it is in the best interest of a risk-neutral expert to behave honestly. An expert's expected score for a real-valued scoring rule $R(\mathbf{q}, \theta_x)$ is

$$\mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \cdot)] = \sum_{k=1}^{n} p_k R(\mathbf{q}, \theta_k). \tag{1}$$

An interesting property of proper scoring rules is that any positive affine transformation of a proper scoring rule is still proper; i.e., the scoring rule $\gamma R(\mathbf{q}, \theta_x) + \lambda$, for $\gamma > 0$ and $\lambda \in \mathbb{R}$, is also proper (Toda 1963). The best-known proper scoring rules, together with their scoring ranges, are

$$\text{logarithmic: } R(\mathbf{q}, \theta_x) = \log q_x \quad (-\infty, 0];$$

$$\text{quadratic: } R(\mathbf{q}, \theta_x) = 2q_x - \sum_{k=1}^{n} q_k^2 \quad [-1, 1];$$

$$\text{spherical: } R(\mathbf{q}, \theta_x) = \frac{q_x}{\sqrt{\sum_{k=1}^{n} q_k^2}} \quad [0, 1].$$

Selten (1998) and Jose (2009) provided axiomatic characterizations of, respectively, the quadratic scoring rule and the spherical scoring rule in terms of desirable properties, e.g., sensitivity to small probability values, symmetry, invariance, etc. One interesting result concerns that the logarithmic scoring rule is the only proper scoring rule (up to a positive affine transformation) that is *local*, i.e., that depends only on the reported probability $q_x$ associated with the observed outcome $\theta_x$ (Winkler 1969). In a seminal work, Savage (1971) showed that any differentiable strictly convex function $J(\mathbf{q})$ that is well behaved at the end points of the scoring range can be used to generate a proper scoring rule. Formally,

$$R(\mathbf{q}, \theta_x) = J(\mathbf{q}) - \left( \sum_{k=1}^{n} \frac{\partial J(\mathbf{q})}{\partial q_k} \times q_k \right) + \frac{\partial J(\mathbf{q})}{\partial q_x}.$$

One can derive the logarithmic scoring rule by using $J(\mathbf{q}) = \sum_{k=1}^{n} q_k \log q_k$, whereas $J(\mathbf{q}) = \sum_{k=1}^{n} q_k^2$ yields the quadratic scoring rule. Gneiting and Raftery (2007) and Schervish (1989) provided more rigorous versions of the characterization by Savage (1971). Proper scoring rules have also been characterized in terms of entropy measures (Jose et al. 2008), decision geometry (Dawid 2007), and by means of convex analysis (Hendrickson and Buehler 1971).

Even though our discussion in this section has been focused on the elicitation of forecasts as discrete probability distributions, it is important to mention that proper scoring rules have also been proposed to elicit continuous probability distribution (Matheson and Winkler 1976), quantiles (Jose and Winkler 2009), and intervals (Winkler 1972). Proper scoring rules can also be adapted to take into account different baseline forecasts (Jose et al. 2009, Forbes 2012).

In terms of implementation, it might be desirable to present a proper scoring rule to less mathematically inclined experts without explicitly introducing the mathematical formulae. For example, Johnstone et al. (2011) suggested using tables and graphs displaying different scores for different assessments. Andersen et al. (2014) suggested using sliders for the experts to choose the desired probability associated with each outcome. When moving the sliders, the underlying software can simultaneously display the payoffs associated with each outcome. Among other benefits, the above approaches might ensure that axioms of probability theory, such as additivity, are not violated.

### 2.1. Choosing a Proper Scoring Rule

Since different proper scoring rules might satisfy different properties other than properness, a question

that then arises is, which proper scoring rule should one choose? In this regard, one particularly relevant question is on whether different proper scoring rules result in different rankings of experts. Bickel (2007) argued that the quadratic and spherical scoring rules often result in extreme ranking differences when compared to the logarithmic scoring rule. Furthermore, Bickel (2007) concluded that, due to being nonlocal, the quadratic and spherical scoring rules allow for the possibility of one expert receiving the highest score when assigning a probability to the observed outcome lower than the probabilities assigned by other experts.

The above arguments by Bickel (2007), together with the fact that the logarithmic scoring rule is the only proper scoring rule that is local, might suggest that the same is superior to the other proper scoring rules, at least when the ranking of experts is an important criterion. Selten (1998), on the other hand, criticized the logarithmic scoring rule by showing that its resulting score is very sensitive to small mistakes for small probabilities. Another drawback of the logarithmic scoring rule is that an expert's score is $-\infty$ when an event occurs that the expert predicted to be impossible. Thus, the logarithmic scoring rule is unbounded and it needs to be truncated in practice, but it is not difficult to show that it will no longer be strictly proper after such a truncation.

The above discussion illustrates that the choice of the most appropriate proper scoring rule is dependent on the desired properties, which in turn is dependent on the underlying context. Merkle and Steyvers (2013) reached the same conclusion when investigating how different proper scoring rules from the beta family influence the rankings of experts. Quoting the authors: "*it is insufficient to use a scoring rule simply because it is strictly proper; instead, it is beneficial to consider the specific way in which the scoring rule rewards and penalizes forecasts*" (Merkle and Steyvers 2013, p. 302; emphasis added). Machete (2013) also suggested that the proper scoring rule one chooses should depend on the application at hand, and an issue to consider may be future decisions associated with high impact, low probability events. In particular, given two forecasts whose errors from the ideal distribution differ only by the sign, Machete (2013) found that the logarithmic scoring rule scores higher the forecast with the highest entropy, the spherical scoring rule scores higher the forecast with the lowest entropy, and the quadratic scoring rule does not distinguish between the two forecasts. In other words, the logarithmic (respectively, spherical) scoring rule should be used when more (respectively, less) uncertainty is desirable.

Johnstone et al. (2011) proposed a different perspective by suggesting strategies for tailoring proper scoring rules to specific domains in a way that aligns the interests of an expert and the underlying decision maker. Along similar lines, Grant and Johnstone (2010) studied how probability forecasts can have different economic values for decision makers with different utility functions. Following a portfolio-theory perspective, Johnstone (2012) elaborated on the economic value of forecasts by suggesting that if a decision maker places bets based on experts' forecasts, then experts must be evaluated relative to each other; i.e., the decision maker should reward experts more for reporting accurate forecasts when other experts perform badly than when most experts do well.

In summary, our reading of the literature on choosing the most appropriate scoring rule indicates that (1) properness is not the only property that matters, (2) one must consider how different proper scoring rules evaluate different forecasts in terms of penalizing errors, and (3) whenever possible, one must use/incorporate information about the decision maker who will eventually use the forecast.

### 2.2. Circumventing Basic Assumptions

It is worth mentioning that recent years have seen a surge in approaches to circumvent the two main assumptions behind proper scoring rules, namely, risk neutrality and the existence of observable outcomes. Regarding the former, Armantier and Treich (2013) showed that eliciting beliefs with proper scoring rules might bias experts' reports. In particular, when the relative risk aversion with respect to the scoring range is increasing (decreasing), raising the scores from a proper scoring rule leads the expert to report more (less) uniform probabilities. After such a negative result, Armantier and Treich (2013) went further to suggest that "*one may want to consider the merits of eliciting beliefs without offering any financial reward for accuracy*" (Armantier and Treich 2013, p. 30; emphasis added).

Considering the potential issues that non-risk-neutral behavior brings to the elicitation process, some

authors suggested strategies to calibrate an expert's forecast by taking into account some components that drive the expert's attitude toward risk and uncertainty, such as utility functions and weighting functions (Winkler 1969, Winkler and Murphy 1970, Offerman et al. 2009, Kothiyal et al. 2011, Carvalho 2015). A different approach that does not depend on the elicitation of utility/weighting functions is to use well-crafted lotteries with payments based on proper scoring rules (Allen 1987, Karni 2009, Hossain and Okui 2013, Schlag and van der Weele 2013, Sandroni and Shmaya 2013). However, the effectiveness of such approaches is debatable (Selten et al. 1999).

Regarding the assumption of observable outcomes, the Bayesian truth serum (Prelec 2004) and the peer prediction (Miller et al. 2005) methods are the two most prominent methods based on proper scoring rules that induce risk-neutral experts to honestly report their private information without relying on the assumption of observable outcomes. Roughly speaking, these methods reward experts by making pairwise comparisons between experts' reports. Carvalho et al. (2016b) elaborated on the connections between proper scoring rules and peer prediction methods.

## 3. Research Methodology

The survey we present in this paper is the outcome of an attempt to collect and study academic publications related to applications of proper scoring rules. To this end, we conducted an initial search using Google Scholar and Web of Science with the exact terms (including quotes) "*proper scoring rules*," "*Brier scoring rule*," "*quadratic scoring rule*," "*logarithmic scoring rule*," and "*spherical scoring rule*." The first term, "*proper scoring rules*," was an obvious choice since it is the main concept in our study. We used the keywords "*quadratic scoring rule*," "*logarithmic scoring rule*," and "*spherical scoring rule*" because they represent three of the most popular proper scoring rules, as we discussed in Section 2. Since the quadratic scoring rule is also known as the "*Brier scoring rule*," we also decided to include the latter term in our initial search.

Next, we reviewed the full text of each publication to eliminate articles unrelated to applications of proper scoring rules, e.g., theoretical articles devoid of any application context and articles referencing, but not applying proper scoring rules. Consequently, our data set includes (1) articles of empirical nature where proper scoring rules were used when eliciting experts' beliefs and/or to evaluate the accuracy of reported forecasts and (2) articles of theoretical nature that have an explicit context, e.g., articles proposing different proper scoring rules tailored to a certain domain.

We also examined the references in each article so as to identify potentially missing publications. Finally, to avoid duplicates in our data set of articles, we removed dissertations and unpublished papers (e.g., working papers and technical notes). Our final data set contains 201 articles published between 1950 and 2015 in academic journals, conference proceedings, or as book chapters. This distinction between venues is later used in our analysis to explain the current publication trends when it comes to applications of proper scoring rules.

For the sake of topic analysis, we classified each article in our data set according to its context. To this end, whenever available, we used the article's keywords as the basis for our classification. For example, the articles we classified under the category of "*prediction markets*" often have keywords such as "*information markets*," "*prediction markets*," "*market scoring rules*," and "*logarithmic market scoring rule*." Alternatively, the articles we classified under the category of "*energy*" have keywords such as "*energy*," "*electricity*," "*smart grids*," and "*smart meters*." When classifying the articles, we tried as much as possible to minimize the final number of categories. However, as we discuss in the next section, some research topics are very specific, ranging from fair division to natural language recognizers. As a consequence, some categories ended up having only a few articles. The appendix shows all the articles in our data set as well as their respective categories.

We conclude this section by noting that, because of the nature of this research endeavor, it is virtually impossible to construct a data set containing all the articles about applications of proper scoring rules. For instance, although all the articles in our data set are written in English, there might exist relevant articles published in other languages and/or in academic outlets not indexed by search engines such as Google Scholar and Web of Science. This review can therefore be characterized as extended, but by no means as exhaustive. Nonetheless, we hope it will serve as a comprehensive basis for understanding applications of proper scoring rules.

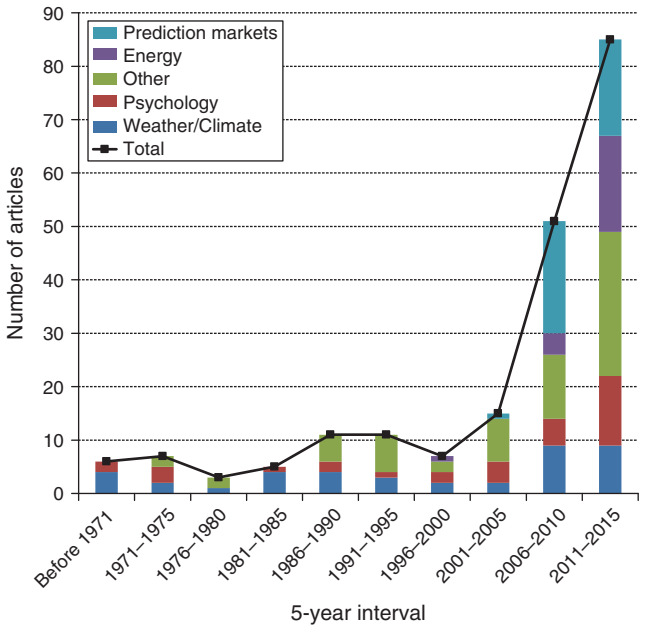
## 4. Data Analysis

We start our analysis by investigating the development of applications of proper scoring rules over the years. Figure 2 shows the publication trend considering five-year intervals. Generally speaking, the number of published articles is growing over the five-year intervals, and the trend suggests that more publications can be expected in years to come.

Looking at Figure 2, we can distinguish between two periods of time when it comes to the number of published articles. The first period goes from 1950 to 2005 when, according to our data set, a total of 65 articles were published. Several seminal articles advancing the theory behind proper scoring rules were published during this first period. When it comes to applications, however, our data set indicates that at most 15 articles were published during each five-year interval. As we elaborate in the following subsection, most of the articles published during this first period were either related to weather/climate or to the psychology field.

Regarding the second period of time, which started in 2006, we note that there was a tremendous increase in the number of published articles about applications of proper scoring rules. In particular, our data set contains 136 articles published between 2006 and 2015, approximately 2.1 times more than what was published during the first period of time. This result is in agreement with what Figure 1 shows, i.e., that the overall interest in proper scoring rules also drastically increased during this second period of time. As we further explain in this section and Section 5, one of the main reasons behind such a surge in popularity is the increasing interest in proper scoring rules coming from relatively new research areas, such as prediction markets and crowdsourcing.

**Figure 2** (Color online) Publication Trend Over Five-Year Intervals

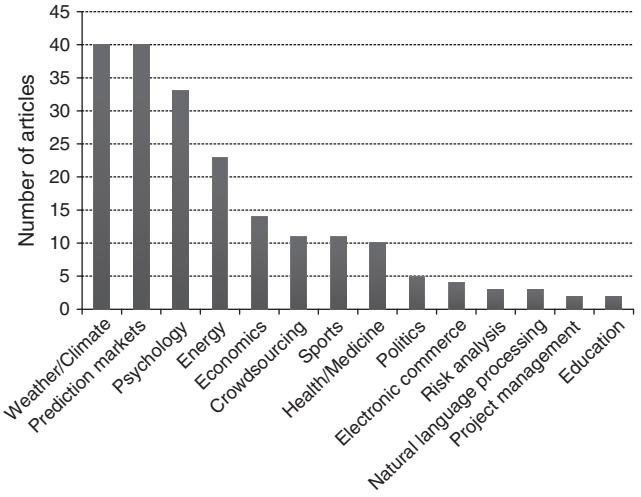



### 4.1. Topic Analysis

In our second analysis, we study the distribution of the articles in our data set over research topics. Figure 3 shows that the most popular topics in our data set are "weather/climate" and "prediction markets," with a total of 40 articles each. What is perhaps surprising is that "psychology" and "energy" are the next two most popular topics with, respectively, 33 and 23 articles. The remaining categories all contain 11 or fewer articles each. Henceforth, we group these categories into a single topic called "other."

Figure 2 shows the evolution of the most popular categories in our data set over five-year intervals. From 1950 to 2000, one can see that most of the articles in our data set are about weather/climate and psychology, followed by articles about other topics such as health/medicine and risk analysis. From 2000 on, there has been a big change in the publication patterns. In particular, recent years have seen a surge of articles about applications of proper scoring rules in prediction

**Figure 3** Frequency of the Research Topics in Our Data Set



markets, energy, and a variety of different areas, ranging from education to natural language processing. In the following subsections, we provide specific information on how proper scoring rules are applied in these areas. In Section 5, we discuss potential causes of the increasing number of publications.

**4.1.1. Weather/Climate.** To the best of our knowledge, the classic paper by Brier (1950) was the first to introduce proper scoring rules in the context of weather forecasting. Specifically, Brier (1950) suggested how to induce honest reporting of weather forecasts and how to evaluate the accuracy of a reported forecast by means of a variant of the quadratic scoring rule. Since Brier's (1950) article, there has been an increasing number of applications of proper scoring rules in the weather and climate domains, with a particular focus on the evaluation of forecasts. Several proper scoring rules have been applied in this context, including variants of the quadratic scoring rule (von Holstein 1971a, Murphy and Winkler 1982, Winkler 1994, Gneiting and Raftery 2007), proper scoring rules that explicitly score reported forecasts against baseline forecasting schemes to derive more informative skill scores (Murphy 1974, Brunet et al. 1988, Winkler 1994, Mason 2004, Ahrens and Walser 2008), and proper scoring rules that take the order of the underlying outcomes into account (Murphy and Daan 1984, Epstein 1988, Grimit et al. 2006, Jaun and Ahrens 2009).

Figure 2 suggests that the number of articles in our data set related to the weather and climate domains is more or less constant over the five-year intervals up to the year 2006. Since 2006, there has been a significant increase in the number of published articles. It is noteworthy that many important theoretical contributions to the field of proper scoring rules were published in weather-related academic journals. However, many such articles are not in our data set because they are devoid of a specific context, i.e., they are purely theoretical articles. Although we analyze the number of articles for different research outlets in Section 4.2, it is worth mentioning here about the high number of articles published in the weather journal *Monthly Weather Review*, which makes up slightly more than 7.4% of the articles in our data set (15 out of 201 articles).

**4.1.2. Prediction Markets.** Hanson (2003) suggested a new family of prediction markets defined in terms of proper scoring rules, which he called *market scoring rules* (*MSRs*). An MSR always has a complete probability distribution (market prices) over the entire outcome space. Any trader can at any time change any part of that distribution as long as he agrees to both pay the scoring rule payment associated with the current probability distribution and receive the scoring rule payment associated with the new probability distribution. For example, if outcome $\theta_x$ is realized, a trader who changes the probability distribution (market prices) from $\mathbf{q}$ to $\mathbf{q}'$ pays $R(\mathbf{q}, \theta_x)$, and receives $R(\mathbf{q}', \theta_x)$, where $R$ is a proper scoring rule. As Hanson (2007) suggested, the traders will eventually reach a consensus regarding the appropriate market prices:

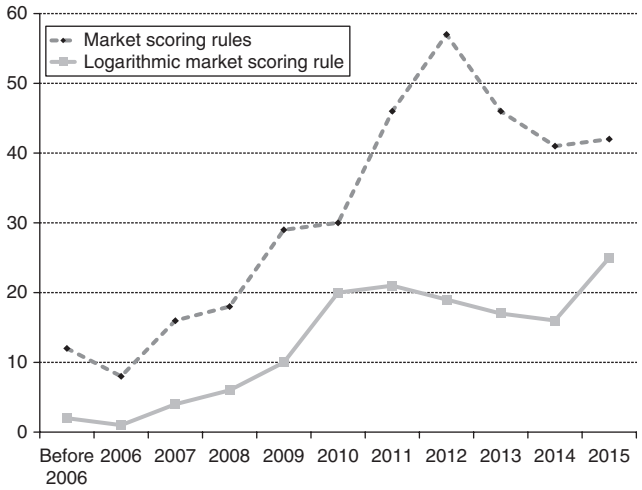
> Market scoring rules produce consensus estimates in the same way that betting markets produce consensus estimates. While each person is always free to change the current estimate, doing so requires taking on more risk, and eventually everyone reaches a limit where they do not want to make further changes, at least not until they receive further information. At this point the market can be said to be in equilibrium. (Hanson 2007, p. 8)

The most prominent market scoring rule, called the *logarithmic market scoring rule* (*LMSR*) (Hanson 2003, 2007), arises when one sets $R(\mathbf{q}, \theta_x) = b \ln q_x$, where $q_x \in (0, 1]$ represents an estimate that outcome $\theta_x$ will occur, and $b \in \mathbb{R}^+$ is a parameter that controls the liquidity of the market. For small values of $b$ (small liquidity), the market prices fluctuate wildly after every trade. Alternatively, the market prices move slowly with larger liquidity. Large liquidity is good for the traders since it stimulates more trades, but it comes at the expense of increasing the market maker's maximum loss exposure. Assuming that the initial market prices have the same value, the LMSR has a worst-case loss bounded by $b \ln n$ (Hanson 2003), where $n$ is the number of outcomes.

In practical applications, asking traders to modify probability distributions might not be very convenient. To solve this problem, Berg and Proebsting (2009) derived the necessary formulae to implement a prediction market equivalent to the LMSR, but in terms of buying and selling Arrow–Debreu contracts, instead of changing probability distributions.

At its core, an MSR is just an application of proper scoring rules to both elicit and aggregate traders' private information. Since suggested by Hanson (2003), there has been a tremendous growth in the number

**Figure 4** **Number of Articles Containing the Terms "Market Scoring Rules" and "Logarithmic Market Scoring Rule" According to Google Scholar**




of articles about MSRs and the LMSR, as Figure 4 shows. However, a careful look at Figure 4 reveals a mild decline in the number of published articles in the very recent years. A potential explanation for such a decline is the rise of new alternatives to MSRs, e.g., Bayesian prediction markets (Brahma et al. 2012). Most of the articles about prediction markets in our data set either describe empirical studies (e.g., see the articles by Ledyard et al. 2009, Dudik et al. 2013, Othman and Sandholm 2013, Slamka et al. 2013) or equilibrium results (e.g., see the articles by Chen et al. 2007, Dimitrov and Sami 2008, Chen et al. 2010, Dimitrov and Sami 2010, Iyer et al. 2010, Ostrovsky 2012, Gao et al. 2013). The proceedings of the academic conference previously known as *ACM Conference on Electronic Commerce*, which is now called the *ACM Conference on Economics and Computation*, is by far the most popular venue for publishing papers on market scoring rules, with a total of 13 articles according to our data set.

**4.1.3. Psychology.** Proper scoring rules and, in particular, variants of the quadratic scoring rule have long been used to reward elicited beliefs in psychological experiments. Some of these studies investigated the belief formation process (Hyndman et al. 2012a), whether the elicited beliefs are consistent with actions and/or models of belief learning (Nyarko and Schotter 2002, Costa-Gomes and Weizsäcker 2008, Danz et al. 2012), and potential hedging confounds during belief elicitation procedures (Blanco et al. 2010).

An early work by Jensen and Peterson (1973) investigated the psychological effects of different proper scoring rules when eliciting beliefs. The authors concluded that different types of scoring rules do not seem to influence probability assessments to any large extent, but they suggested that suboptimal strategies seemed to be employed when the rule contained both positive and negative scores. More recently, some researchers studied the effects of non-risk-neutral behavior on how experts report their beliefs under proper scoring rules. For example, Armantier and Treich (2013) characterized how proper scoring rules might bias reported beliefs for different scoring ranges, when a risk-averse expert has a financial stake in the event he is predicting, and when the expert can hedge his prediction by taking an additional action whose payoff depends on the outcome of the event. Among other empirical findings, Armantier and Treich (2013) showed that beliefs elicited with proper scoring rules are biased toward the probability value of 0.5 in settings involving binary outcomes, and that the strength of this result is positively correlated with the amount involved in the payments resulting from the proper scoring rule.

The fact that (human) experts might not be risk neutral has led many researchers to suggest different approaches for taking experts' risk attitudes into account when eliciting beliefs by means of proper scoring rules (Offerman et al. 2009, Kothiyal et al. 2011, Carvalho 2015). However, the suggested procedures are dependent on the assumed decision theory. Carvalho (2015) showed how different decision theories (rank-affected multiplicative weights versus prospect theory) might result in different calibration of beliefs. This point raises an important, yet still open question: which decision theory is the most appropriate theory when eliciting beliefs using proper scoring rules?

**4.1.4. Energy.** The fourth most popular research topic in our data set is the energy domain. In particular, there has been a range of applications of proper scoring rules connected to the concept of smart grids. For example, Rose et al. (2012) proposed a scoring rule-based mechanism that rewards household owners based on the precision of their reported information, which an aggregator eventually uses to procure electricity. Akasiadis and Chalkiadakis (2013), on the other hand, suggested using a proper scoring rule to incentivize agents to honestly report their reduction capacity

when forming energy cooperatives and shifting their electricity consumption.

Another interesting application was proposed by Chakraborty et al. (2013b). The authors presented a pricing scheme for smart houses where a provider monitors the network load and proposes a day-ahead pricing to the consumers. The consumers respond to that pricing proposal by providing a probabilistic prediction regarding the usage of smart devices. The provider incentivizes the consumers to report honestly by offering rewards based on a proper scoring rule.

A common keyword in the above articles is "smart." Recent developments in information technology and communication allow for different users, such as householders, to be "smart" by better monitoring and controlling electricity load. This opens up the possibility for such users to educate themselves about current usage and to formulate potential forecasts about future electricity consumption. In turn, these forecasts are valuable for electricity providers to adjust their procurement strategies or production plans. That is where proper scoring rules become very useful as a technique to both evaluate and elicit truthful forecasts.

We believe that the energy domain will continue to be one of the most active research areas in the near future when it comes to applications of proper scoring rules for two reasons. First, as Figure 2 shows, most of the articles related to the energy domain in our data set were published in the last five years. Second, key concepts such as smart grids and demand-side management, which naturally depend on electricity consumers providing reliable probabilistic information for electricity providers, are still in their infancy. This brings a lot of opportunities for creative research is terms of using proper scoring rules to reward consumers for providing information.

**4.1.5. Other Topics.** In what follows, we provide a brief overview of applications of proper scoring rules in research areas other than the four most popular areas discussed previously.

*Crowdsourcing.* The practice of leveraging collective intelligence by outsourcing problems and tasks to a crowd is often referred to as crowdsourcing. In a traditional crowdsourcing process, a requester elicits and aggregates information from an undefined crowd to solve a certain task. Some researchers have questioned the veracity and quality of the data obtained by means of crowdsourcing (Buhrmester et al. 2011, Carvalho et al. 2016b). It comes as no surprise that one way to induce honest reporting as well as to reward the crowd members for the reported information is by means of proper scoring rules. For example, proper scoring rules have been used to reward crowd members participating in community sensing (Faltings et al. 2012, 2014) and in routing tasks (Zhang et al. 2012).

*Economics.* Proper scoring rules have been used during the elicitation of forecasts regarding macroeconomic indicators (O'Carroll 1977, Casillas-Olvera and Bessler 2006) and the likelihood of bankruptcy (Johnstone et al. 2013), in financial markets (von Holstein 1972, Yates et al. 1991, Muradoglu and Onkal 1994, Lopez 2001, Lad et al. 2012), and in insurance (Gschlößl and Czado 2007). Another interesting application in the economics area was suggested by Carvalho and Larson (2010, 2011, 2012). Specifically, the authors investigated how to share a divisible good among a set of experts based on peer evaluations. In this context, the authors used proper scoring rules to induce honest reporting of peer evaluations.

*Education.* One highly promising research area for applications of proper scoring rules, but virtually unexplored according to our data set, concerns education. Because of constraints on the available resources (personnel, time, etc.), the grading process is one of the biggest challenges faced by instructors. Exams based on multiple-choice questions can alleviate the process because they can be quickly graded by computers. In this setting, students can be allowed to encode their uncertainty by assigning a probability value to each possible answer, as opposed to having to choose a single answer. This was the context Bickel (2010) studied. In particular, the author discussed how different scoring rules provide insights for both students and instructors in testing situations.

An interesting research direction in the context of education is to better understand which proper scoring rule is the most appropriate for a given multiple-choice question. For instance, it might be the case that the answers in the multiple-choice question can be naturally ordered, in which case proper scoring rules that take such an ordering into account are more appropriate (for example, see the work by Epstein 1969). Moreover, instructors might want to rank students, which in turn might be affected by the choice of the proper scoring rule, as we discussed in Section 2.1.

*Electronic Commerce.* Product reviews and ratings have become ubiquitous on electronic commerce websites aimed at consumers. In this context, it is clearly desirable that consumers report their true reviews/ratings, otherwise future consumers might end up basing their purchasing decisions on erroneous information. To this end, proper scoring rules have been used to provide rewards in order to induce honest reporting. An important issue in this context is the lack of a ground truth, i.e., the observed outcome required by proper scoring rules is nonexistent. Miller et al. (2005) proposed a way to circumvent this problem by, roughly speaking, using a random rater's rating as the observed outcome. This resulted in the now celebrated peer prediction method.

*Health/Medicine.* The provision of accurate probabilistic assessments of future events is a fundamental task for health workers collaborating in clinical or experimental medicine. Regarding this topic, proper scoring rules have been used to evaluate the accuracy of clinical diagnosis (Linnet 1988, 1989), clinicians' subjective estimates regarding some diseases (Dolan et al. 1986), and physicians' probabilistic estimates of survival in intensive care units (Winkler and Poses 1993).

*Natural Language Processing.* A surprising application of proper scoring rules was found in the field of natural language processing. Specifically, proper scoring rules and, in particular, variants of the logarithmic scoring rule have been used to evaluate the accuracy of confidence scores in the field of speaker detection (Brümmer and du Preez 2006, Campbell et al. 2006) and language recognition (Brummer and Van Leeuwen 2006).

*Risk Analysis.* A crucial component of risk analysis concerns the assessment of the likelihood that certain events of interest will occur. In this context, proper scoring rules have been used during the elicitation of probability distributions when analyzing environmental risks (Cunningham and Martell 1976), when measuring the cost of refurbishing pumping stations and the length of unrecorded sewers (Garthwaite and O'Hagan 2000), and when eliciting judgments about personal exposure to benzene (Walker et al. 2003).

*Sports.* Generally speaking, the domain of sports is very attractive for research on probabilistic predictions because of the biases that naturally arise as well as because of the often intense media coverage and scrutiny of the strengths and weaknesses of the teams and individual players, which provide useful information for the general public. As suggested by Winkler (1971), proper scoring rules might be useful in this context to induce individuals to make careful probability assessments. Recognizing this, some websites featuring sports contests, such as ProbabilityFootball.com, reward individuals' predictions based on proper scoring rules. As a consequence, their public data have been extensively used in research about proper scoring rules, e.g., see the work by Chen et al. (2005), Carvalho and Larson (2013), and Carvalho et al. (2016a).

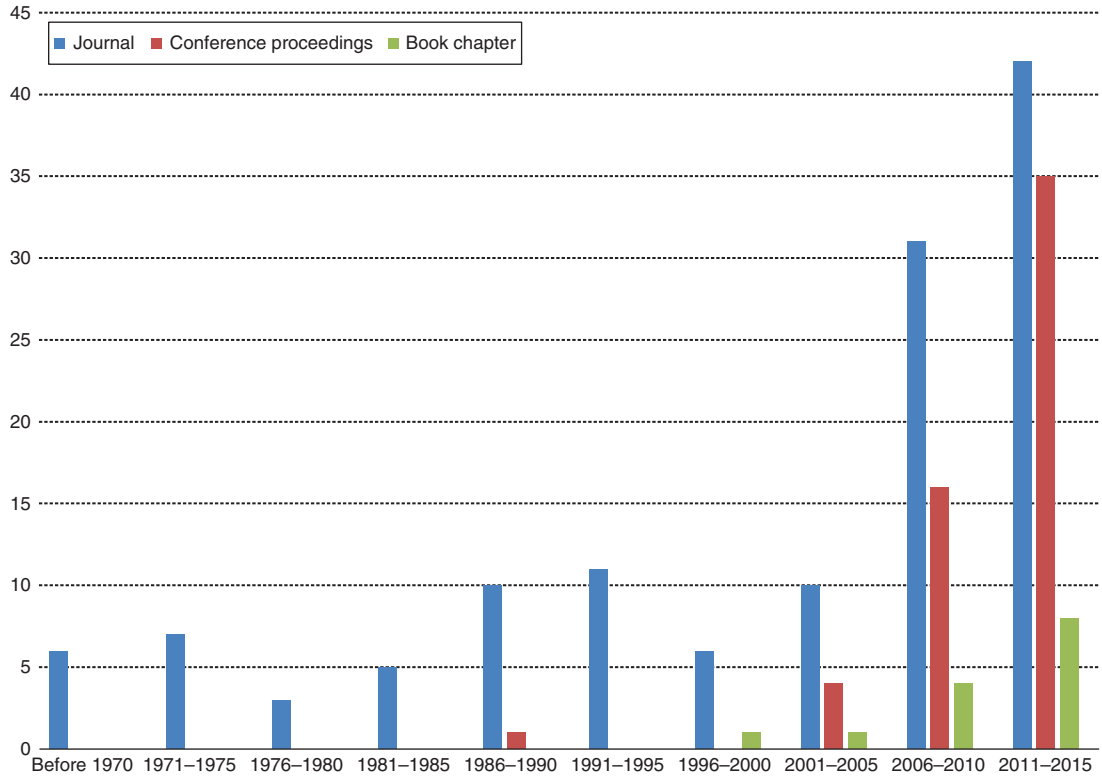
*Politics.* Regarding politics, our data set indicates that proper scoring rules have been applied to evaluate the accuracy of predictions about presidential elections (Heath and Tversky 1991, Pennock et al. 2002), in a geopolitical forecasting contest (Mellers et al. 2014), and to evaluate the opinions from policy makers regarding the occurrence of political events (Tetlock 2005).

*Project Management.* Our data set contains two articles about applications of proper scoring rules in project management. Bhola et al. (1992) suggested using a proper scoring rule for evaluating expert opinions to aid with the management of projects in a specific company. Bacon et al. (2012) studied a principal–agent setting where a worker and a manager may each have information about the likely completion time of a task. Moreover, the worker might also affect the completion time by choosing a level of effort. Bacon et al. (2012) suggested a family of scoring rules for the worker and manager that satisfies three properties: (1) information is truthfully reported, (2) the worker has incentives to complete tasks as quickly as possible, and (3) collusion is not profitable.

### 4.2. Analysis of Publication Outlets

Our last analysis is about the venues where articles on applications of proper scoring rules have been published. The main goal behind this analysis is to connect different research communities by indicating where to find/expect articles on proper scoring rules. In particular, we found that 131, 56, and 14 articles were published in, respectively, academic journals, conference proceedings, and as book chapters. From these numbers, one can immediately see that the number of journal publications is greater than the number of publications in conference proceedings and book chapters

**Figure 5** **(Color online) Number of Publications in Different Venues Over Five-Year Intervals**



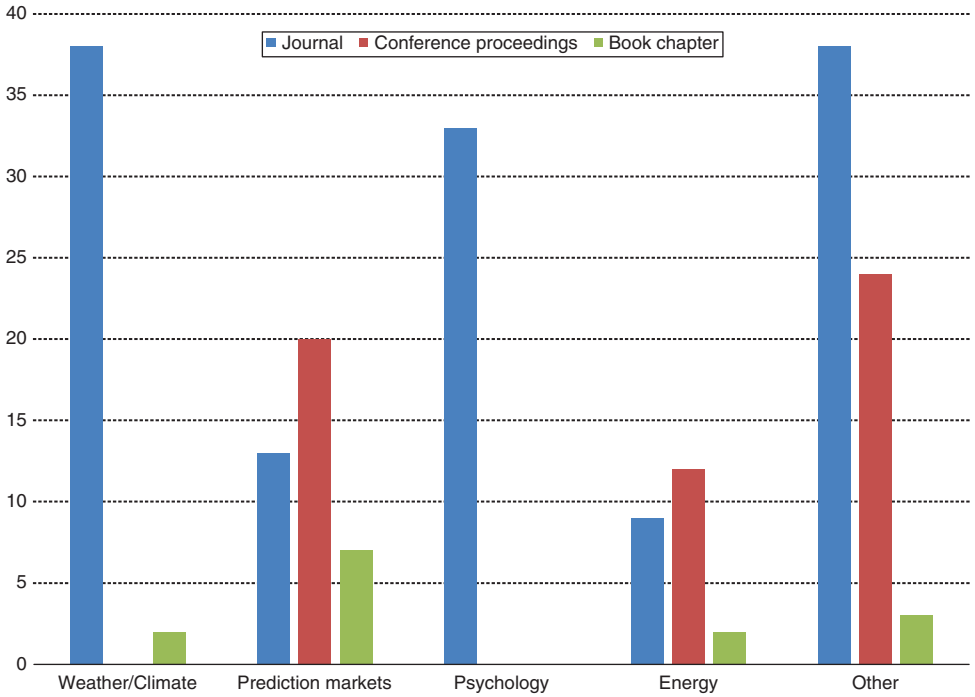

combined. Figure 5 shows, however, that this result has been drastically changing over the years. In particular, the number of articles published in conference proceedings is growing tremendously and, for the last five-year interval, it is almost the same as the number of papers published in academic journals (35 against 42).

Figure 6 shows the number of articles per topic in our data set published in different venues. It is noticeable that almost all the articles about weather/climate, a traditional application domain, were published in academic journals. On the other hand, articles in emerging application domains, such as prediction markets and energy, were mostly published in conference proceedings.

After taking a closer look at the articles in our data set, we notice that the computer science and, more specifically, the artificial intelligence communities are behind the growth in the number of articles published in conference proceedings. In particular, the top conferences ranked by the number of published articles about applications of proper scoring rules are the *ACM Conference on Electronic Commerce*, which is now called the *ACM Conference on Economics and Computation* (17 articles); the *International Conference on Autonomous Agents and Multiagent Systems* (10 articles); and the *Conference on Uncertainty in Artificial Intelligence* (5 articles). Regarding academic journals, the most popular venues according to our data set are the *Monthly Weather Review* (15 articles); *Organizational Behavior and Human Performance*, which is now called *Organizational Behavior and Human Decision Processes* (7 articles); and the *Journal of Economic Behavior & Organization* (6 articles).

One can argue that the increasing number of articles in conference proceedings is attributable to these conferences being relatively younger than most well-established journals. We note, however, that conferences are the primary venue for publishing research in computer science (Vardi 2010), where premier conferences tend to be decades old. For example, at the time of writing, the ACM Conference on Economics and Computation and the Conference on Uncertainty in Artificial Intelligence are, respectively, in their 17th

**Figure 6** **(Color online) Number of Articles per Topic Published in Different Venues**




and 32nd editions. Even though these conferences are many years old, articles on proper scoring rules have only recently been published in their proceedings. This leads us to conclude that the interest of the computer science community in proper scoring rules is a recent phenomenon. We discuss in the next section a potential explanation behind it.

One might wonder whether there is any substantial difference between articles published in AI conference proceedings and articles published in academic journals in other fields. We cannot find any significant difference when proper scoring rules are used to evaluate the accuracy of forecasts. We do note, however, a different perspective when the main goal behind the use of proper scoring rules is to induce honest reporting of beliefs. Before discussing this difference, it is important to note that the field of artificial intelligence aims at building rational agents who can perceive the environment and take actions toward specific goals. As noted by Parkes and Wellman (2015), theories of normative behavior may end up being more relevant when dealing with artificial agents than human beings since one can always predefine the behavior of the former.

In light of the above discussion, we find that the AI community tends to treat proper scoring rules as black boxes that are able to induce the underlying agents to behave honestly, no matter the context or which scoring rule is used. In particular, this happens because the assumption of risk neutrality is taken for granted. Once again, this is a fair assumption in AI because, different from human beings, one can impose a certain decision model on artificial agents.

It is also interesting to mention that the AI community has been at the forefront of research on how to induce honest reporting of beliefs without the assumption of observable outcomes. Many important refinements of the two most prominent methods we discussed in Section 2.2, namely, the Bayesian truth serum and the peer prediction method, were proposed by AI researchers, e.g., the robust Bayesian truth serum (Witkowski and Parkes 2012, Radanovic and Faltings 2013) and the minimal peer prediction mechanism (Radanovic and Faltings 2015). Since experimental work involving these methods is still lacking, we see this as an opportunity for multidisciplinary collaboration involving the AI community and researchers from areas such as decision analysis and behavioral economics.

## 5. Concluding Remarks

Proper scoring rules are traditional techniques to measure the accuracy of forecasts as well as to induce risk-neutral experts to honestly report their beliefs, what Winkler and Murphy (1969) called, respectively, the substantive and normative standards of goodness. In this paper, we provided an overview of applications of proper scoring rules; i.e., we examined scenarios where proper scoring rules were applied in a normative and/or in a substantive way. To this end, we collected a total of 201 articles published between 1950 and 2015.

Our first finding was that there has been a tremendous increase in the number of published articles about proper scoring rules over the years, in particular after the year of 2005. Regarding application domains, we found that the four most popular domains are the weather/climate, prediction markets, psychology, and energy. Furthermore, we also found that proper scoring rules have been applied to a variety of other domains, including crowdsourcing, economics, education, electronic commerce, health/medicine, natural language processing, risk analysis, sports, politics, and project management.

We noted an interesting change in the publication patterns over the years, in that many articles about proper scoring rules are now being published in conference proceedings. A closer inspection of this result showed that most of these conferences are linked to the artificial intelligence community. We believe this result has the potential for connecting different research communities by suggesting where researchers can expect to find or publish articles related to proper scoring rules.

By no means do we claim that this review is exhaustive, but we hope it will provide academics and practitioners with a starting point for further study of the relevant literature. We conclude this paper with an attempt to understand the reason behind the growing interest in proper scoring rules. In particular, our reading of the literature suggests that the *wisdom-of-crowds phenomenon* is a driving force behind the recent popularity of proper scoring rules.

Roughly speaking, the concept of wisdom of crowds means that when information is elicited from a relevant crowd and appropriately aggregated, the resulting aggregate information is, at least in expectation, more accurate than the information from any single individual. The wisdom-of-crowds topic has received substantial attention in the business and academic worlds since the publication of Surowiecki's (2005) book *The Wisdom of Crowds* and the article by Howe (2006), which defined the related term *crowdsourcing*. It is interesting to note that both Surowiecki's (2005) book and Howe's (2006) article were published around the same time that proper scoring rules started gaining tremendous popularity.

The two keywords in the above definition of wisdom of crowds are *elicitation* and *aggregation*. Regarding elicitation, the crowd members might be self-interested, meaning that they might behave strategically when reporting their individual, private information. Hence, a suitable incentive scheme must be used to incentivize honest reporting. Proper scoring rules provide a compelling way for eliciting forecasts in an incentive-compatible manner. This is the perspective taken by many authors when using proper scoring rules in relatively new application areas, e.g., Bacon et al. (2012) in project management, Zhang et al. (2012) in crowdsourcing, and Akasiadis and Chalkiadakis (2013) in the energy domain.

Regarding aggregation, it is well known in management science and operations research that combining predictions from multiple different sources often leads to improved forecasting performance (Clemen 1989, Hendry and Clements 2004). As discussed in Section 4.1.2, market scoring rules allow a decision maker to elicit and aggregate information from a potentially large group of people. The seminal paper by Hanson (2003), which defined market scoring rules, was published a few years before the sudden increase in interest in proper scoring rules. Since then, market scoring rules opened up a new stream of research regarding the use of proper scoring rules.

### Acknowledgments

The author acknowledges an associate editor and three anonymous reviewers for the constructive comments. The author also thanks Ben Bode for useful discussions and Jeroen van Peer for helping with data collection.

### Appendix

Table A.1 shows the articles in our data set as well as their respective topics.

**Table A.1 Classification of the Articles in Our Data Set per Topic**

| Topic | Articles |
|---|---|
| Crowdsourcing | Jurca and Faltings (2009), Faltings et al. (2012), Kamar and Horvitz (2012), Zhang et al. (2012), Ray et al. (2013), Sakurai et al. (2013), Cao et al. (2014), Faltings et al. (2014), Oka et al. (2014), Sakurai et al. (2015), Ugander et al. (2015) |
| Economics | von Holstein (1972), O'Carroll (1977), Yates et al. (1991), Muradoglu and Onkal (1994), Hirtle and Lopez (1999), Lopez (2001), Casillas-Olvera and Bessler (2006), Gschlößl and Czado (2007), Carvalho and Larson (2010, 2011, 2012), Diebold and Mariano (2012), Lad et al. (2012), Johnstone et al. (2013) |
| Education | Bickel (2007, 2010) |
| Electronic commerce | Miller et al. (2005), Gerding et al. (2010), Cai et al. (2013), Radanovic and Faltings (2015) |
| Energy | Woo et al. (1998), Gneiting et al. (2007), Pinson et al. (2007), Petroliagis et al. (2010), Thorarinsdottir and Gneiting (2010), Gneiting and Ranjan (2011), Rose et al. (2011), Chakraborty and Ito (2012), Friederichs and Thorarinsdottir (2012), Robu et al. (2012), Rose et al. (2012), Akasiadis and Chalkiadakis (2013), Bagchi et al. (2013), Chakraborty et al. (2013a, b), Egri and Váncza (2013), Chakraborty et al. (2014), Mhanna et al. (2014a, b), Chakraborty and Ito (2015), Hara and Ito (2015), Mhanna et al. (2015), Scheuerer and Hamill (2015) |
| Health/Medicine | Dolan et al. (1986), Spiegelhalter (1986), Linnet (1988, 1989), Spiegelhalter et al. (1990), Bernardo and Muñoz (1993), Winkler and Poses (1993), Madigan et al. (1995), Dawid and Musio (2013), Conigliani et al. (2015) |
| Natural language processing | Brümmer and du Preez (2006), Brummer and Van Leeuwen (2006), Campbell et al. (2006) |
| Politics | Heath and Tversky (1991), Pennock et al. (2002), Tetlock (2005), Merkle and Steyvers (2013), Mellers et al. (2014) |
| Prediction markets | Hanson (2003), Abramovicz (2006, 2007), Chen (2007), Chen et al. (2007), Hanson (2007), Nikolova and Sami (2007), Chen et al. (2008), Dimitrov and Sami (2008), Agrawal et al. (2009), Conitzer (2009), Gao et al. (2009), Guo and Pennock (2009), Ledyard et al. (2009), Shi et al. (2009), Chen et al. (2010), Chen and Pennock (2010), Chen and Vaughan (2010a, b), Dimitrov and Sami (2010), Iyer et al. (2010), Othman and Sandholm (2010), Abernethy et al. (2011), Chen et al. (2011), Othman and Sandholm (2011), Chen et al. (2012), Ostrovsky (2012), Abernethy et al. (2013), Chen et al. (2013), Dudik et al. (2013), Gao et al. (2013), Jumadinova and Dasgupta (2013), Li and Vaughan (2013), Othman et al. (2013), Othman and Sandholm (2013), Slamka et al. (2013), Abernethy et al. (2014), Chen et al. (2014), Abernethy and Johnson-Roberson (2015), Chakraborty et al. (2015) |
| Project management | Bhola et al. (1992), Bacon et al. (2012) |
| Psychology | Phillips and Edwards (1966), Schum et al. (1967), von Holstein (1971b, c), Jensen and Peterson (1973), Fischer (1982), Tetlock and Kim (1987), Nelson and Bessler (1989), van Lenthe (1994), Offerman et al. (1996), Friedman and Massaro (1998), Huck and Weizsäcker (2002), Nyarko and Schotter (2002), Guerra and Zizzo (2004), Offerman and Sonnemans (2004), Costa-Gomes and Weizsäcker (2008), Offerman et al. (2009), Palfrey and Wang (2009), Rutström and Wilcox (2009), Blanco et al. (2010), Wang (2011), Danz et al. (2012), Hyndman et al. (2012a, b), Koessler et al. (2012), Armantier and Treich (2013), Hossain and Okui (2013), Manski and Neri (2013), Andersen et al. (2014), Blanco et al. (2014), Harrison et al. (2014), Trautmann and van de Kuilen (2014), Offerman and Palley (2015) |
| Risk analysis | Cunningham and Martell (1976), Garthwaite and O'Hagan (2000), Walker et al. (2003) |

**Table A.1 (Continued)**

| Topic | Articles |
|---|---|
| Sports | Winkler (1971), Lawrence et al. (2002), Debnath et al. (2003), Chen et al. (2005), Grant and Johnstone (2010), Štrumbelj and Šikonja (2010), Constantinou and Fenton (2012), Carvalho and Larson (2013), Deloatch et al. (2013), Carvalho et al. (2015, 2016a) |
| Weather/Climate | Brier (1950), Sanders (1963), Winkler and Murphy (1969), Glahn and Jorgensen (1970), von Holstein (1971a), Murphy (1974), Charba and Klein (1980), Murphy and Winkler (1982, 1984), Murphy and Daan (1984), Murphy (1985), Brunet et al. (1988), Epstein (1988), Murphy et al. (1989), Murphy (1990), Murphy and Winkler (1992), Murphy (1993), Winkler (1994), Katz and Murphy (1997), Wilson et al. (1999), Roulston and Smith (2002), Mason (2004), Grimit et al. (2006), Friederichs and Hense (2007), Gneiting and Raftery (2007), Ahrens and Walser (2008), Bröcker and Smith (2008), Gneiting et al. (2008), Jaun and Ahrens (2009), Weijs et al. (2010a, b), Chmielecki and Raftery (2011), Wilks (2011), Ehm and Gneiting (2012), Fricker et al. (2013), Lerch and Thorarinsdottir (2013), Thorarinsdottir et al. (2013), Christensen (2015), Christensen et al. (2015), Smith et al. (2015) |

## References

Abernethy J, Johnson-Roberson M (2015) Financialized methods for market-based multi-sensor fusion. *2015 IEEE/RSJ Internat. Conf. Intelligent Robots and Systems* (Institute of Electrical and Electronics Engineers, Hamburg, Germany), 900–907.

Abernethy J, Chen Y, Vaughan JM (2011) An optimization-based framework for automated market-making. *Proc. 12th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 297–306.

Abernethy J, Chen Y, Vaughan JW (2013) Efficient market making via convex optimization, and a connection to online learning. *ACM Trans. Econom. Comput.* 1(2):1–39.

Abernethy J, Kutty S, Lahaie S, Sami R (2014) Information aggregation in exponential family markets. *Proc. 15th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 395–412.

Abramowicz M (2006) Deliberative information markets for small groups. Hahn R, Tetlock P, eds. *Information Markets: A New Way of Making Decisions* (Aei Press, Washington, DC), 101–125.

Abramovicz M (2007) The hidden beauty of the quadratic market scoring rule: A uniform liquidity market maker, with variations. *J. Prediction Markets* 1(2):111–125.

Agrawal S, Delage E, Peters M, Wang Z, Ye Y (2009) A unified framework for dynamic pari-mutuel information market design. *Proc. 10th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 255–264.

Ahrens B, Walser A (2008) Information-based skill scores for probabilistic forecasts. *Monthly Weather Rev.* 136(1):352–363.

Akasiadis C, Chalkiadakis G (2013) Agent cooperatives for effective power consumption shifting. *27th AAAI Conf. Artificial Intelligence*, 1263–1269.

Allen F (1987) Discovering personal probabilities when utility functions are unknown. *Management Sci.* 33(4):542–544.

Andersen S, Fountain J, Harrison GW, Rutström EE (2014) Estimating subjective probabilities. *J. Risk Uncertainty* 48(3):207–229.

Armantier O, Treich N (2013) Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *Eur. Econom. Rev.* 62(August):17–40.

Bacon DF, Chen Y, Kash I, Parkes DC, Rao M, Sridharan M (2012) Predicting your own effort. *Proc. 11th International Conf. Autonomous Agents and Multiagent Systems*, 695–702.

Bagchi D, Biswas S, Narahari Y, Viswanadham N, Suresh P, Subrahmanya SV (2013) Incentive compatible green procurement using scoring rules. *Proc. 2013 IEEE Internat. Conf. Automation Sci. Engrg.*, 504–509.

Berg H, Proebsting TA (2009) Hanson's automated market maker. *J. Prediction Markets* 3(1):45–59.

Bernardo JM, Muñoz J (1993) Bayesian analysis of population evolution. *Statistician* 42(5):541–550.

Bhola B, Cooke RM, Blaauw HG, Kok M (1992) Expert opinion in project management. *Eur. J. Oper. Res.* 57(1):24–31.

Bickel JE (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Anal.* 4(2):49–65.

Bickel JE (2010) Scoring rules and decision analysis education. *Decision Anal.* 7(4):346–357.

Blanco M, Engelmann D, Koch AK, Normann H-T (2010) Belief elicitation in experiments: Is there a hedging problem? *Experiment. Econom.* 13(4):412–438.

Blanco M, Engelmann D, Koch AK, Normann HT (2014) Preferences and beliefs in a sequential social dilemma: A within-subjects analysis. *Games Econom. Behav.* 87:122–135.

Brahma A, Chakraborty M, Das S, Lavoie A, Magdon-Ismail M (2012) A Bayesian market maker. *Proc. 13th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 215–232.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Bröcker J, Smith LA (2008) From ensemble forecasts to predictive distribution functions. *Tellus A* 60(4):663–678.

Brümmer N, du Preez J (2006) Application-independent evaluation of speaker detection. *Comput. Speech Language* 20(2):230–275.

Brummer N, Van Leeuwen D (2006) On calibration of language recognition scores. *Proc. IEEE Speaker and Language Recognition Workshop*, 1–8.

Brunet N, Verret R, Yacowar N (1988) An objective comparison of model output statistics and "perfect prog" systems in producing numerical weather element forecasts. *Weather and Forecasting* 3(4):273–283.

Buhrmester MD, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psych. Sci.* 6(1):3–5.

Cai Y, Mahdian M, Mehta A, Waggoner B (2013) Designing markets for daily deals. Chen Y, Immorlica N, eds. *Web and Internet Economics*, Lecture Notes in Computer Science, Vol. 8289 (Springer-Verlag, New York), 82–95.

Campbell WM, Brady KJ, Campbell JP, Granville R, Reynolds DA (2006) Understanding scores in forensic speaker recognition. *Proc. IEEE Speaker and Language Recognition Workshop*, 1–8.

Cao CC, Chen L, Jagadish HV (2014) From labor to trader: Opinion elicitation via online crowds as a market. *Proc. 20th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York), 1067–1076.

Carvalho A (2015) Tailored proper scoring rules elicit decision weights. *Judgment Decision Making* 10(1):86–96.

Carvalho A, Larson K (2010) Sharing a reward based on peer evaluations. *Proc. 9th Internat. Conf. Autonomous Agents and Multiagent Systems*, 1455–1456.

Carvalho A, Larson K (2011) A truth serum for sharing rewards. *Proc. 10th Internat. Conf. Autonomous Agents and Multiagent Systems*, 635–642.

Carvalho A, Larson K (2012) Sharing rewards among strangers based on peer evaluations. *Decision Anal.* 9(3):253–273.

Carvalho A, Larson K (2013) A consensual linear opinion pool. *Proc. 23rd Internat. Joint Conf. Artificial Intelligence*, 2518–2524.

Carvalho A, Dimitrov S, Larson K (2015) A study on the influence of the number of MTurkers on the quality of the aggregate output. Bulling N, ed. *Multi-Agent Systems*, Lecture Notes in Computer Science, Vol. 8953 (Springer-Verlag, New York), 285–300.

Carvalho A, Dimitrov S, Larson K (2016a) How many crowdsourced workers should a requester hire? *Ann. Math. Artificial Intelligence* 78(1):45–72.

Carvalho A, Dimitrov S, Larson K (2016b) Inducing honest reporting of private information in the presence of social projection. *Decision*, ePub ahead of print March 14, http://dx.doi.org/10.1037/dec0000052.

Casillas-Olvera G, Bessler DA (2006) Probability forecasting and central bank accountability. *J. Policy Modeling* 28(2):223–234.

Chakraborty S, Ito T (2012) Smart house load management scheme using scoring rule based optimal time dependent pricing. *Proc. 2012 Joint Agent Workshop and Sympos.* 1–8.

Chakraborty S, Ito T (2015) Hierarchical scoring rule based smart dynamic electricity pricing scheme. Bai Q, Ren F, Zhang M, Ito T, Tang X, eds. *Smart Modeling and Simulation for Complex Systems*, Studies in Computational Intelligence, Vol. 564 (Springer, Japan, Tokyo), 113–131.

Chakraborty M, Das S, Peabody J (2015) Price evolution in a continuous double auction prediction market with a scoring-rule based market maker. *Proc. 29th AAAI Conf. Artificial Intelligence*, 835–841.

Chakraborty S, Ito T, Hara K (2013a) Incentive based smart pricing scheme using scoring rule. *Proc. 4th IEEE/PES Innovative Smart Grid Technologies Europe*, 1–5.

Chakraborty S, Ito T, Senjyu T (2014) Smart pricing scheme: A multilayered scoring rule application. *Expert Systems with Applications* 41(8):3726–3735.

Chakraborty S, Ito T, Kanamori R, Senjyu T (2013b) Application of incentive based scoring rule deciding pricing for smart houses. *Proc. 2013 IEEE Power and Energy Soc. General Meeting*, 1–5.

Charba JP, Klein WH (1980) Skill in precipitation forecasting in the national weather service. *Bull. Amer. Meteorological Soc.* 61(12):1546–1555.

Chen Y (2007) A utility framework for bounded-loss market makers. *Proc. 23rd Conf. Uncertainty in Artificial Intelligence*, 49–56.

Chen Y, Pennock DM (2010) Designing markets for prediction. *AI Magazine* 31(4):42–52.

Chen Y, Vaughan JW (2010a) A new understanding of prediction markets via no-regret learning. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 189–198.

Chen Y, Vaughan JW (2010b) Connections between markets and learning. *ACM SIGecom Exchanges* 9(1):6.

Chen Y, Ruberry M, Vaughan JW (2012) Designing informative securities. *Proc. 28th Conf. Uncertainty in Artificial Intelligence*, 185–195.

Chen Y, Ruberry M, Vaughan JM (2013) Cost function market makers for measurable spaces. *Proc. 14th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 785–802.

Chen Y, Chu CH, Mullen T, Pennock DM (2005) Information markets vs. opinion pools: An empirical comparison. *Proc. 6th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 58–67.

Chen Y, Gao XA, Goldstein R, Kash IA (2014) Market manipulation with outside incentives. *Autonomous Agents and Multi-Agent Systems* 29(2):230–265.

Chen Y, Kash I, Ruberry M, Shnayder V (2011) Decision markets with good incentives. Chen N, Elkind E, Koutsoupias E, eds. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 7090 (Springer-Verlag, New York), 72–83.

Chen Y, Fortnow L, Lambert N, Pennock DM, Wortman J (2008) Complexity of combinatorial market makers. *Proc. 9th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 190–199.

Chen Y, Reeves DM, Pennock DM, Hanson RD, Fortnow L, Gonen R (2007) Bluffing and strategic reticence in prediction markets. Deng X, Graham F, eds. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 4858 (Springer, Berlin), 70–81.

Chen Y, Dimitrov S, Sami R, Reeves DM, Pennock DM, Hanson RD, Fortnow L, Gonen R (2010) Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica* 58(4): 930–969.

Chmielecki RM, Raftery AE (2011) Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Rev.* 139(5):1626–1636.

Christensen HM (2015) Decomposition of a new proper score for verification of ensemble forecasts. *Monthly Weather Rev.* 143(5):1517–1532.

Christensen HM, Moroz IM, Palmer TN (2015) Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quart. J. Roy. Meteorological Soc.* 141(687):538–549.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Conigliani C, Manca A, Tancredi A (2015) Prediction of patient-reported outcome measures via multivariate ordered probit models. *J. Roy. Statist. Soc. Ser. A* 178(3):567–591.

Conitzer V (2009) Prediction markets, mechanism design, and cooperative game theory. *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, 101–108.

Constantinou AC, Fenton NE (2012) Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *J. Quant. Anal. Sports* 8(1):1.

Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, Oxford, UK).

Costa-Gomes MA, Weizsäcker G (2008) Stated beliefs and play in normal-form games. *Rev. Econom. Stud.* 75(3):729–762.

Cunningham AA, Martell DL (1976) The use of subjective probability assessments to predict forest fire occurrence. *Canadian J. Forest Res.* 6(3):348–356.

Danz DN, Fehr D, Kübler D (2012) Information and beliefs in a repeated normal-form game. *Experiment. Econom.* 15(4): 622–640.

Dawid AP (2007) The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* 59(1):77–93.

Dawid AP, Musio M (2013) Estimation of spatial processes using local scoring rules. *AStA Adv. Statist. Anal.* 97(2):173–179.

Debnath S, Pennock DM, Giles CL, Lawrence S (2003) Information incorporation in online in-game sports betting markets. *Proc. 4th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 258–259.

Deloatch R, Marmarchi A, Kirlik A (2013) Testing the conditions for acquiring intuitive expertise in judgment evidence from a study of NCAA basketball tournament predictions. *Proc. Human Factors and Ergonomics Soc. Annual Meeting*, Vol. 57, 290–294.

Diebold FX, Mariano RS (2012) Comparing predictive accuracy. *J. Bus. Econom. Statist.* 20(1):134–144.

Dimitrov S, Sami R (2008) Non-myopic strategies in prediction markets. *Proc. 9th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 200–209.

Dimitrov S, Sami R (2010) Composition of markets with conflicting incentives. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 53–62.

Dolan JG, Bordley DR, Mushlin AI (1986) An evaluation of clinicians' subjective prior probability estimates. *Medical Decision Making* 6(4):216–223.

Dudik M, Lahaie S, Pennock DM, Rothschild D (2013) A combinatorial prediction market for the US elections. *Proc. 14th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 341–358.

Egri P, Váncza J (2013) Efficient mechanism for aggregate demand prediction in the smart grid. *Multiagent System Technologies*, Lecture Notes in Computer Science, Vol. 8076 (Springer, Berlin), 250–263.

Ehm W, Gneiting T (2012) Local proper scoring rules of order two. *Ann. Statist.* 40(1):609–637.

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorology* 8(6):985–987.

Epstein ES (1988) Long-range weather prediction: limits of predictability and beyond. *Weather and Forecasting* 3(1):69–75.

Faltings B, Li JJ, Jurca R (2012) Eliciting truthful measurements from a community of sensors. *Proc. 3rd Internat. Conf. Internet of Things*, 47–54.

Faltings B, Li JJ, Jurca R (2014) Incentive mechanisms for community sensing. *IEEE Trans. Comput.* 63(1):115–128.

Fischer GW (1982) Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting. *Organ. Behav. Human Performance* 29(3):352–369.

Forbes PGM (2012) Compatible weighted proper scoring rules. *Biometrika* 99(4):989–994.

Fricker TE, Ferro CAT, Stephenson DB (2013) Three recommendations for evaluating climate predictions. *Meteorological Appl.* 20(2):246–255.

Friederichs P, Hense A (2007) Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Rev.* 135(6):2365–2378.

Friederichs P, Thorarinsdottir TL (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* 23(7):579–594.

Friedman D, Massaro DW (1998) Understanding variability in binary and continuous choice. *Psychonomic Bull. Rev.* 5(3): 370–389.

Gao X, Chen Y, Pennock DM (2009) Betting on the real line. Leonardi S, ed. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 5929 (Springer, Berlin), 553–560.

Gao XA, Zhang J, Chen Y (2013) What you jointly know determines how you act: Strategic interactions in prediction markets. *Proc. 14th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 489–506.

Garthwaite PH, O'Hagan A (2000) Quantifying expert opinion in the UK water industry: An experimental study. *J. Roy. Statist. Soc. Ser. D* 49(4):455–477.

Gerding EH, Larson K, Jennings NR (2010) Eliciting expert advice in service-oriented computing. David E, Gerding E, Sarne D, Shehory O, eds. *Agent-Mediated Electronic Commerce*, Lecture Notes in Business Information Processing, Vol. 59 (Springer, Berlin), 29–43.

Glahn HR, Jorgensen DL (1970) Climatological aspects of the brier *P*-score. *Monthly Weather Rev.* 98(2):136–141.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102(477):359–378.

Gneiting T, Ranjan R (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* 29(3):411–422.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *J. Roy. Statist. Soc. Ser. B* 69(2): 243–268.

Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* 17(2):211–235.

Grant A, Johnstone D (2010) Finding profitable forecast combinations using probability scoring rules. *Internat. J. Forecasting* 26(3):498–510.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteorological Soc.* 132:1–17.

Gschlößl S, Czado C (2007) Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial J.* 2007(3):202–225.

Guerra G, Zizzo DJ (2004) Trust Responsiveness and Beliefs. *J. Econom. Behav. Organ.* 55(1):25–30.

Guo M, Pennock DM (2009) Combinatorial prediction markets for event hierarchies. *Proc. 8th Internat. Conf. Autonomous Agents and Multiagent Systems*, 201–208.

Hanson R (2003) Combinatorial information market design. *Inform. Systems Frontiers* 5(1):107–119.

Hanson R (2007) Logarithmic market scoring rules for modular combinatorial information aggregation. *J. Prediction Markets* 1(1):3–15.

Hara K, Ito T (2015) A scoring rule-based truthful demand response mechanism. *2015 IEEE/ACIS 14th Internat. Conf. Comput. Inform. Sci.* (Institute of Electrical and Electronics Engineers), 355–360.

Harrison GW, Martinez-Correa J, Swarthout JT (2014) Eliciting subjective probabilities with binary lotteries. *J. Econom. Behav. Organ.* 101(May):128–140.

Heath C, Tversky A (1991) Preference and belief: Ambiguity and competence in choice under uncertainty. *J. Risk Uncertainty* 4(1):5–28.

Hendrickson AD, Buehler RJ (1971) Proper scores for probability forecasters. *Ann. Math. Statist.* 1916–1921.

Hendry DF, Clements MP (2004) Pooling of forecasts. *Econometrics J.* 7(1):1–31.

Hirtle B, Lopez JA (1999) Supervisory information and the frequency of bank examinations. *Econom. Policy Rev.* 5(1).

Hossain T, Okui R (2013) The binarized scoring rule. *Rev. Econom. Stud.* 80(3):984–1001.

Howe J (2006) The rise of crowdsourcing. *Wired* 14(6):1–4.

Huck S, Weizsäcker G (2002) Do players correctly estimate what others do? Evidence of conservatism in beliefs. *J. Econom. Behav. Organ.* 47(1):71–85.

Hyndman K, Özbay EY, Schotter A, Ehrblatt W (2012a) Belief formation: An experiment with outside observers. *Experiment. Econom.* 15(1):176–203.

Hyndman K, Ozbay EY, Schotter A, Ehrblatt WZ (2012b) Convergence: An experimental study of teaching and learning in repeated games. *J. Eur. Econom. Assoc.* 10(3):573–604.

Iyer K, Johari R, Moallemi CC (2010) Information aggregation in smooth markets. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 199–206.

Jaun S, Ahrens B (2009) Evaluation of a probabilistic hydrometeorological forecast system. *Hydrology and Earth System Sci.* 13(7):1031–1043.

Jensen FA, Peterson CR (1973) Psychological effects of proper scoring rules. *Organ. Behav. Human Performance* 9(2):307–317.

Johnstone DJ (2012) Log-optimal economic evaluation of probability forecasts. *J. Roy. Statist. Soc. Ser. A* 175(3):661–689.

Johnstone DJ, Jose VRR, Winkler RL (2011) Tailored scoring rules for probabilities. *Decision Anal.* 8(4):256–268.

Johnstone DJ, Jones S, Jose VRR, Peat M (2013) Measures of the economic value of probabilities of bankruptcy. *J. Roy. Statist. Soc. Ser. A* 176(3):635–653.

Jose VRR (2009) A characterization for the spherical scoring rule. *Theory Decision* 66(3):263–281.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57(5):1287–1297.

Jose VRR, Nau RF, Winkler RL (2008) Scoring rules, generalized entropy, and utility maximization. *Oper. Res.* 56(5):1146–1157.

Jose VRR, Nau RF, Winkler RL (2009) Sensitivity to distance and baseline distributions in forecast evaluation. *Management Sci.* 55(4):582–590.

Jumadinova J, Dasgupta P (2013) Prediction market-based information aggregation for multi-sensor information processing. David E, Kiekintveld C, Robu V, Shehory O, Stein S, eds. *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*, Lecture Notes in Business Information Processing, Vol. 136 (Springer, Berlin), 75–89.

Jurca R, Faltings B (2009) Mechanisms for making crowds truthful. *J. Artificial Intelligence Res.* 34(1):209–253.

Kamar E, Horvitz E (2012) Incentives for truthful reporting in crowdsourcing. *Proc. 11th Internat. Conf. Autonomous Agents and Multiagent Systems*, 1329–1330.

Karni E (2009) A mechanism for eliciting probabilities. *Econometrica* 77(2):603–606.

Katz RW, Murphy AH (1997) *Economic Value of Weather and Climate Forecasts* (Cambridge University Press, Cambridge, UK).

Koessler F, Noussair C, Ziegelmeyer A (2012) Information aggregation and belief elicitation in experimental parimutuel betting markets. *J. Econom. Behav. Organ.* 83(2):195–208.

Kothiyal A, Spinu V, Wakker PP (2011) Comonotonic proper scoring rules to measure ambiguity and subjective beliefs. *J. Multi-Criteria Decision Anal.* 17(3–4):101–113.

Lad F, Sanfilippo G, Agrò G (2012) Completing the logarithmic scoring rule for assessing probability distributions. *AIP Conf. Proc.* 1490(March):13–30.

Lawrence S, Glover EJ, Giles CL (2002) Characterizing efficiency and information incorporation in sports betting markets. *Proc. 9th Res. Sympos. Emerging Electronic Markets*, 45–52.

Ledyard J, Hanson R, Ishikida T (2009) An experimental test of combinatorial information markets. *J. Econom. Behav. Organ.* 69(2):182–189.

Lerch S, Thorarinsdottir TL (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* 65, http://dx.doi.org/10.3402/tellusa.v65i0.21206.

Li X, Vaughan JW (2013) An axiomatic characterization of adaptive-liquidity market makers. *Proc. 14th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 657–674.

Linnet K (1988) A review on the methodology for assessing diagnostic tests. *Clinical Chemistry* 34(7):1379–1386.

Linnet K (1989) Assessing diagnostic tests by a strictly proper scoring rule. *Statist. Medicine* 8(5):609–618.

Lopez JA (2001) Evaluating the predictive accuracy of volatility models. *J. Forecasting* 20(2):87–109.

Machete RL (2013) Contrasting probabilistic scoring rules. *J. Statist. Planning Inference* 143(10):1781–1790.

Madigan D, Gavrin J, Raftery AE (1995) Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Comm. Statist. - Theory Methods* 24(9):2271–2292.

Mahmoud E (1984) Accuracy in forecasting: A survey. *J. Forecasting* 3(2):139–159.

Manski CF, Neri C (2013) First- and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games Econom. Behav.* 81(September):232–254.

Mason SJ (2004) On using "climatology" as a reference strategy in the brier and ranked probability skill scores. *Monthly Weather Rev.* 132(7):1891–1895.

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. *Management Sci.* 22(10):1087–1096.

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, Moore D, Atanasov P, Swift SA (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psych. Sci.* 25(5):1106–1115.

Merkle EC, Steyvers M (2013) Choosing a strictly proper scoring rule. *Decision Anal.* 10(4):292–304.

Mhanna S, Verbic G, Chapman AC (2014a) Guidelines for realistic grounding of mechanism design in demand response. *Proc. 2014 Australasian Universities Power Engrg. Conf.*, 1–6.

Mhanna S, Verbic G, Chapman AC (2014b) Towards a realistic implementation of mechanism design in demand response aggregation. *IEEE Power Systems Comput. Conf.*, 1–7.

Mhanna S, Verbic G, Chapman AC (2015) A faithful distributed mechanism for demand response aggregation. *IEEE Trans. Smart Grid* 7(3):1743–1753.

Miller N, Resnick P, Zeckhauser R (2005) Eliciting informative feedback: The peer-prediction method. *Management Sci.* 51(9):1359–1373.

Muradoglu G, Onkal D (1994) An exploratory analysis of portfolio managers' probabilistic forecasts of stock prices. *J. Forecasting* 13(7):565–578.

Murphy AH (1974) A sample skill score for probability forecasts. *Monthly Weather Rev.* 102(1):48–55.

Murphy AH (1985) Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Monthly Weather Rev.* 113(3):362–369.

Murphy AH (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* 8(2):281–293.

Murphy AH, Daan H (1984) Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Monthly Weather Rev.* 112(3):413–423.

Murphy AH, Winkler RL (1982) Subjective probabilistic tornado forecasts: Some experimental results. *Monthly Weather Rev.* 110(9):1288–1297.

Murphy AH, Winkler RL (1984) Probability forecasting in meteorology. *J. Amer. Statist. Assoc.* 79(387):489–500.

Murphy AH, Winkler RL (1992) Diagnostic verification of probability forecasts. *Internat. J. Forecasting* 7(4):435–455.

Murphy AH, Brown BG, Chen YS (1989) Diagnostic verification of temperature forecasts. *Weather and Forecasting* 4(4):485–501.

Murphy JM (1990) Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Royal Meteorological Soc.* 116(491):89–125.

Nakazono Y (2013) Strategic behavior of Federal Open Market Committee board members: Evidence from members' forecasts. *J. Econom. Behav. Organ.* 93(September):62–70.

Nelson RG, Bessler DA (1989) Subjective probabilities and scoring rules: Experimental evidence. *Amer. J. Agricultural Econom.* 71(2):363–369.

Nikolova E, Sami R (2007) A strategic model for information markets. *Proc. 8th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 316–325.

Nyarko Y, Schotter A (2002) An experimental study of belief learning using elicited beliefs. *Econometrica* 70(3):971–1005.

O'Carroll FM (1977) Subjective probabilities and short-term economic forecasts: An empirical investigation. *Appl. Statist.* 26(3):269–278.

Offerman T, Palley AB (2015) Lossed in translation: An off-the-shelf method to recover probabilistic beliefs from loss-averse agents. *Experiment. Econom.* 19(1):1–30.

Offerman T, Sonnemans J (2004) What's causing overreaction? An experimental investigation of recency and the hot-hand effect. *Scandinavian J. Econom.* 106(3):533–553.

Offerman T, Sonnemans J, Schram A (1996) Value orientations, expectations and voluntary contributions in public goods. *Econom. J.* 106(437):817–845.

Offerman T, Sonnemans J, Van De Kuilen G, Wakker PP (2009) A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Rev. Econom. Stud.* 76(4):1461–1489.

Oka M, Todo T, Sakurai Y, Yokoo M (2014) Predicting own action: Self-fulfilling prophecy induced by proper scoring rules. *Proc. 2nd AAAI Conf. Human Comput. Crowdsourcing*.

Ostrovsky M (2012) Information aggregation in dynamic markets with strategic traders. *Econometrica* 80(6):2595–2647.

Othman A, Sandholm T (2010) Decision rules and decision markets. *Proc. 9th Internat. Conf. Autonomous Agents and Multiagent Systems*, 625–632.

Othman A, Sandholm T (2011) Liquidity-sensitive automated market makers via homogeneous risk measures. Chen N, Elkind E, Koutsoupias E, eds. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 7090 (Springer-Verlag, New York), 314–325.

Othman A, Sandholm T (2013) The Gates Hillman prediction market. *Rev. Econom. Design* 17(2):95–128.

Othman A, Pennock DM, Reeves DM, Sandholm T (2013) A practical liquidity-sensitive automated market maker. *ACM Trans. Econom. Comput.* 1(3):14.

Palfrey TR, Wang SW (2009) On eliciting beliefs in strategic games. *J. Econom. Behav. Organ.* 71(2):98–109.

Parkes DC, Wellman MP (2015) Economic reasoning and artificial intelligence. *Science* 349(6245):267–272.

Pennock DM, Debnath S, Glover EJ, Giles CL (2002) Modeling information incorporation in markets, with application to detecting and explaining events. *Proc. 18th Conf. Uncertainty in Artificial Intelligence*, 405–413.

Petroliagis TI, Tambke J, Heinemann D, Denhard M, Hagedorn R (2010) How well can we forecast winds at different heights? An assessment of ECMWF IFS & EPS skill of forecasting wind fields at different model levels. *Proc. Eur. Wind Energy Assoc. Conf.*

Phillips LD, Edwards W (1966) Conservatism in a simple probability inference task. *J. Experiment. Psych.* 72(3):346–354.

Pinson P, Nielsen HA, Møller JK, Madsen H, Kariniotakis GN (2007) Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy* 10(6):497–516.

Prelec D (2004) A Bayesian truth serum for subjective data. *Science* 306(5695):462–466.

Radanovic G, Faltings B (2013) A robust Bayesian truth serum for non-binary signals. *Proc. Twenty-Seventh AAAI Conf. Artificial Intelligence*, 833–839.

Radanovic G, Faltings B (2015) Incentives for subjective evaluations with private beliefs. *Proc. 19th AAAI Conf. Artificial Intelligence*, 1014–1020.

Ray R, Vallam RD, Narahari Y (2013) Eliciting high quality feedback from crowdsourced tree networks using continuous scoring rules. *Proc. 12th Internat. Conf. Autonomous Agents and Multiagent Systems*, 279–286.

Robu V, Kota R, Chalkiadakis G, Rogers A, Jennings NR (2012) Cooperative virtual power plant formation using scoring rules. *Proc. 11th Internat. Conf. Autonomous Agents and Multiagent Systems*, 1165–1166.

Rose H, Rogers A, Gerding EH (2011) Mechanism design for aggregated demand prediction in the smart grid. *Proc. Workshops at the 25th AAAI Conf. Artificial Intelligence.*

Rose H, Rogers A, Gerding EH (2012) A scoring rule-based mechanism for aggregate demand prediction in the smart grid. *Proc. 11th Internat. Conf. Autonomous Agents and Multiagent Systems*, 661–668.

Roulston MS, Smith LA (2002) Evaluating probabilistic forecasts using information theory. *Monthly Weather Rev.* 130(6): 1653–1660.

Rutström EE, Wilcox NT (2009) Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games Econom. Behav.* 67(2):616–632.

Sakurai Y, Shinoda M, Oyama S, Yokoo M (2015) Flexible reward plans for crowdsourced tasks. Chen Q, Torroni P, Villata S, Hsu J, Omicini A, eds. *Proc. 18th Internat. Conf. Principles and Practice of Multi-Agent Systems* (Springer International Publishing, Switzerland), 400–415.

Sakurai Y, Okimoto T, Oka M, Shinoda M, Yokoo M (2013) Ability grouping of crowd workers via reward discrimination. *Proc. 1st AAAI Conf. Human Comput. Crowdsourcing*, 147–155.

Sanders F (1963) On subjective probability forecasting. *J. Appl. Meteorology* 2(2):191–201.

Sandroni A, Shmaya E (2013) Eliciting beliefs by paying in chance. *Econom. Theory Bull.* 1(1):33–37.

Savage LJ (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66(336):783–801.

Schervish MJ (1989) A general method for comparing probability assessors. *Ann. Statist.* 17(4):1856–1879.

Scheuerer M, Hamill TM (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Rev.* 143(4):1321–1334.

Schlag KH, van der Weele JJ (2013) Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Econom. Lett.* 3(1):38–42.

Schum DA, Goldstein IL, Howell WC, Southard JF (1967) Subjective probability revisions under several cost-payoff arrangements. *Organ. Behav. Human Performance* 2(1):84–104.

Selten R (1998) Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* 1(1):43–62.

Selten R, Sadrieh A, Abbink K (1999) Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory Decision* 46(3):211–249.

Shi P, Conitzer V, Guo M (2009) Prediction mechanisms that do not incentivize undesirable actions. Leonardi S, ed. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 5929 (Springer, Berlin), 89–100.

Slamka C, Skiera B, Spann M (2013) Prediction market performance and market liquidity: A comparison of automated market makers. *IEEE Trans. Engrg. Management* 60(1):169–185.

Smith LA, Suckling EB, Thompson EL, Maynard T, Du H (2015) Towards improving the framework for probabilistic forecast evaluation. *Climatic Change* 132(1):31–45.

Spiegelhalter DJ (1986) Probabilistic prediction in patient management and clinical trials. *Statist. Medicine* 5(5):421–433.

Spiegelhalter DJ, Franklin RCG, Bull K (1990) Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. *Proc. 5th Annual Conf. Uncertainty in Artificial Intelligence*, 285–294.

Štrumbelj E, Šikonja MR (2010) Online bookmakers' odds as forecasts: The case of European soccer leagues. *Internat. J. Forecasting* 26(3):482–488.

Surowiecki J (2005) *The Wisdom of Crowds* (Anchor, New York).

Tetlock P (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press, Princeton, NJ).

Tetlock PE, Kim JI (1987) Accountability and judgment processes in a personality prediction task. *J. Personality Soc. Psych.* 52(4): 700–709.

Thorarinsdottir TL, Gneiting T (2010) Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* 173(2): 371–388.

Thorarinsdottir TL, Gneiting T, Gissibl N (2013) Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantification* 1(1):522–534.

Toda M (1963) Measurement of subjective probability distributions. Technical Report ESD-TDR-63-407, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, Bedford, MA.

Trautmann ST, van de Kuilen G (2014) Belief elicitation: A horse race among truth serums. *Econom. J.* 125(589):2116–2135.

Ugander J, Drapeau R, Guestrin C (2015) The wisdom of multiple guesses. *Proc. 16th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 643–660.

van Lenthe J (1994) Scoring-rule feedforward and the elicitation of subjective probability distributions. *Organ. Behav. Human Decision Processes* 59(2):188–209.

Vardi MY (2010) Revisiting the publication culture in computing research. *Comm. ACM* 53(3):5.

von Holstein CAS (1971a) An experiment in probabilistic weather forecasting. *J. Appl. Meteorology* 10(4):635–645.

von Holstein CAS (1971b) The effect of learning on the assessment of subjective probability distributions. *Organ. Behav. Human Performance* 6(3):304–315.

von Holstein CAS (1971c) Two techniques for assessment of subjective probability distributions—An experimental study. *Acta Psychologica* 35(6):478–494.

von Holstein CAS (1972) Probabilistic forecasting: An experiment related to the stock market. *Organ. Behav. Human Performance* 8(1):139–158.

Walker KD, Catalano P, Hammitt JK, Evans JS (2003) Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene. *J. Exposure Sci. Environ. Epidemiology* 13(1):1–16.

Wang SW (2011) Incentive effects: The case of belief elicitation from individuals in groups. *Econom. Lett.* 111(1):30–33.

Weijs SV, Schoups G, van De Giesen N (2010a) Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sci.* 14:2545–2558.

Weijs SV, van Nooijen R, van de Giesen N (2010b) Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Rev.* 138(9):3387–3399.

Wilks DS (2011) *Statistical Methods in the Atmospheric Sciences* (Academic Press, Cambridge, MA).

Wilson LJ, Burrows WR, Lanzinger A (1999) A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Rev.* 127(6):956–970.

Winkler RL (1969) Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* 64(327):1073–1078.

Winkler RL (1971) Probabilistic prediction: Some experimental results. *J. Amer. Statist. Assoc.* 66(336):675–685.

Winkler RL (1972) A decision-theoretic approach to interval estimation. *J. Amer. Statist. Assoc.* 67(337):187–191.

Winkler RL (1994) Evaluating probabilities: Asymmetric scoring rules. *Management Sci.* 40(11):1395–1405.

Winkler RL, Murphy AH (1969) "Good" probability assessors. *J. Appl. Meteorology* 7(5):751–758.

Winkler RL, Murphy AH (1970) Nonlinear utility and the probability score. *J. Appl. Meteorology* 9(1):143–148.

Winkler RL, Poses RM (1993) Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Sci.* 39(12):1526–1543.

Witkowski J, Parkes DC (2012) A robust Bayesian truth serum for small populations. *Proc. 26th AAAI Conf. Artificial Intelligence*.

Woo CK, Horowitz I, Martin J (1998) Reliability differentiation of electricity transmission. *J. Regulatory Econom.* 13(3):277–292.

Yates JF, McDaniel LS, Brown ES (1991) Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organ. Behav. Human Decision Processes* 49(1):60–79.

Zhang H, Horvitz E, Chen Y, Parkes DC (2012) Task routing for prediction tasks. *Proc. 11th Internat. Conf. Autonomous Agents and Multiagent Systems*, 889–896.

**Arthur Carvalho** is an assistant professor at the Farmer School of Business, Miami University. The main focus of his research is on how to leverage the wisdom of crowds to assist decision makers in making better decisions. Some of the research questions he has been working on include (1) how to elicit and aggregate crowd members' subjective and private information in an optimal way; and (2) from a decision-making perspective, how to optimally build a crowd by taking into account diversity of information and crowd members' skills. Dr. Carvalho received his MMath and Ph.D. in computer science from the University of Waterloo (Canada) in 2010 and 2014, respectively.