DECISION ANALYSIS
Vol. 4, No. 2, June 2007, pp. 49–65
ISSN 1545-8490 | EISSN 1545-8504 | 07 | 0402 | 0049

DOI 10.1287/deca.1070.0089

Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules

J. Eric Bickel
Department of Industrial and Systems Engineering, Texas A&M University, 236-B Zachry Engineering Center,
3131 TAMU, College Station, Texas 77843-3131, ebickel@tamu.edu

Strictly proper scoring rules continue to play an important role in probability assessment. Although many such rules have been developed, relatively little guidance exists as to which rule is the most appropriate. In this paper, we discuss two important properties of quadratic, spherical, and logarithmic scoring rules. From an ex post perspective, we compare their rank order properties and conclude that both quadratic and spherical scoring perform poorly in this regard, relative to logarithmic. Second, from an ex ante perspective, we demonstrate that in many situations, logarithmic scoring is the method least affected by a nonlinear utility function. These results suggest that logarithmic scoring is superior when rank order results are important and/or when the assessor has a nonlinear utility function. In addition to these results, and perhaps more important, we demonstrate that nonlinear utility induces relatively little deviation from the optimal assessment under an assumption of risk neutrality. These results provide both comfort and guidance to those who would like to use scoring rules as part of the assessment process.

Key words: scoring rules; Brier score; scoring rule comparison; probability assessment; probability forecasts; truthful revelation
History: Received on January 26, 2007. Accepted by L. Robin Keller on May 14, 2007, after 1 revision.

1. Introduction

Consider the assessment of a probability distribution by an individual A, the assessor, over n mutually exclusive and collectively exhaustive statements, where $n > 1$. Let $\mathbf{p} = (p_1, \ldots, p_n)$ be an n-vector of probabilities representing A's private beliefs, where p_i is the probability A assigns that statement i is true or will occur, and the sum of these probabilities is equal to 1. Let A's public assessment or response be given by $\mathbf{r} = (r_1, \ldots, r_n)$, where r_i is the stated probability that statement i is correct, and the sum of these probabilities is equal to 1.

If A is rewarded or scored according to some function R, then A's expected score when he or she assigns $\mathbf{r}$ and believes $\mathbf{p}$ is $\bar{R}(\mathbf{r} \mid \mathbf{p}) = E_{\mathbf{p}}[R_i(\mathbf{r})] = \sum_i p_i R_i(\mathbf{r})$, where E is the expectation operator and R_i is the score received for assigning $\mathbf{r}$ when statement i is correct. A strictly proper scoring rule T is a scoring function such that A strictly maximizes his or her expected score by setting $\mathbf{r} = \mathbf{r}^* = \mathbf{p}$; that is, $\bar{T}(\mathbf{p} \mid \mathbf{p}) > \bar{T}(\mathbf{r} \mid \mathbf{p})$ for all $\mathbf{r} \neq \mathbf{p}$ and $\bar{T}(\mathbf{p} \mid \mathbf{p}) = \bar{T}(\mathbf{r}^* \mid \mathbf{p})$ when $\mathbf{r}^* = \mathbf{p}$ (Toda 1963, Roby 1965, Shuford et al. 1966, Winkler 1968). Many scoring rules have been developed. Three of the most popular are

$$\text{Quadratic (Q):} \quad Q_i(\mathbf{r}) = 2r_i - \mathbf{r} \cdot \mathbf{r} \in [-1, 1] \tag{1}$$

$$\text{Spherical (S):} \quad S_i(\mathbf{r}) = r_i / (\mathbf{r} \cdot \mathbf{r})^{1/2} \in [0, 1] \tag{2}$$

$$\text{Logarithmic (L):} \quad L_i(\mathbf{r}) = \ln(r_i) \in (-\infty, 0], \tag{3}$$

where, again, the subscript i signifies that this is the score obtained by assigning $\mathbf{r}$ when statement i is correct.

In some cases it is convenient to scale or normalize the scoring rules. Toda (1963) proved that a linear transformation of a strictly proper scoring rule is also strictly proper. For example, the Brier score (Brier 1950), which is popular in meteorological contexts, is $B_i(\mathbf{r}) = 1 - Q_i(\mathbf{r})$.

Given the variety of scoring rules, the question naturally arises as to which scoring rule should be used. Scoring rules have ex ante and ex post properties (Winkler 1996). Ex ante properties are related to the scoring rule's ability to encourage $\mathbf{r}^* = \mathbf{p}$, while ex post properties are concerned with the scoring rule's usefulness in evaluating assessment performance.

Shuford et al. (1966) proved that logarithmic is the only proper scoring rule the value of which depends only on the probability assigned to the correct statement, when there are more than two statements. This is referred to as the *local* property and has two important implications. First, such a rule should be easier for individuals to understand. For example, a two-dimensional chart can be provided that details the score for any set of n assignments, which is only possible for other rules in special circumstances (e.g., A assigns r_i to the correct answer and $(1-r_i)(n-1)^{-1}$ to each of the remaining statements). Second, and perhaps more important if one is using the output of a scoring rule to evaluate the ability of the assessor, logarithmic is the only scoring rule consistent with the likelihood principle or the use of Bayes factors to update the weights assigned to different experts or forecasting systems (Winkler 1969, 1996).

However, Friedman (1979, 1983) argued that not only should scoring rules induce honest assessments, but they should encourage $\mathbf{r}$ to be close to $\mathbf{p}$ in terms of a distance metric, a property Friedman terms *effective*. Friedman proves that both quadratic and spherical are effective and conjectures that logarithmic is not effective for any metric. Nau (1985) countered that effectiveness only adds a transitivity property among probability distributions, which is difficult to justify. Selten (1998) defined the expected score loss by assigning $\mathbf{r}$ instead of $\mathbf{p}$ to be $\overline{T}^{-}(\mathbf{r} \mid \mathbf{p}) = \overline{T}(\mathbf{p} \mid \mathbf{p}) - \overline{T}(\mathbf{r} \mid \mathbf{p})$ and introduces an additional property he calls *neutrality*, which requires that the expected score loss satisfy the symmetry property of a distance measure such that $\overline{T}^{-}(\mathbf{p} \mid \mathbf{r}) = \overline{T}^{-}(\mathbf{r} \mid \mathbf{p})$. Selten proved that Q is the only scoring rule that satisfies neutrality.[1] Earlier, Savage (1971) had shown that Q is the only rule scoring rule with a symmetric loss function in the case of a simple dichotomy.

Because L is local, it will always assign a higher score to higher assignments on the correct statement. Q and S do not share this property when there are more than two statements. One implication of this feature is that one assessor may assign a *higher* (*lower*) probability than another assessor to the correct statement but receive a *lower* (*higher*) score. Whether this is a concern depends on the context. In academic testing situations, for example, students are likely to perceive such a result as being unfair. Another implication of nonlocality is that, when there are more than two statements, different scoring rules may generate different rank orderings among assessors for the same set of assessments.

A few authors have studied the rank order properties of the scoring rules. In an assessment study involving American football, Winkler (1971) reported that all three scoring rules yielded similar rankings when averaged over several assessment tasks. Staël von Holstein (1970) performed two different assessment experiments, with between 5 and 9 statements, and reported that Q, S, and L produced similar rankings when considering average scores. However, close inspection of his results shows that some individuals' rank changed by up to 50% by being scored with Q or S instead of L.

The proper scoring rules discussed above assume that A's objective is to maximize his or her expected score, which implicitly assumes the utility function is linear in his or her score. If this is not the case, it may no longer be optimal to set $\mathbf{r}^* = \mathbf{p}$. If A's utility function u is known and has an inverse u^{-1}, then $u^{-1} \circ T$ is a proper scoring rule under u (Winkler 1969).

The difficulty in practice is that neither the functional form nor the parameters of u may be known. This suggests that we should study the effect of nonlinear utility functions on $\mathbf{r}^*$ under Q, S, and L scoring. Perhaps under certain circumstances this impact is "small." Winkler and Murphy (1970) illustrated the effect of quadratic and exponential utility under Q scoring with two statements and found that a risk-preferring individual moves closer to deterministic assessments (i.e., closer to 0 or 1), and a risk-averse individual hedges his or her assessments closer to uniform. Murphy and Winkler (1971) calculated the optimal adjustment under quadratic utility and Q scoring when $n = 2$ and $p_1 = 0.9$ and found $p_1 - r_1^* = 0.2593$, which is a significant deviation from the risk-neutral solution.

[1] Though not specifically discussed by Selten, L does not satisfy neutrality because the expected score loss in this case is equal to the Kullback-Leibler distance between $\mathbf{r}$ and $\mathbf{p}$. It is well known that KL is not a true distance measure because it is not symmetric and does not satisfy the triangle inequality (Cover and Thomas 1991).

In this paper, we investigate two properties of scoring rules, one ex post and one ex ante, which we believe are important in certain situations. First, we analytically compare the ex post rank order properties of Q, S, and L, which provides context for considering the empirical studies discussed above. Our results demonstrate that the rank order differences among Q, S, and L can be significant. Second, we quantify the ex ante impact of nonlinear utility functions under many different conditions. We study the absolute impact on $\mathbf{r}^*$ as well as the relative performance of Q, S, and L. This analysis yields two important findings: (1) In many situations, the deviation from the risk-neutral solution is minor, and (2) L is the least affected by nonlinear utility, under an assumption of exponential utility.

The remainder of this paper is organized as follows. Section 2 discusses the normalization of the scoring rules so that we can properly compare them. Section 3 investigates the rank order properties of the scoring rules. Section 4 analyzes the performance of Q, S, and L under nonlinear utility. Finally, §5 concludes.

2. Normalization

To compare the scoring rules, which have different ranges, we must normalize them in some cases. For example, the optimal response for different scoring rules under an assumption of nonlinear utility will depend on the range of possible scoring outcomes and hence the normalization scheme. Unfortunately, normalizing the rules is not straightforward, because L is unbounded below. We adopt the perspective of an analyst who is considering the use of Q, S, or L scoring for a particular assessment task and suggest a normalization scheme defined over the range $p_i \in (0, 1]$ that we believe is both natural and reasonable. We will use the script characters $\mathcal{T}$, $\mathcal{Q}$, $\mathcal{S}$, and $\mathcal{L}$ to denote the *normalized* rules and the nonscript characters T, Q, S, and L to denote the *core* scoring rule. We will refer to the normalized rules only in those cases where the results depend on the normalization scheme.

Two normalization approaches are possible: ex ante or ex post. An ex ante normalization ensures that the pairs $(\mathbf{r}_\alpha, \mathbf{p}_\alpha)$ and $(\mathbf{r}_\beta, \mathbf{p}_\beta)$ yield expected scores of α and β, respectively, such that $\overline{\mathcal{T}}(\mathbf{r}_\alpha \mid \mathbf{p}_\alpha) = a + b\overline{T}(\mathbf{r}_\alpha \mid \mathbf{p}_\alpha) = \alpha$ and $\overline{\mathcal{T}}(\mathbf{r}_\beta \mid \mathbf{p}_\beta) = a + b\overline{T}(\mathbf{r}_\beta \mid \mathbf{p}_\beta) = \beta$ (Murphy and Winkler 1970). An ex post normalization ensures that an assignment of $\mathbf{r}_\alpha$ yields a score of α when statement j is correct, and assignment of $\mathbf{r}_\beta$ yields a score of β when statement k is correct, noting that j may equal k. For example, consider a statement with three possible answers: (j, k, l). We may wish to award the assessor a score of α if he or she assigns $(0.1, 0.8, 0.1)$ and statement j is correct and a score of β when he or she assigns $(0.4, 0.5, 0.1)$ and statement k is correct. Formally, we are free to choose a and b such that $\mathcal{T}_j(\mathbf{r}_\alpha) = a + bT_j(\mathbf{r}_\alpha) = \alpha$ and $\mathcal{T}_k(\mathbf{r}_\beta) = a + bT_k(\mathbf{r}_\beta) = \beta$. We focus on ex post normalization because we believe it enables assessors to better understand the implications of their assignments because a given score is a function only of the response. The ex post normalization factors $a_{\mathcal{T}}$ and $b_{\mathcal{T}}$ for the scoring rule T are

$$a_{\mathcal{T}} = \frac{\alpha T_k(\mathbf{r}_\beta) - \beta T_j(\mathbf{r}_\alpha)}{T_k(\mathbf{r}_\beta) - T_j(\mathbf{r}_\alpha)} \tag{4}$$

$$b_{\mathcal{T}} = \frac{\beta - \alpha}{T_k(\mathbf{r}_\beta) - T_j(\mathbf{r}_\alpha)} \tag{5}$$

(see appendix). As can be seen from Equation (5), the ratio between bs for any two scoring rules is independent of α and β. This will become important in our subsequent analysis.

Suppose an analyst is considering the use of Q, S, or L scoring in a particular assessment situation. In the context of ex post normalization, it is natural to assign the assessor the maximum number of points, α, when he or she assigns 1 to the correct statement. Likewise, assigning a score of β when the assessor assigns a uniform distribution is also appealing. In this case, the normalization constants given by Equations (4) and (5) ensure that the scoring rules assign the same score to perfect knowledge or ignorance. Normalization of the scoring rules is difficult and many approaches are possible, but we believe the normalization discussed here is quite reasonable. In fact, any normalization of the scoring rules that does not assign identical scores to perfect knowledge or ignorance seems difficult to justify.

Following this normalization, we let $\mathbf{r}_\alpha = (r_1, \ldots, r_n)$, where $r_i = 1$ if statement i is true and 0 otherwise, and $\mathbf{r}_\beta = (n^{-1}, \ldots, n^{-1})$. We will generally take $\beta = 0$, which has the benefit of assuring the assessor a score

Table 1 Natural Normalization Factors ($\beta = 0$)

Scoring rule	$a_{\mathcal{T}}$	$b_{\mathcal{T}}$
Quadratic (Q)	$\frac{-1}{n-1}\alpha$	$\frac{n}{n-1}\alpha$
Spherical (S)	$\frac{-1}{\sqrt{n}-1}\alpha$	$\frac{\sqrt{n}}{\sqrt{n}-1}\alpha$
Logarithmic (L)	α	$\frac{\alpha}{\ln n}$

of 0 for a statement of ignorance. However, our results are unchanged for different values of β. With these assumptions, a and b for each scoring rule are shown in Table 1.

Figure 1 displays the scoring rules normalized according to Table 1 for $n = 2$. Notice that the rules yield identical scores when $r_i = 1$ and $r_i = n^{-1} = 0.5$.

3. Rank Order Properties

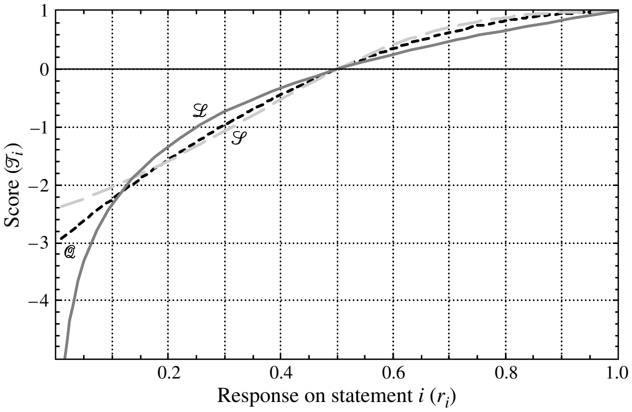
A positive linear transformation of a scoring rule will not change the rank ordering of the scores for a particular set of assessments. Therefore, the rank order properties between scoring rules are independent of the linear normalization scheme. However, to facilitate comparisons, we will use the normalization scheme defined by the parameters in Table 1 with $\alpha = 1$.

Before proceeding, we define the following vectors to streamline our notation:

- *ith unit n vector*: $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{in})$, where

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

Figure 1 Normalized Scoring Rules (Two Statements)



- *ith zero n vector*: $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{in})$, where

$$\zeta_{ij} = \begin{cases} 0 & \text{for } i = j \\ 1 & \text{for } i \neq j \end{cases}$$

- *uniform n vector*: $\mathbf{u}_n = n^{-1}\mathbf{1} = (n^{-1}, \ldots, n^{-1})$.

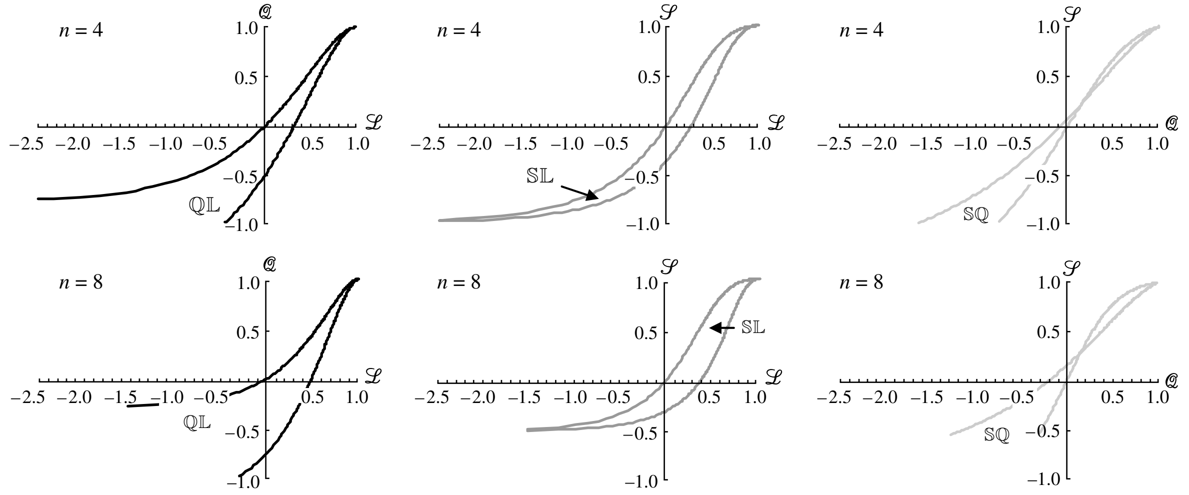
$\boldsymbol{\delta}_i$ is a vector of 0s with 1 in the ith spot. $\boldsymbol{\zeta}_i$ is a vector of 1s with a 0 in the ith spot.

Assume that statement i is correct and that A has assigned $r_i = \hat{r}$, thereby earning the normalized logarithmic score $\mathcal{L}_i(\hat{r})$. Because quadratic and spherical scoring are not local, there are a range of $\mathcal{Q}_i$ and $\mathcal{S}_i$ values that satisfy $\mathcal{L}_i(\hat{r})$. $\mathcal{Q}_i$ and $\mathcal{S}_i$ will achieve maxima (minima) when $\mathbf{r} \cdot \mathbf{r}$ achieves a minimum (maximum) for a given r_i (see Equations (1) and (2)). Because $\mathbf{r} \cdot \mathbf{r}$ achieves a minimum equal to n^{-1} when $\mathbf{r} = \mathbf{u}_n$ and a maximum equal to 1 when $\mathbf{r} = \boldsymbol{\delta}_i$, $\mathcal{Q}_i$ and $\mathcal{S}_i$ will achieve minima when $\mathbf{r} = \mathbf{r}^- = r_i\boldsymbol{\delta}_i + (1 - r_i)\boldsymbol{\delta}_j$ and will achieve maxima when $\mathbf{r} = \mathbf{r}^+ = r_i\boldsymbol{\delta}_i + (1 - r_i)(n - 1)^{-1}\boldsymbol{\zeta}_i$. The minimum values are then $\mathcal{Q}_i^- = \mathcal{Q}_i(\mathbf{r}^-)$ and $\mathcal{S}_i^- = \mathcal{S}_i(\mathbf{r}^-)$ with $r_i^- = \hat{r}$, where r_i^- is the r_i associated with $\mathbf{r}^-$. The maximum values are $\mathcal{Q}_i^+ = \mathcal{Q}_i(\mathbf{r}^+)$ and $\mathcal{S}_i^+ = \mathcal{S}_i(\mathbf{r}^+)$ with $r_i^+ = \hat{r}$, where r_i^+ is the r_i associated with $\mathbf{r}^+$. The set of points satisfying these criteria in the case of $\mathcal{Q}$-$\mathcal{L}$ is $\mathbb{QL} = \{(l, q)\colon l = \mathcal{L}_i(r_i), q \in [\mathcal{Q}_i^-, \mathcal{Q}_i^+], r_i \in [0, 1]\}$. Similarly for $\mathcal{S}$-$\mathcal{L}$ and $\mathcal{S}$-$\mathcal{Q}$ we have $\mathbb{SL} = \{(l, s)\colon l = \mathcal{L}_i(r_i), s \in [\mathcal{S}_i^-, \mathcal{S}_i^+], r_i \in [0, 1]\}$ and $\mathbb{SQ} = \{(q, s)\colon q \in [\mathcal{Q}_i^-, \mathcal{Q}_i^+], s \in [\mathcal{S}_i^-, \mathcal{S}_i^+], r_i \in [0, 1]\}$.

When $n = 2$, then $\mathbf{r}^- = \mathbf{r}^+$ and all three scoring rules will produce the same rank ordering. When there are more than two statements, the rank ordering among the rules will no longer be perfect. Figure 2 displays the relationships between the rules when $n = 4$ and $n = 8$.

As can be seen in Figure 2, the relationship among the scoring rules is not perfect and degrades with increasing n. $\mathcal{S}$ and $\mathcal{Q}$ are the most closely related because they are both functions of $\mathbf{r} \cdot \mathbf{r}$, and $\mathcal{L}$ is only a function of r_i. In fact, there are always two responses r_i^- and r_i^+ such that $\mathcal{Q}_i(\mathbf{r}^-) = \mathcal{Q}_i(\mathbf{r}^+)$ and $\mathcal{S}_i(\mathbf{r}^-) = \mathcal{S}_i(\mathbf{r}^+)$, which are the crossover points shown in the $\mathcal{S}$-$\mathcal{Q}$ plots. When $n = 4$ and $\mathcal{L} > 0$, $\mathcal{Q}$ and $\mathcal{L}$ tend to be more closely related than $\mathcal{S}$ and $\mathcal{L}$. However, the relationship between $\mathcal{Q}$ and $\mathcal{L}$ degrades quickly as $\mathcal{L}$ decreases, while that of $\mathcal{S}$ and $\mathcal{L}$ improves. When $n = 8$, $\mathcal{Q}$ and $\mathcal{L}$ are less closely related than $\mathcal{S}$ and $\mathcal{L}$ for $\mathcal{L} > 0$.

Figure 2 Relationships between Scoring Rules



3.1. Rank Correlations

Although the rank correlations among the scoring rules will not be perfect for more than two statements, the magnitude of these correlations is an empirical matter. However, to get a sense for what might be expected in practice, we simulate an assessment situation by sampling from the sets displayed in Figure 2. Each sampled point represents the response of a single assessor. We sample N points, which yields a rank correlation among the N assessors. We repeat this sampling 500 times in an effort to quantify the range of rank correlations one might observe in practice.

The difficultly is choosing a sampling procedure. One approach is to uniformly simulate r_i to determine $\mathscr{L}_i(r_i)$ and then uniformly sample from the $\mathscr{Q}$ and $\mathscr{S}$ bounds corresponding to $\mathscr{L}_i(r_i)$. This implies a uniform sampling of r_i but a nonuniform sampling from the remaining responses because of the nonlinearity of $\mathscr{Q}$ and $\mathscr{S}$. Another approach is to uniformly simulate r_i to determine $\mathscr{L}_i(r_i)$ and then uniformly sample from the remaining possible responses, ensuring that the sum of the responses does not exceed 1. The difficulty in this case is that as n increases, the remaining responses become nearly identical because we must normalize their sum to equal $1 - r_i$, which will bias our simulations toward the upper boundaries in Figure 2. Therefore, to estimate the rank correlations one might encounter in practice, we generated a set of simulated data by sampling from $\mathbb{QL}$, $\mathbb{SL}$, and $\mathbb{SQ}$ according to the following procedure:

$\mathbb{QL}$ *and* $\mathbb{SL}$ *Sampling*: Draw a sample l uniformly from $\mathscr{L}_i \in [0, 1]$, which implies $r_i = \mathscr{L}_i^{-1}(l)$. We believe the lower limit of 0 is reasonable because individuals are likely to assign an r_i greater than n^{-1} to the correct statement in many situations. Given r_i, we then uniformly sample $q \in [\mathscr{Q}_i^-, \mathscr{Q}_i^+]$ and $s \in [\mathscr{S}_i^-, \mathscr{S}_i^+]$.

$\mathbb{SQ}$ *Sampling*: Simulating $\mathbb{SQ}$ is more difficult. To ensure that we sample from the same $\mathscr{Q}$ range as above, we cannot uniformly sample q from $\mathscr{Q}_i \in [0, 1]$ because $\mathscr{Q}_i = 0$ is the maximum $\mathscr{Q}_i$ when $\mathscr{L}_i = 0$. Rather, we first determine $\mathscr{Q}_i^-$ when $r_i^- = n^{-1}$, which is equal to $a_Q + b_Q(2n^{-1} - n^{-2}(2 - 2n + n^2))$. For example, when $n = 4$, $\mathscr{Q}_i^- = -0.5$ (see Figure 2). We then uniformly sample q from $\mathscr{Q}_i \in [\mathscr{Q}_i^-, 1]$. Next, we find the $\mathbf{r}^-$ that satisfies $\mathscr{Q}_i^- = q$ and the $\mathbf{r}^+$ that satisfies $\mathscr{Q}_i^+ = q$. Finally, we sample uniformly from $s \in [\mathscr{S}_i^-, \mathscr{S}_i^+]$.

Table 2 displays the mean rank correlations of the samples for 2–8 statements and 10, 100, and 200 assessors. The mean rank correlations are quite high, being greater than 0.920 in all cases. Rank correlations decrease with the number of statements and increase with the number of assessors, which is understandable, given the behavior depicted in Figure 2.[2] As one might expect, based on Figure 2, S-Q generates the largest rank correlations. Q-L tends to produce higher

[2] The standard deviations of these estimates are available as an online supplement on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html).

Table 2 Simulated Rank Correlations (500 Simulations)

	$\mathscr{Q}$ vs. $\mathscr{L}$			$\mathscr{S}$ vs. $\mathscr{L}$			$\mathscr{L}$ vs. $\mathscr{Q}$		
Statements	Number of assessors (N) =			Number of assessors (N) =			Number of assessors (N) =		
n	10	100	200	10	100	200	10	100	200
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4	0.963	0.982	0.983	0.959	0.979	0.980	0.987	0.996	0.997
6	0.940	0.966	0.968	0.938	0.966	0.967	0.975	0.989	0.991
8	0.927	0.954	0.956	0.926	0.958	0.959	0.962	0.984	0.984

rank correlations than S-L except for large n. This is undoubtedly a function of our simulation procedure, which does not include logarithmic scores below 0. We next compare these results to the empirical studies discussed in §1.

Staël von Holstein (1970) tested the performance of the quadratic scoring rule in an experiment involving stock price forecasting with five statements ($n = 5$) and 72 participants ($N = 72$). Like Winker, Staël von Holstein compared the quadratic scores to logarithmic and spherical scores and reports that the rank correlations in 10 different assessment sessions varied from 0.768 to 0.938 for Q-L and from 0.949 to 0.988 for S-Q. He did not report the correlation for S-L. When scores were averaged over the 10 sessions the rank correlations were 0.864 for Q-L and 0.855 for S-Q.

In a second experiment, 30 participants provided weather forecasts (n = 3–8) and received quadratic scoring rule feedback. Staël von Holstein (1970) found the following correlations: Q-L, 0.876; S-L, 0.830; S-Q, 0.982.[3]

We add to these empirical studies the results of several academic testing situations. The midterm exam in the introductory graduate decision analysis course at Stanford University is a 15-question multiple-choice test. Students assign a probability to each of four possible answers ($n = 4$) and are scored according to the logarithmic rule. Figure 3 compares the scores that would have been obtained for each of the 15 questions under Q, S, and L scoring for one particular year of Stanford data. In this case the normalization constants α and β were equal to $100/15$ and 0, respectively. As can be seen in Figure 3, the theoretical bounds between the scoring rules are achieved in practice. For this single year the rank correlations averaged over the 15 questions were 0.983 for Q-L, 0.991 for S-L, and 0.993 for S-Q. Based on five years of Stanford test results involving 1,030 students, we find the following rank correlation ranges for individual questions: Q-L, 0.906–0.999; S-L, 0.974–0.999; S-Q, 0.948–1.000. When the students' 15 individual question scores are aggregated into a total exam score, we find overall rank correlations of Q-L, 0.991; S-L, 0.992; S-Q, 0.997.[4]

The combination of our simulation results and the empirical studies demonstrates that one should expect high rank correlations in actual assessment situations.

3.2. Rank Differences

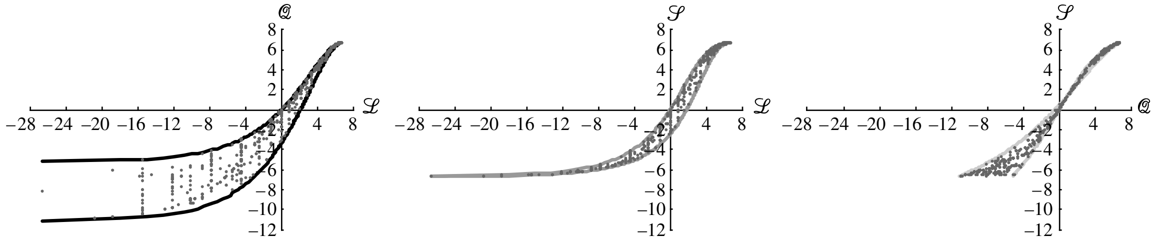
While the rank correlations are high, it is important to bear in mind that they average over many samples. It is possible for the rank order of a particular subset of assessments to be quite different even though the overall rank correlations are high. For example, based on the rank correlations in his second experiment, Staël von Holstein (1970) concludes that the different scoring rules yield essentially the same rankings. However, close inspection of his Table 11.5 shows that one individual's ranking increased by 11 of 30 (37%) by using Q instead of L and by 15 spots (50%) by using S instead of L. Another individual lost 9 spots (30%) by being ranked with either Q or S instead of L.

We computed the 10th and 90th percentile ranking differences for our Stanford data set. Over five years, on a individual question basis, 10% of the students would have lost an average of 5.1% in rank

[3] Staël von Holstein reports that rank correlations for S-L and S-Q were 0.818 and 0.959, respectively. However, direct calculation using the data he reports in Table 11.5 of his paper yields the values reported here. Staël von Holstein does not report Q-L, but it can be calculated from the data he provides.

[4] Please see the online supplement on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html) for a detailed breakdown of performance by year.

Figure 3 Comparison of Scoring Rules for Stanford Data



by being scored with Q rather than L, 5.3% by being scored with S rather than L, and 1.3% by being scored with S rather than Q. On the other hand, 10% of the students would have gained 1.0% by being scored with Q rather than L, 1.2% by being scored with S rather than L, and 1.1% by being scored with S rather than Q. The gains on an individual question basis are smaller because, as can be seen in Figure 3, scores tend to cluster at the upper boundary of the scoring ranges. In addition to these averages, the performance on particular questions can be quite poor. For example, on one particular question 10% of the students would have lost at least 21% in rank by being scored with Q instead of L.[5] These differences suggest that the ranking similarity among the scoring rules is not as strong as might be assumed from the rank correlations alone.

To investigate this behavior over a range of situations, we determined the 10th and 90th percentile ranking differences for each set of scoring rules for our simulated data. The results appear in Table 3, and, again, the differences are quite large. For example, when $n = 4$ and $N = 200$, 10% of the assessors would have lost at least 6.9% by being ranked by Q instead of L or at least 7.5% by being ranked by S instead of L. Conversely, when $n = 4$ and $N = 200$, 10% of assessors would have gained at least 6.7% if they were ranked by Q instead of L and at least 7.7% if ranked S instead of L. This ranking performance worsens with increasing n. The effect of changes in the number of assessors is not as clear; increasing N reduces percentage losses in rank, but increases percentage gains in rank. S-Q generally results in smaller changes in rank.[6]

The ranges in Table 3 are consistent with and help to explain the change in rankings observed in Staël von Holstein (1970) and our Stanford data set; such differences are to be expected.

3.3. Score Equality

The fact that Q and S are not local scoring rules introduces the possibility of two assessors assigning the same probability to the correct statement but receiving different scores, or receiving the same score for the different assignments. Figure 4 displays the range of possible scores for quadratic and spherical scoring. These ranges can be quite wide. For example, when $n = 4$ the minimum $\mathscr{Q}$ score when the assessor assigns 0.6 to the correct answer is 0.57 and the maximum is 0.72, over a 25% increase for the same r_i. Likewise, an assessor could also achieve a score of 0.57 with $r_i = 0.51$, a 15% decrease compared to 0.6. The ranges are even wider when $n = 8$. For example, in the case of spherical scoring and $n = 8$, an assessor could achieve a score of 0.74 with r_i assignments between 0.36 and 0.60.

An analysis of the Stanford data set demonstrates that this impact can be material. For example, if $\mathscr{Q}$ or $\mathscr{S}$ scoring had been used instead of $\mathscr{L}$, the maximum response difference to achieve the same or a higher score would have averaged 0.06 and 0.10, respectively. These differences varied by question and the maximum difference on one particular question was 0.14 under $\mathscr{Q}$ scoring and 0.21 under $\mathscr{S}$ scoring. Spherical scoring tended to produce the largest and most extreme differences, which is understandable given

[5] Please see the online supplement on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html) for a detailed breakdown of performance by year.

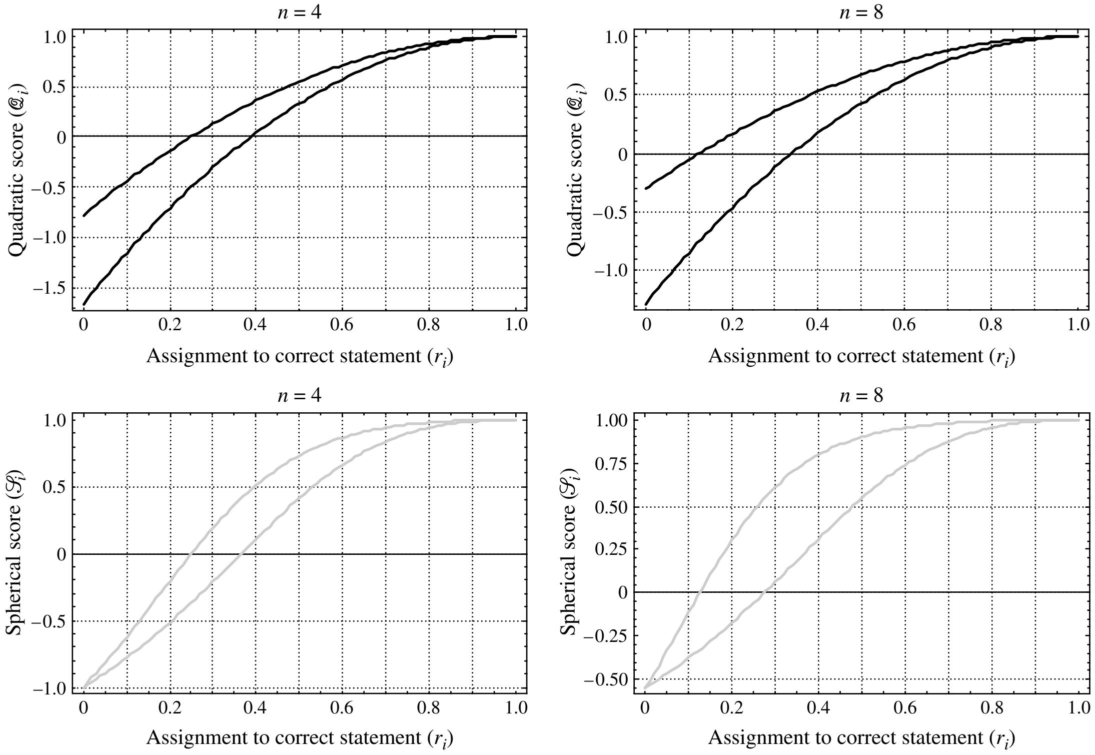
[6] The standard deviations of these estimates are available in the online supplement on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html).

Table 3 Simulated Ranking Difference (500 Simulations)

Statements	$\mathcal{Q}$ vs. $\mathcal{L}$ Number of assessors =			$\mathcal{S}$ vs. $\mathcal{L}$ Number of assessors =			$\mathcal{S}$ vs. $\mathcal{Q}$ Number of assessors =		
n	10 (%)	100 (%)	200 (%)	10 (%)	100 (%)	200 (%)	10 (%)	100 (%)	200 (%)
				Loss (10th percentile)					
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	−12.4	−7.1	−6.9	−13.9	−7.8	−7.5	−6.5	−2.9	−2.6
6	−14.9	−9.8	−9.4	−16.2	−10.1	−9.8	−9.7	−4.9	−4.4
8	−17.2	−11.3	−11.1	−17.5	−11.4	−11.1	−12.8	−6.1	−5.7
				Gain (90th percentile)					
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	5.6	6.7	6.7	6.8	7.7	7.7	1.9	2.8	2.6
6	7.7	9.4	9.2	9.3	10.1	10.1	3.7	4.6	4.4
8	9.0	10.8	10.8	10.4	11.2	11.3	5.5	5.8	5.7

Figure 4 in the case of $n = 4$. In addition to their magnitude, the maximum response differences tended to involve a large number of students. For example, the average number of students involved in an incident comprising the maximum response difference was 5.6% for $\mathcal{Q}$ scoring and 8.2% for $\mathcal{S}$ scoring. However, on one question 28.4% of the class would have been involved in such an incident under $\mathcal{Q}$ scoring and 30.7% under $\mathcal{S}$ scoring. Clearly, situations such as these could be quite difficult for the instructor to manage, with nearly one-third of the class pointing out that they received a lower score on a particular question than another student even though they assigned a higher probability on the correct statement.

Figure 4 Maximum Response Differences to Maintain Score Equality



Problems such as these occurred in all five years of the Stanford data on each of the 15 questions.[7]

We believe the behavior discussed in this section, a consequence of nonlocality, could make the scoring appear somewhat arbitrary, which would make it difficult to defend in certain situations.

3.4. Summary

Even though the rank correlations between the scoring rules are generally quite high, as demonstrated in §3.1, quadratic and spherical scoring often results in extreme ranking differences compared to logarithmic scoring, as demonstrated in §3.2, which will always rank assessors according to the probability they assigned to the correct statement. In addition, quadratic and spherical scoring introduce the possibility of one assessor assigning a lower probability than other assessors to the correct statement, but receiving a higher score. As demonstrated in §3.3, the range over which this can occur is so large that an assessor may accuse the scoring system of being arbitrary.

The rank order and score equality features of quadratic and spherical scoring discussed in this section are a consequence of nonlocality. Whether these issues are a concern depends on the context. In situations where rank order results are important (e.g., academic testing situations), we believe the results presented here support the use of logarithmic scoring.[8]

4. Nonlinear Utility

As discussed in §1, the honest revelation property of proper scoring rules is based on an assumption that the assessor, A, maximizes her expected score, or has a linear utility function. If this is not the case, then honest revelation is not optimal.[9] To see this, assume A's utility function $u(\mathcal{T})$ is a strictly increasing function of $\mathcal{T}$, where $\mathcal{T}$ is $\mathcal{Q}$, $\mathcal{S}$, or $\mathcal{L}$. A solves

$$\begin{aligned} \max_{\mathbf{r}} \quad & \sum_{i=1}^{n} p_i u(\mathcal{T}_i(\mathbf{r})) \\ \text{s.t.} \quad & \sum_{i=1}^{n} r_i = 1 \\ & r_i > 0 \quad \forall i, \end{aligned} \tag{6}$$

which yields the first-order condition (see appendix)

$$r_i^* = p_i \frac{u'(\mathcal{T}_i(\mathbf{r}^*))}{\sum_{j=1}^{n} p_j u'(\mathcal{T}_j(\mathbf{r}^*))} = p_i h_i^*, \tag{7}$$

where u' is the marginal utility and h_i^* is the *optimal hedge ratio*, which is always positive.

Define $\Delta_i = p_i - r_i^* = p_i(1 - h_i^*)$ as the *optimal reduction* for statement i. These reductions may also be negative in the case where A overstates his or her belief by setting $r_i^* > p_i$. If u is linear, then u' is constant, and $h_i^* = 1$ and $r_i^* = p_i$, which is the standard proper scoring rule result. Likewise, if $p_i = 1$, then $h_i^* = 1$ and $\Delta_i = 0$, since A's risk aversion is not material when he or she is sure of the correct statement.

Whether Equation (7) constitutes a maximum depends on the concavity of $u \circ \mathcal{T}_i$. The second derivative of $u \circ \mathcal{T}_i$ with respect to r_i is $u'' \cdot (\partial\mathcal{T}_i/\partial r_i)^2 + u' \cdot \partial^2\mathcal{T}_i/\partial r_i^2$, which must be less than 0 for a maximum. In the cases of Q, S, and L, $\partial^2\mathcal{T}_i/\partial r_i^2$ is negative (see appendix). Therefore, the term $u' \cdot \partial^2\mathcal{T}_i/\partial r_i^2$ is negative and Equation (7) will constitute a maximum if A is risk averse ($u'' < 0$). Because the constraints in Equation (6) form a convex feasible region, this maximum is unique. If A is risk preferring, then a maximum will be obtained only if $-u'/u'' < (\partial\mathcal{T}_i/\partial r_i)^2/(\partial^2\mathcal{T}_i/\partial r_i^2)$, where the inequality has been reversed because $\partial^2\mathcal{T}_i/\partial r_i^2$ is negative. This places a constraint on A's risk tolerance function, in that A has to be sufficiently risk preferring for a maximum to be obtained. In this paper, we only address risk-averse assessors, by assuming $u'' < 0$.

Equation (7) is troubling for two reasons. First, the optimal response is confounded by A's utility function through the composite function $u' \circ \mathcal{T}_i$, interfering with the honest revelation property of the scoring rule. Second, it is transcendental and solutions require numerical methods, because A's assignment to any

[7] Please see the online supplement on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html) for a detailed breakdown of performance by year.

[8] If the assessor is concerned about rank order performance, then the scoring rules discussed in this paper are no longer proper (Lichtendahl and Winkler 2007).

[9] We are using "honest" as a shorthand for $\mathbf{r}^* = \mathbf{p}$. Clearly, an assessor with nonlinear utility who assigns $\mathbf{r}^* \neq \mathbf{p}$ is being rational and not necessarily dishonest, because the specification of $\mathbf{r}$ is an exercise in decision making and not strictly a probability assessment.

statement i depends on his or her assignments to the other statements. It is unlikely that individuals can reliably carry out such a calculation without computer assistance. Given these complexities, it would help if we could determine how different the optimal response under a nonlinear utility function is from the risk-neutral solution of $\mathbf{r}^* = \mathbf{p}$. Perhaps under conditions that are likely to hold in actual assessments, the risk-neutral response is close to the optimal response under nonlinear utility. If so, this would simplify A's assessment task and encourage the use of scoring rules in the assessment process. We address this issue in the remainder of this section.

4.1. Maximum Optimal Reduction

Suppose A believes that the probability statement i is correct is p_i and is contemplating an ith response r_i. The reduction Δ_i will reach a maximum when A believes the remaining non-i statements are equally likely and a minimum when he or she believes only one of the non-i statements could be true. Believing the non-i statements are equally likely is the worst or most risky situation for the assessor because it forces him or her to spread the remaining assignment out equally and thereby requires a larger reduction in r_i to properly hedge the response. Thus, A's reduction will be maximal when his or her response on each of the remaining $n-1$ statements is uniform and equal to $(1-r_i)(n-1)^{-1}$, that is, when $\mathbf{r} = \mathbf{r}^+$. Another way to see this is to note that Δ_i will reach a maximum for a given p_i when h_i^* achieves a minimum. By assigning $\mathbf{r}^+$, A maximizes $\mathcal{T}_i$ and therefore minimizes the numerator of h_i^*, because u' is decreasing. Furthermore, an assignment of $\mathbf{r}^+$ maximizes the denominator of h_i^* (see appendix).

Since A's response on the non-i statements is uniform, h_i^* is independent of A's probability assessment for each of the remaining statements, as long as they sum to $1-r_i$. This is quite helpful because it allows us to investigate the *maximum optimal reduction* Δ_i^+ by investigating responses involving only r_i and p_i, essentially reducing the n-dimensional assessment of $\mathbf{r}$ to the simple case of $n=2$. We can then think of A as assigning r_i to statement i and $(1-r_i)\cdot(n-1)^{-1}$ to each of the remaining $n-1$ statements whose total probability is $1-p_i$. In this case, the scoring rules become:

Q:

$$Q_i(r_i) = 2r_i - (r_i^2 + (n-1)(1-r_i)^2(n-1)^{-2})$$
$$= 2r_i - (r_i^2 + (1-r_i)^2(n-1)^{-1}),$$
$$Q_{j\neq i}(r_i) = 2(1-r_i)(n-1)^{-1} - (r_i^2 + (1-r_i)^2(n-1)^{-1})$$

S:

$$S_i(r_i) = r_i/(r_i^2 + (n-1)(1-r_i)^2(n-1)^{-2})^{1/2}$$
$$= r_i/(r_i^2 + (1-r_i)^2(n-1)^{-1})^{1/2}$$
$$S_{j\neq i}(r_i) = (1-r_i)(n-1)^{-1}/(r_i^2 + (1-r_i)^2(n-1)^{-1})^{1/2}$$

L:

$$L_i(r_i) = \ln(r_i)$$
$$L_{j\neq i}(r_i) = \ln((1-r_i)(n-1)^{-1}).$$

Because each of the remaining $n-1$ statements yields the same score if correct, Equation (7) reduces to

$$r_i^* = p_i \frac{u'(\mathcal{T}_i(r_i^*))}{p_i u'(\mathcal{T}_i(r_i^*)) + (1-p_i)u'(\mathcal{T}_{j\neq i}(r_i^*))}. \tag{8}$$

4.2. Exponential Utility

To obtain some numerical results, assume $u(\mathcal{T}) = -\exp[-R^{-1}\mathcal{T}]$, where $R > 0$ is A's risk tolerance.[10] Equation (7) then becomes[11]

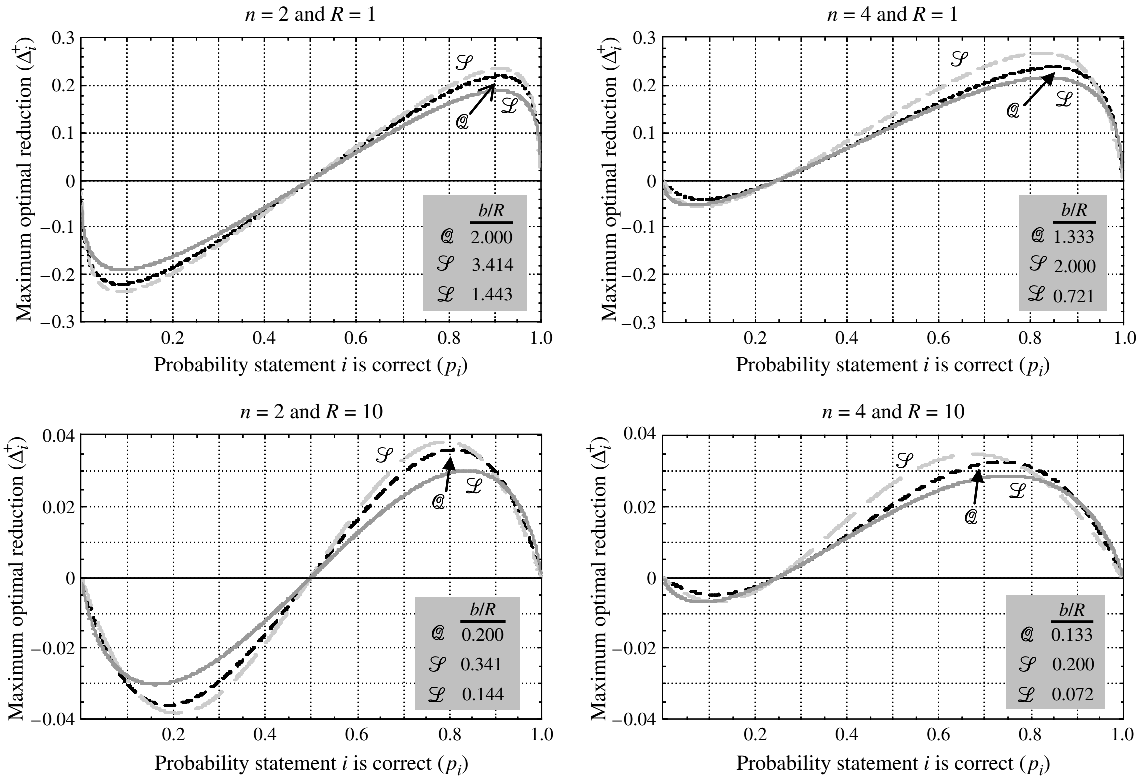
$$r_i^* = p_i \frac{\exp(-b_{\mathcal{T}} R^{-1} T_i(\mathbf{r}^*))}{\sum_{j=1}^n p_j \exp(-b_{\mathcal{T}} R^{-1} T_j(\mathbf{r}^*))}. \tag{9}$$

Equation (9) provides several insights. First, under exponential utility, the optimal response is independent of the $a_{\mathcal{T}}$ used in the normalization scheme. This stems from the "delta property" exhibited by the exponential utility function: If we add an amount a to all outcomes of a gamble, then the gamble's certain equivalent increases by a. Second, the term $b_{\mathcal{T}} R^{-1} T_i(\mathbf{r}^*)$ can be interpreted in several ways: (1) $b_{\mathcal{T}} T_i(\mathbf{r}^*)$ could be viewed as scaling the core scoring rule T, (2) $b_{\mathcal{T}} R^{-1}$ could be interpreted as an effective risk aversion with larger bs, leading to greater

[10] It has been demonstrated under a variety of conditions that exponential utility is a good approximation of the decision maker's true utility function (Kirkwood 2004).

[11] In the case of L scoring, this reduces to $r_i^* = p_i(r_i^*)^{-b_{\mathcal{L}}R^{-1}} \cdot (\sum_{j=1}^n p_j(r_j^*)^{-b_{\mathcal{L}}R^{-1}})^{-1}$.

Figure 5 Maximum Reductions under Exponential Utility



risk aversion, or (3) $b_{\mathcal{T}}R^{-1}$ could be considered a normalization factor that is scaled by A's risk tolerance. In this sense, $b_{\mathcal{T}}R^{-1}$ measures the "concern" induced by a particular set of statements. For example, since Q ranges from −1 to 1, $b_{\mathcal{T}}R^{-1}$ measures the range of possible losses and gains as a fraction (or multiple) of A's risk tolerance.

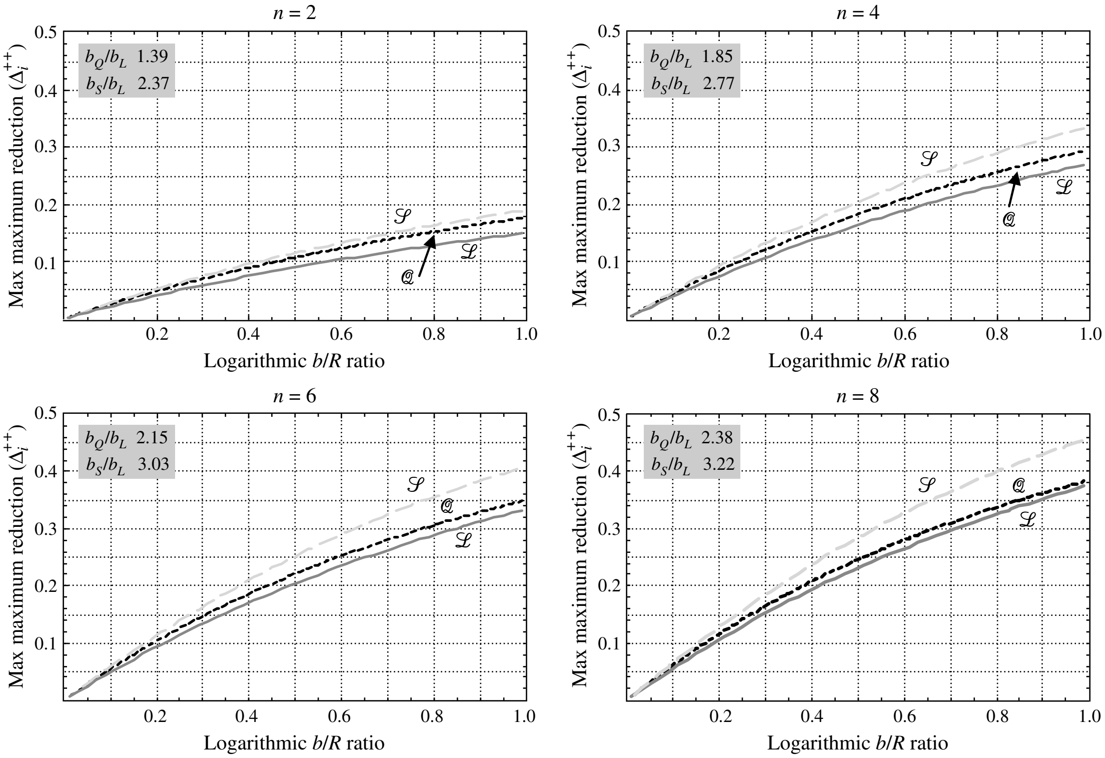
This latter interpretation is particularly attractive because it serves to explain some experimental results involving scoring rules. For example, Jensen and Peterson (1973) compared subjects' responses under Q, S, or L scoring and found that the particular rule used made little difference, while the "steepness" of the linear transformation between the core scoring rule T and the normalized rule $\mathcal{T}$ (i.e., $b_{\mathcal{T}}$) was significant. Specifically, Jensen and Peterson found that making the function steeper tended to reduce the response assigned to the most likely statement.

Illustrative Example. Unfortunately, the transcendental nature of Equation (9) hinders the drawing of general conclusions regarding the scoring rules. However, we can produce some interesting numerical results.[12]

Suppose A faces two different assessment situations involving the same stakes, but in one case there are two statements ($n = 2$) and in the other there are four ($n = 4$). Assume further that A is being scored under Q, S, or L scoring, using the normalization in Table 1 with $\alpha = 1$. Figure 5 displays A's maximum optimal reduction as a function of p_i when $R = 1$ and $R = 10$; recall that these reductions are the largest possible for a given n. Note that the reductions are 0 when A is certain about statement i ($p_i = 0$ or 1) or when $p_i = n^{-1}$. In all cases, $\mathcal{S}$ has the largest b/R ratio and $\mathcal{L}$ has the lowest, which is a result of the normalization presented in Table 1, which implies $b_{\mathcal{S}} > b_{\mathcal{Q}} > b_{\mathcal{L}}$. For example, in the case of $n = 2$, $b_{\mathcal{Q}}/b_{\mathcal{L}} = 2\ln 2 \approx 1.39$, $b_{\mathcal{S}}/b_{\mathcal{L}} = \sqrt{2}\ln 2/(\sqrt{2} - 1) \approx 2.37$, and $b_{\mathcal{S}}/b_{\mathcal{Q}} = \sqrt{2}/(2(\sqrt{2} - 1)) \approx 1.71$; S has to be "bent" or

[12] These results were obtained with the software system Mathematica (http://www.wolfram.com).

Figure 6 **Sensitivity of Maximum Maximum-Optimal-Reductions**



"stretched" more than the other rules given that its natural range is $[0, 1]$.

When $R = 1$, the reductions reach a maximum of approximately 0.20. However, as demonstrated by the b/R ratios, which range from 0.7 to 3.4, these reductions take place in an environment where the potential losses and gains are on the order of A's risk tolerance or much larger. When $R = 10$, the reductions are less than 0.04, even though b/R ranges from 0.07 to 0.34, which is still relatively large.

The effect of increasing n is not clear; it increases the reduction when $R = 1$ but lowers it when $R = 10$. This is because increasing n lowers the probability assigned to the non-i statements, which should serve to increase the reduction, but also lowers $b_{\mathcal{T}}$, which serves to lower the reduction.

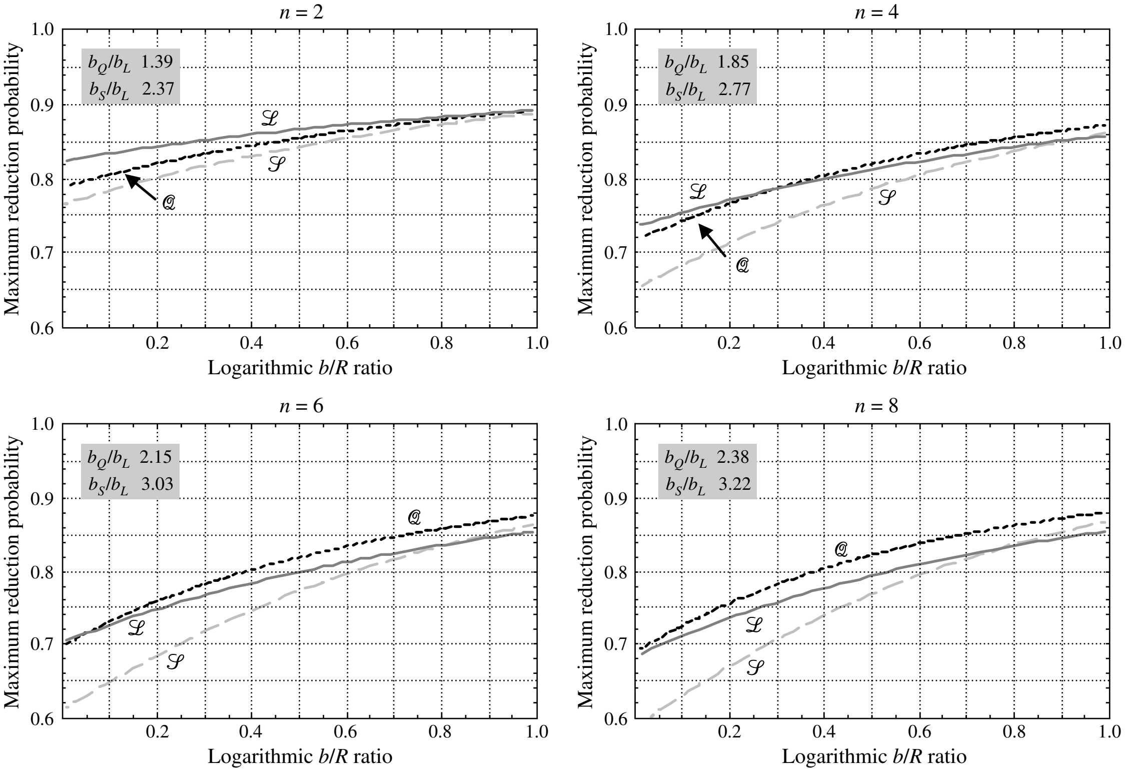
Another interesting feature of Figure 5 is that in every case, $\mathcal{L}$ generates the lowest *maximum maximum-optimal-reduction* Δ_i^{++}. For example, when $n = 2$ and $R = 10$, the maximum optimal reduction reaches a maximum of approximately 0.03 for $\mathcal{L}$ but is larger than 0.03 for $\mathcal{Q}$ and $\mathcal{S}$. This performance holds over a wide range of beliefs; $\mathcal{Q}$ and $\mathcal{S}$ only generate lower reductions for p_i near 0 or 1. This might be surprising, given that $\mathcal{L}$ holds the prospect of an infinitely negative score. $\mathcal{S}$ performs the worst in the cases depicted in Figure 5, with larger reductions occurring at less extreme values of p_i.

Figure 6 displays the maximum maximum-optimal-reduction as a function of the number of statements, n, and the logarithmic $b_{\mathcal{L}}/R$ ratio. The $b_{\mathcal{T}}/R$ ratios for quadratic and spherical scoring are determined using the $b_{\mathcal{Q}}/b_{\mathcal{L}}$ and $b_{\mathcal{S}}/b_{\mathcal{L}}$ ratios implicit in Table 1, which ensures that the rules are comparable.

Figure 6 demonstrates that $\mathcal{L}$ induces lower reductions than either $\mathcal{Q}$ or $\mathcal{S}$ over a wide range of scenarios, while $\mathcal{S}$ generates the largest. Because the b ratios are independent of the α and β used in the normalization scheme, these results hold for *any* normalization that yields α whenever A assigns 1 to the correct statement and β whenever he or she assigns $\mathbf{u}_n$.

Unfortunately, these results do not hold for all possible normalizations. Assume we have set $\mathbf{r}_\alpha = \boldsymbol{\delta}_i$ but have not constrained $\mathbf{r}_\beta$. In this case, $b_{\mathcal{Q}}/b_{\mathcal{L}} = L_i(\mathbf{r}_\beta)(Q_i(\mathbf{r}_\beta) - 1)^{-1}$ and $b_{\mathcal{S}}/b_{\mathcal{L}} = L_i(\mathbf{r}_\beta)(S_i(\mathbf{r}_\beta) - 1)^{-1}$. These ratios will achieve a minimum when Q_i and

Figure 7 Probability at Which Largest Maximum Reduction Occurs



S_i are minimized because $L_i \le 0$. Q_i and S_i are minimized when $\mathbf{r}_\beta = \mathbf{r}^-$. Because the non-$i$ response is concentrated at a single point, the minimum values for b_Q/b_L and b_S/b_L are independent of n. Numerical optimization yields a minimum for b_Q/b_L of approximately 1.23 when $r_i^- = 0.28$ and 1.98 for b_S/b_L when $r_i^- = 0.32$. A repeat of the analysis in Figure 6 demonstrates that $\mathscr{Q}$ induces the lowest reductions in this case. However, such a normalization scheme seems artificial.

As shown in Figure 5, in addition to lower reductions, $\mathscr{L}$'s maximum reduction tends to be shifted toward more extreme values of p_i. This is important in practice because it suggests that $\mathscr{Q}$ and $\mathscr{S}$ have wider ranges of larger deviations from honest assessments. Figure 7 plots the probability at which Δ_i^{++} occurs, for the range of parameters in Figure 6. When $n = 2$, the maximum reductions for $\mathscr{L}$ occur at higher probabilities than for either $\mathscr{Q}$ or $\mathscr{S}$. Thus, $\mathscr{L}$ tends to induce lower reductions that are shifted toward more extreme assessments. This may be quite beneficial in practice, because it results in a smaller region in which nonlinear utility has a material effect. When $n = 4$, $\mathscr{L}$'s largest maximum reductions again occur to the right of $\mathscr{Q}$ as long as $b_\mathscr{L}/R$ is less than about 0.25. For higher values of n, $\mathscr{Q}$ tends to outperform $\mathscr{L}$ or $\mathscr{S}$ on this measure.

4.3. General Logarithmic Scoring Results

We believe the features of logarithmic scoring discussed in §3 and §4.2 make a strong case for its use in practice. Therefore, in this section we offer some general results that aim to provide guidance for the use of logarithmic scoring in actual assessment situations.

Figure 5 suggests that even relatively high stakes should result in modest deviations from the optimal risk-neutral response.[13] Figure 8 plots the maximum maximum-optimal reduction for $\mathscr{L}$ scoring as a function of $b_\mathscr{L}/R$ and n. Suppose we wanted to ensure that all reductions are less than 0.03, believing that individuals cannot assesses their beliefs more precisely

[13] Descriptively, individuals may reduce their assessments more than this (Phillips and Edwards 1966).

Figure 8 **General Logarithmic Scoring Results**

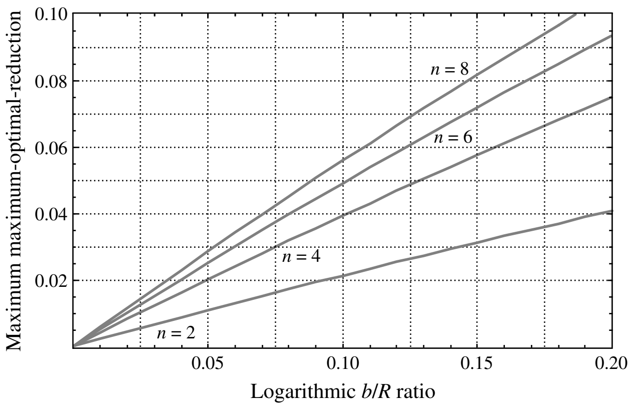
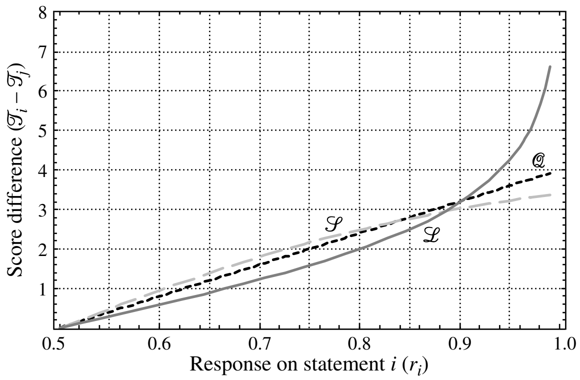


Figure 9 **Differences in Normalized Scores** $(n = 2)$



than this. Then $b_{\mathcal{L}}/R$ should be less than about 0.14 when $n = 2$ and less than about 0.075 when $n = 4$. This seems quite reasonable in practice when monetary amounts are involved. For example, if A's risk tolerance is \$10,000, then b should be held below \$1,400 when $n = 2$ and below \$750 when $n = 4$. It is unlikely that assessments made as part of a decision analysis project or an assessment experiment would involve monetary amounts of this magnitude. Even if $b_{\mathcal{L}}/R = 0.025$, Δ_i^{++} is only 0.01 when $n = 4$. Because this is the *largest* possible reduction under an exponential utility function, the actual optimal reductions under the range of possible beliefs will be even lower. In non-monetary settings, such as academic testing situations, Figure 8 can be used to determine the largest value of $b_{\mathcal{L}}$that should be used, given beliefs about A's risk tolerance.

4.4. Why L Performs Well

As discussed above, the transcendental nature of Equations (7), (8), and (9) makes it difficult to draw general conclusions regarding the scoring rules under nonlinear utility (e.g., why L scoring induces the lowest deviation from the optimal risk-neutral response). However, we can gain some insight by investigating the behavior of h_i^*. From Equation (8) we have

$$(h_i^*)^{-1} = \frac{p_i u'(\mathcal{T}_i(r_i^*)) + (1-p_i)u'(\mathcal{T}_{j\neq i}(r_i^*))}{u'(\mathcal{T}_i(r_i^*))} = p_i + (1-p_i)\frac{u'(\mathcal{T}_{j\neq i}(r_i^*))}{u'(\mathcal{T}_i(r_i^*))}.$$

The maximum reduction is driven by the ratio of A's marginal utility at the points $\mathcal{T}_j$ and $\mathcal{T}_i$. As this ratio is increased, $(h_i^*)^{-1}$ increases and therefore h_i^* decreases. In the case of exponential utility this ratio is

$$\frac{u'(\mathcal{T}_{j\neq i}(r_i^*))}{u'(\mathcal{T}_i(r_i^*))} = \frac{\exp[-b_{\mathcal{T}}R^{-1}T_j(r_i^*)]}{\exp[-b_{\mathcal{T}}R^{-1}T_i(r_i^*)]} = \exp[b_{\mathcal{T}}R^{-1}(T_i(r_i^*) - T_j(r_i^*))].$$

Under exponential utility, the deviation from the risk-neutral assessment depends on the difference between scores T_i and T_j, scaled by $b_{\mathcal{T}}R^{-1}$. $\mathcal{Q}$ and $\mathcal{L}$, for example, will generate the same h_i^* when $b_{\mathcal{Q}}R^{-1}(Q_i(r_i^*) - Q_j(r_i^*)) = b_{\mathcal{L}}R^{-1}(L_i(r_i^*) - L_j(r_i^*))$, which is independent of R. Therefore, A's risk tolerance governs the magnitude of h_i^* under the different rules, but not if these values are equal. The critical term is then $b_{\mathcal{T}}(T_i(r_i^*) - T_j(r_i^*)) = \mathcal{T}_i(r_i^*) - \mathcal{T}_j(r_i^*)$. Figure 9 plots the difference between the normalized scoring rules from Figure 1 for the case where $n = 2$, and the possible responses are $r_i \geq 0.5$ and $r_j = 1 - r_i$. $\mathcal{L}$ produces lower score differences than either $\mathcal{S}$ or $\mathcal{Q}$, provided r_i is less than 0.885 or 0.904, respectively. $\mathcal{Q}$ yields lower score differences than $\mathcal{S}$, provided $r_i \leq 0.838$.

5. Conclusions

Strictly proper scoring rules continue to play an important role in probability assessment. Analysts have many rules from which to choose. In this paper we have quantified the implications of the nonlocality feature of quadratic and spherical scoring and demonstrated the superior performance of logarithmic scoring under nonlinear utility.

Because of its local property, logarithmic scoring will always assign a higher score to higher-probability assignments on the correct statement. Both spherical and quadratic (and therefore Brier) scoring can perform poorly in this regard. Where rank ordering is important, as it is in academic settings, this argues strongly for the logarithmic rule.

In addition, logarithmic scoring induces lower deviations from honest assessments under exponential utility in a wide range of scenarios. This conclusion is somewhat surprising, given that logarithmic may yield an infinitely negative score. This behavior is important in practice because the risk-neutral assumption is an ideal that may not be realized in actual assessment situations. If the analyst is concerned about the impact of nonlinear utility, then the logarithmic rule will reduce the magnitude of this problem.

Finally, we have demonstrated that the overall impact of nonlinear utility is quite small. In most assessment situations, deviations from honest assessments should be small, which should provide comfort to those who would like to use strictly proper scoring rules as part of the probability assessment process.

An online supplement to this paper is available on the *Decision Analysis* website (http://da.pubs.informs.org/online-supp.html).

Acknowledgments
The author thanks the following individuals: Ross Shachter for a stimulating discussion after the author's INFORMS 2005 talk on scoring rules that spurred the author's interest in the rank order properties; Thomas Seyller for providing three years of Stanford scoring data; Guy Curry for a helpful conversation regarding the optimal response for the spherical scoring rule; and Jim Smith, Bob Winkler, two referees, and the associate editor for helpful comments and suggestions on a draft of this paper.

Appendix

Ex Post Normalization

We want to normalize the rules such that $\mathscr{T}_j(\mathbf{r}_\alpha) = a + bT_j(\mathbf{r}_\alpha) = \alpha$ and $\mathscr{T}_k(\mathbf{r}_\beta) = a + bT_k(\mathbf{r}_\beta) = \beta$. From the α condition we have $a = \alpha - bT_j(\mathbf{r}_\alpha)$. From the β condition we have $\mathscr{T}_k(\mathbf{r}_\beta) = \alpha + b[T_k(\mathbf{r}_\beta) - T_j(\mathbf{r}_\alpha)] = \beta$ or $b = (\beta-\alpha)[T_k(\mathbf{r}_\beta) - T_j(\mathbf{r}_\alpha)]^{-1}$. Therefore, $a = \alpha - (\beta-\alpha)[T_k(\mathbf{r}_\beta) - T_j(\mathbf{r}_\alpha)]^{-1}\mathrm{T}_\mathrm{j}(\mathbf{r}_\alpha) = (\alpha T_\mathrm{k}(\mathbf{r}_\beta) - \beta T_\mathrm{j}(\mathbf{r}_\alpha))[\mathrm{T}_\mathrm{k}(\mathbf{r}_\beta) - T_\mathrm{j}(\mathbf{r}_\alpha)]^{-1}$.

Nonlinear Utility Solution

Q: Form the Lagrangian

$$\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{n} p_i u(\mathscr{Q}_i(\mathbf{r})) + \lambda\left(\sum_{i=1}^{n} r_i - 1\right) + \boldsymbol{\mu}\cdot\mathbf{r}.$$

Rewriting in terms of the core scoring rule, we have

$$\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{n} p_i u\left(a + b\left(2r_i - \sum_{j=1}^{n} r_j^2\right)\right) + \lambda\left(\sum_{i=1}^{n} r_i - 1\right) + \boldsymbol{\mu}\cdot\mathbf{r}.$$

Differentiating $\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu})$ with respect to r_i and setting equal to 0, we obtain

$$2b\left[p_i u'(\mathscr{Q}_i(\mathbf{r}^*)) - r_i^* \sum_{j=1}^{n} p_j u'(\mathscr{Q}_j(\mathbf{r}^*))\right] + \lambda + \mu_i = 0 \quad \text{or}$$

$$r_i^* = \frac{p_i u'(\mathscr{Q}_i(\mathbf{r}^*)) + (\lambda + \mu_i)/2b}{\sum_{j=1}^{n} p_j u'(\mathscr{Q}_j(\mathbf{r}^*))}.$$

From the equality constraint we have $\sum_{i=1}^{n} r_i^* - 1 = 0$. Substitution of r_i^* yields the condition that $\mu_i^* = -\lambda^*$. Therefore, $r_i^* = p_i u'(\mathscr{Q}_i(\mathbf{r}^*))/\sum_{j=1}^{n} p_j u'(\mathscr{Q}_j(\mathbf{r}^*))$.

S: This derivation follows the proof from Toda (1963) that S is strictly proper under risk neutrality. Because $S(\mathbf{r}) = S(a\cdot\mathbf{r})$, a multiple of $\mathbf{r}$ yields the same score as $\mathbf{r}$ and a family of optimal solutions exists. Therefore, we are free to set $S_i(\mathbf{r}) = r_i/k$ under the constraint that $k = (\sum_{i=1}^{n} r_i^2)^{1/2}$ or $k^2 = \sum_{i=1}^{n} r_i^2$. Assume the constraint $r_i > 0$ is not binding at the optimal and form the Lagrangian

$$\Gamma(\mathbf{r}, \lambda) = \sum_{i=1}^{n} p_i u(r_i k^{-1}) - \lambda\left(\sum_{i=1}^{n} r_i^2 - k^2\right).$$

Differentiating $\Gamma(\mathbf{r}, \lambda)$ with respect to r_i and setting it equal to 0, we obtain $r_i^* = p_i u'(\mathscr{S}_i(\mathbf{r}^*))/(2\lambda k)$. From the equality constraint, we have $\sum_{i=1}^{n} (r_i^*)^2 - k^2 = 0$. Substitution of r_i^* yields $(\lambda^*)^2 = \sum_{i=1}^{n} [p_i u'(\mathscr{S}_i(\mathbf{r}^*))]^2/(4k^4)$. From r_i^* we see that $\lambda > 0$, and therefore, $\lambda^* = (\sum_{i=1}^{n} [p_i u'(\mathscr{S}_i(\mathbf{r}^*))]^2)^{1/2}/(2k^2)$. Substitution of λ^* into r_i^* yields

$$r_i^* = p_i u'(\mathscr{S}_i(\mathbf{r}^*))k\left(\sum_{j=1}^{n} [p_j u'(\mathscr{S}_j(\mathbf{r}^*))]^2\right)^{-1/2}.$$

Because k is an arbitrary constant, we can let

$$k = \left(\sum_{j=1}^{n} [p_j u'(\mathscr{S}_j(\mathbf{r}^*))]^2\right)^{1/2} \left(\sum_{j=1}^{n} p_j u'(\mathscr{S}_j(\mathbf{r}^*))\right)^{-1},$$

in which case $r_i^* = p_i u'(\mathscr{S}_i(\mathbf{r}^*))/(\sum_{j=1}^{n} p_j u'(\mathscr{S}_j(\mathbf{r}^*)))$. Because $r_i^* > 0$, we see that the nonnegativity constraint is not binding.

L: Form the Lagrangian

$$\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{n} p_i u(\mathscr{L}_i(\mathbf{r})) + \lambda\left(\sum_{i=1}^{n} r_i - 1\right) + \boldsymbol{\mu}\cdot\mathbf{r}.$$

Rewriting in terms of the core scoring rule, we have

$$\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu}) = \sum_{i=1}^{n} p_i u(a + b\ln r_i) + \lambda\left(\sum_{i=1}^{n} r_i - 1\right) + \boldsymbol{\mu}\cdot\mathbf{r}.$$

Differentiating $\Gamma(\mathbf{r}, \lambda, \boldsymbol{\mu})$ with respect to r_i and setting equal to 0, we obtain

$$bp_i u'(\mathscr{L}_i(\mathbf{r}^*))\frac{1}{r_i^*} + \lambda + \mu_i = 0 \quad \text{or} \quad r_i^* = p_i \frac{b}{-(\lambda + \mu_i)} u'(\mathscr{L}_i(\mathbf{r}^*)).$$

From the equality constraint, we have $\sum_{i=1}^{n} r_i^* - 1 = 0$. Substitution of r_i^* yields the condition that $-(\lambda + \mu_i) = b\sum_{i=1}^{n} p_i u'(\mathscr{L}_i(\mathbf{r}^*))$. Therefore,

$$r_i^* = \frac{p_i u'(\mathscr{L}_i(\mathbf{r}^*))}{\sum_{j=1}^{n} \cdot p_j u'(\mathscr{L}_j(\mathbf{r}^*))}.$$

Concavity of Scoring Rules

Q: $Q_i(\mathbf{r}) = 2r_i - \mathbf{r}\cdot\mathbf{r}$. $\partial Q_i(\mathbf{r})/\partial r_i = 2 - 2r_i = 2(1 - r_i) > 0$ and $\partial^2 Q_i(\mathbf{r})/\partial r_i^2 = -2 < 0$.

S: $S_i(\mathbf{r}) = r_i/(\mathbf{r}\cdot\mathbf{r})^{1/2}$.

$$\begin{aligned}\frac{\partial S_i(\mathbf{r})}{\partial r_i} &= \frac{(\mathbf{r}\cdot\mathbf{r})^{1/2} - r_i^2(\mathbf{r}\cdot\mathbf{r})^{-1/2}}{\mathbf{r}\cdot\mathbf{r}}\\ &= (\mathbf{r}\cdot\mathbf{r})^{-1/2}[1 - r_i^2(\mathbf{r}\cdot\mathbf{r})^{-1}],\end{aligned}$$

which will be positive when $r_i \neq 1$ or $\mathbf{r} \neq \mathbf{u}_n$, because $r_i^2(\mathbf{r}\cdot\mathbf{r})^{-1} < 1$.

$$\begin{aligned}\partial^2 S_i(\mathbf{r})/\partial r_i^2 &= -r_i(\mathbf{r}\cdot\mathbf{r})^{-3/2} - 2r_i(\mathbf{r}\cdot\mathbf{r})^{-3/2} + 3r_i^3(\mathbf{r}\cdot\mathbf{r})^{-5/2}\\ &= -3r_i(\mathbf{r}\cdot\mathbf{r})^{-3/2} + 3r_i^3(\mathbf{r}\cdot\mathbf{r})^{-5/2}\\ &= 3r_i(\mathbf{r}\cdot\mathbf{r})^{-3/2}(r_i^2(\mathbf{r}\cdot\mathbf{r})^{-1} - 1),\end{aligned}$$

which is negative as long as $r_i \neq 1$ or $\mathbf{r} \neq \mathbf{u}_n$.

L: $L_i(\mathbf{r}) = \ln(r_i)$. $\partial L_i(\mathbf{r})/\partial r_i = r_i^{-1} > 0$ and $\partial^2 L_i(\mathbf{r})/\partial r_i^2 = -r_i^{-2} < 0$.

Maximum Optimal Reduction

The denominator of h_i^* can be written as $p_i u'(\mathcal{T}_i(\mathbf{r}^*)) + \sum_{j=1,\, j\neq i}^{n} p_j u'(\mathcal{T}_j(\mathbf{r}^*))$. We have a belief p_i and are considering the assignment of r_i. We seek the set of beliefs and responses on the non-i statements that will force the denominator of h_i^* to be maximal. Formally, we solve

$$\begin{aligned}\max_{p_j,\, r_j^*}\ & p_i u'(\mathcal{T}_i(\mathbf{r}^*)) + \sum_{j=1,\, j\neq i}^{n} p_j u'(\mathcal{T}_j(\mathbf{r}^*))\\ \text{s.t.}\ & \sum_{j=1,\, j\neq i}^{n} p_j = 1 - p_i\\ & \sum_{j=1,\, j\neq i}^{n} r_j^* = 1 - r_i.\end{aligned}$$

Forming the Lagrangian, we have

$$\begin{aligned}\Gamma(\mathbf{p}, \mathbf{r}^*, \lambda, \mu) = {}& p_i u'(\mathcal{T}_i(\mathbf{r}^*)) + \sum_{j=1,\, j\neq i}^{n} p_j u'(\mathcal{T}_j(\mathbf{r}^*))\\ &+ \lambda\Bigg(\sum_{j=1,\, j\neq i}^{n} p_j - 1 - p_i\Bigg) + \mu\Bigg(\sum_{j=1,\, j\neq i}^{n} r_j^* - 1 - r_i\Bigg).\end{aligned}$$

Differentiating $\Gamma(\mathbf{p}, \mathbf{r}^*, \lambda, \mu)$ with respect to p_j we obtain $\partial\Gamma/\partial p_j = u'(\mathcal{T}_j(\mathbf{r}^*)) + \lambda$. Setting this equal to 0 yields the condition $u'(\mathcal{T}_j(\mathbf{r}^*)) = -\lambda$. If u' has an inverse, which we assume, then $\mathcal{T}_j(\mathbf{r}^*) = u'^{-1}(-\lambda)$, which is a constant, and the scores on the non-i statements are equal. This holds for any r_j^*, and therefore the assignments on the remaining statements are equal and $\mathbf{r} = \mathbf{r}^+$. The numerator of h_i^* is minimal when $\mathbf{r} = \mathbf{r}^+$ and the denominator is maximal. The fact that the response on the non-i statements is equal means the specific p_i values are immaterial, as long as they sum to $1 - p_i$. The mixed partial derivatives $\partial^2\Gamma/\partial p_j\partial r_j$ and $\partial^2\Gamma/\partial r_j\partial p_j$ are identical and equal to $u''(\mathcal{T}_j(\mathbf{r}^*))(\partial\mathcal{T}_j/\partial r_j) < 0$. $\partial^2\Gamma/\partial^2 p_j = 0$ and, therefore, the Hessian

$$\begin{bmatrix} \dfrac{\partial^2\Gamma}{\partial p_i^2} & \dfrac{\partial^2\Gamma}{\partial p_i\partial r_i}\\[2ex] \dfrac{\partial^2\Gamma}{\partial r_i\partial p_i} & \dfrac{\partial^2\Gamma}{\partial p_i^2}\end{bmatrix}$$

is negative semidefinite and the first-order conditions yield a maximum that is not strict. The maximum is not strict because there is an infinite number of probability assignments to the non-i statements that yield the maximum reduction as long as the non-i responses are forced to be equal. Therefore, the maximum reduction will be achieved when assignments to the non-i statements are equal. This would be achieved if the assessor believed the non-i statements to be equally likely or if we force the assessor to evenly distribute the non-i responses irrespective of his or her beliefs. Either interpretation generates the largest reductions and reduces the problem to the selection of r_i.

References

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

Cover, T. M., J. A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, New York.

Friedman, D. 1979. An effective scoring rule for probability distributions. Discussion Paper 164, University of California, Los Angeles, Los Angeles, CA.

Friedman, D. 1983. Effective scoring rules for probabilistic forecasts. *Management Sci.* **29**(4) 447–454.

Jensen, F. A., C. R. Peterson. 1973. Psychological effects of proper scoring rules. *Organ. Behav. Human Performance* **9** 307–317.

Kirkwood, C. W. 2004. Approximating risk aversion in decision analysis applications. *Decision Anal.* **1**(1) 51–67.

Lichtendahl, K. C., Jr., R. L. Winkler. 2007. Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.* Forthcoming.

Murphy, A. H., R. L. Winkler. 1970. Scoring rules in probability assessment and evaluation. *Acta Psych.* **34** 273–286.

Murphy, A. H., R. L. Winkler. 1971. Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteorological Soc.* **52**(4) 239–248.

Nau, R. F. 1985. Should scoring rules be "effective"? *Management Sci.* **31**(5) 527–535.

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

Roby, T. B. 1965. Belief states: A preliminary empirical study. *Behavioral Sci.* **10**(3) 255–270.

Savage. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801.

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

Shuford, E. H., Jr., A. Albert, H. E. Massengill. 1966. Admissible probability measurement procedures. *Psychometrika* **31**(2) 125–145.

Staël von Holstein, C.-A. S. 1970. *Assessment and Evaluation of Subjective Probability Distributions*. The Economic Research Institute at the Stockholm School of Economics, Stockholm, Sweden.

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

Winkler, R. L. 1968. "Good" probability assessors. *J. Appl. Meteorology* **7** 751–758.

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

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

Winkler, R. L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1) 1–60.

Winkler, R. L., A. H. Murphy. 1970. Nonlinear utility and the probability score. *J. Appl. Meteorology* **9** 143–148.