DECISION ANALYSIS
Vol. 7, No. 4, December 2010, pp. 346–357
ISSN 1545-8490 | EISSN 1545-8504 | 10 | 0704 | 0346

DOI 10.1287/deca.1100.0184

Scoring Rules and Decision Analysis Education

J. Eric Bickel
Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, Texas 78712,
ebickel@mail.utexas.edu

Experiential learning is perhaps the most effective way to teach. One example is the scoring procedure used for exams in some decision analysis programs. Under this grading scheme, students take a multiple-choice exam, but rather than simply marking which answer they think is correct, they must assign a probability to each possible answer. The exam is then scored with a special scoring rule, under which students' best strategy is to avoid guessing and instead assign their true beliefs. Such a scoring function is known as a strictly proper scoring rule. In this paper, we discuss several different scoring rules and demonstrate how their use in testing situations provides insights for both students and instructors.

Key words: scoring rules; education; probability assessment
History: Received on December 7, 2009. Accepted on June 28, 2010, after 1 revision. Published online in *Articles in Advance* September 3, 2010.

1. Background

In several decision analysis programs (e.g., those at Stanford University, the Darden School of Business, the University of Illinois at Urbana-Champaign, and the University of Texas at Austin), a portion of students' grades are based on their ability to provide high-quality probability assessments. Specifically, at each university, except Darden, students take multiple-choice exams or quizzes, but rather than simply marking the answer that they think is correct (or most likely to be correct), they must assign a probability to each possible answer.[1] Such an exam should better reveal the students' mastery of the subject, but how should the instructor assign scores in this situation?[2]

Formally, consider the assessment of a probability distribution by a student over n mutually exclusive and collectively exhaustive answers, where $n > 1$. Let $\mathbf{p} = (p_1, \ldots, p_n)$ be an n-vector of probabilities representing the student's private beliefs, where p_i is the probability the student assigns to answer i being correct, and the sum of these probabilities is equal to one. These beliefs represent the student's "true" state of knowledge, but are not directly observable by the instructor. Let the student's public assessment or response be given by $\mathbf{r} = (r_1, \ldots, r_n)$, where r_i is the stated probability (the student's answer) that answer i is correct, and the sum of these responses is equal to one.

Students are likely to have many objectives in such a situation, ranging from learning the material to obtaining a good grade. We assume that students seek to maximize their total course points. This simplification seems reasonable, particularly in programs that subdivide letter grades (e.g., B+, A−, A).

If the student is scored according to some function R, then her expected score when she assigns $\mathbf{r}$ and believes $\mathbf{p}$ is

$$\bar{R}(\mathbf{r} \mid \mathbf{p}) = E_p[R(\mathbf{r})] = \sum_i p_i R_i(\mathbf{r}), \qquad (1)$$

where R_i is the score received for assigning $\mathbf{r}$ when statement i is correct. If the student seeks to maximize her expected score, then the optimal response is

$$\mathbf{r}^* = \arg\max_r \bar{R}(\mathbf{r} \mid \mathbf{p}). \qquad (2)$$

1.1. The Problem with Standard Multiple-Choice Exams

If the scoring rule is linear, then the optimal response is to assign 1.0 to the answer that student believes

[1] At Darden, students either forecast whether or not an event will occur or provide quantiles for a continuous quantity.

[2] A summary version of this work appeared in Bickel (2009).

is most likely, which is the best strategy in traditional multiple-choice exams. Thus, students A and B believing $\mathbf{p}_A = (0.85, 0.15)$ and $\mathbf{p}_B = (0.51, 0.49)$, respectively, would both assign $\mathbf{r}_A = \mathbf{r}_B = (1, 0)$ and receive the same score. Likewise, students believing $\mathbf{p}_A = (0.51, 0.49)$ and $\mathbf{p}_B = (0.49, 0.51)$ would assign $\mathbf{r}_A = (1, 0)$ and $\mathbf{r}_B = (0, 1)$, respectively, and receive very different scores even though their knowledge is almost identical.

This insensitivity of a student's score to her knowledge is a major limitation of standard multiple-choice exams. These exams are not incentive compatible in that they do not encourage the students' responses to reflect their beliefs. A set of incentive-compatible scoring rules does exist and is discussed next.

2. Strictly Proper Scoring Rules

A *strictly proper scoring* rule T is a scoring function such that the student strictly maximizes her expected score by setting $\mathbf{r}^* = \mathbf{p}$; that is, $\overline{T}(\mathbf{r} \mid \mathbf{p}) < \overline{T}(\mathbf{p} \mid \mathbf{p})$ for all $\mathbf{r} \neq \mathbf{p}$ and $\overline{T}(\mathbf{r}^* \mid \mathbf{p}) = \overline{T}(\mathbf{p} \mid \mathbf{p})$ when $\mathbf{r}^* = \mathbf{p}$ (Toda 1963, Roby 1965, Shuford et al. 1966, Winkler 1968). Many strictly proper scoring rules have been developed. Three of the most popular are given below:

$$\textbf{Quadratic } (Q)\textbf{:} \quad Q_i(\mathbf{r}) = 2r_i - \mathbf{r} \cdot \mathbf{r} \in [-1, 1]; \tag{3}$$

$$\textbf{Spherical } (S)\textbf{:} \quad S_i(\mathbf{r}) = r_i/(\mathbf{r} \cdot \mathbf{r})^{1/2} \in [0, 1]; \tag{4}$$

$$\textbf{Logarithmic } (L)\textbf{:} \quad L_i(\mathbf{r}) = \ln(r_i) \in (-\infty, 0]. \tag{5}$$

The range of possible scores differs considerably. For example, logarithmic scoring holds the possibility of an infinitely negative score. Although this may seem like a defect, we will argue that this feature is a *benefit* of log scoring.

Any linear transformation of a strictly proper scoring rule is also strictly proper (Toda 1963).[3] Thus, the rules given in Equations (3)–(5) can be scaled to provide the maximum number of points desired on the exam. They can also be scaled such that a particular type of assessment, such as uniform, receives a particular score. Please see Bickel (2007) for a full discussion.

2.1. Deciding Among Scoring Rules

Given the rich choice of scoring rules, which one should be used? Scoring rules have ex ante and ex post properties (Winkler 1996). Ex ante they encourage the assessor to set their response equal to their belief. Ex post they can be used to evaluate assessment performance. In a classroom setting, instructors are likely to be interested in both uses. For example, the instructor wants to encourage truthful response, but also wants to evaluate the students based on their assessment.

Any strictly proper scoring rule provides the proper ex ante incentive—namely, to set $\mathbf{r}$ equal to $\mathbf{p}$. However, the rules differ in their ex post properties. The primary distinction between the rules comes down to whether or not one believes the score should depend only upon the probability assigned to the correct answer or if it should also depend upon probabilities assigned to answers that were incorrect, or events that failed to take place. In what follows, we provide a brief summary of three important properties: locality, sensitivity to distance, and sensitivity to nonlinear objectives. For a more detailed discussion, the interested reader should see Winkler (1996) and Bickel (2007). As the reader will see, we have a strong, and we believe reasonable, preference for logarithmic scoring.

2.1.1. Locality. Shuford et al. (1966) proved that when there are more than two possible answers, the logarithmic rule is the only proper scoring rule whose value depends only upon the probability assigned to the correct answer. This is referred to as the *local* property. From an ex post perspective, the local property is reminiscent of the likelihood principle (Winkler 1969, 1996). The likelihood principle states that when drawing a statistical inference (e.g., the instructor inferring the student's degree of mastery), all that should matter (normatively) is the likelihood of the observation, not the likelihood assigned to events that failed to occur. Of course, the likelihood principle follows automatically in a Bayesian setting. Because Bayesian thinking is fundamental to decision analysis, the use of a scoring rule that satisfies this property is pedagogically consistent. As Bernardo and Smith (2000, p. 72) wrote:

> It is intuitively clear that the preferences of an individual scientist faced with a pure inference problem

[3] The Brier score, which is heavily used in meteorology to measure forecast accuracy, is simply one minus the quadratic score. Under Brier scoring, one seeks to minimize, rather than maximize, their score.

> should correspond to the ordering induced by a local score function.... The individual scientist should be assessed (i.e., scored) only on the basis of his or her reported judgment about the plausibility of [the correct answer or the event that occurred].

From a practical perspective, locality or the lack thereof has three important implications:

1. Local rules should be easier for students to understand. For example, a two-dimensional chart can be provided that details their score for any set of assignments, which is only possible for other rules in special circumstances (e.g., the student assigns r_i to the correct answer and equally distributes the remaining probability among the $n-1$ answers).

2. A local rule will always assign higher scores to higher probability assignments to the correct answer. Q and S do not share this property when there are more than two answers. One implication of this feature is that one student may assign a higher (lower) probability than another student to the correct answer, but receive a lower (higher) score. For example, under Q scoring, with $n=4$, it is possible for an assignment on the correct answer of anywhere between 0.27 and 0.40 to earn the same score.

Bickel (2007) studied this problem in detail based on five years of Stanford midterm test data. The students in this sample were scored using the logarithmic scoring rule. However, if they had been scored under Q about 6% of the students on any given question would have received a *lower* score than another student even though they assigned a *higher* probability to the correct answer. In the case of S scoring, this percentage would have been about 8%. On one particular question, over 28% of the class would have been involved in such an incident under Q scoring and over 30% under S scoring. Instructors might find these situations difficult to manage, with almost one-third of the class pointing out that they received a lower score than other students on a particular question even though they assigned a higher probability to the correct answer.

3. Different nonlocal rules may generate different rank orderings among students for the same set of assessments. Bickel (2007) closely examined the rank order properties of Q, S, and L in actual testing situations and found that Q and S performed poorly in this regard. For example, about 10% of students would have fallen in rank about 7.5% if they had been scored with either Q or S instead of L. In other words, the choice of the scoring rule could lower students' scores almost a full letter grade (about 10%). Logarithmic scoring will always rank students based on the probability they assigned to the correct answer.

2.1.2. Sensitivity to Distance. Suppose the instructor believes that among the incorrect answers some are "more" incorrect than others. For example, one of the incorrect answers might stem from a common calculation error, whereas another may represent a fundamental misunderstanding of course material. In other words, there is an ordering among the possible answers. Scoring rules that can account for this are *sensitive to distance*. By their very nature, rules that are sensitive to distance are not local. Several authors have explored scoring rules that are sensitive to distance (Epstein 1969, Staël von Holstein 1970, Jose et al. 2009). The ordering of events is straightforward in situations where there is a natural ordering (e.g., the score in a sporting event or the closing price of a stock market index), but seems to be more difficult in a classroom setting. For example, the instructor would be required to construct an ordering of the answers in terms of their degree of "correctness." This task strikes us as difficult and, to our knowledge, rules that are sensitive to distance have never been used as part of the grading process in an educational setting. This may be a fruitful area of research.

2.1.3. Sensitivity to Nonlinear Objectives. The proof that students should respond truthfully is based on an assumption that they seek to maximize their expected score. If instead students are risk averse over the total number of points they earn in the course, then Q, S, and L are no longer strictly proper. However, Bickel (2007) demonstrated that logarithmic scoring is the least affected by this, which is surprising given that it introduces the possibility of an infinitely negative score. We will more fully discuss the issue of risk aversion and the negative infinity "problem" later in the paper.

In addition to risk aversion, competition among students also induces a nonlinear preference over course points. For example, if the students view themselves as competing for a limited number of A's, then they care about their rank relative to other students.

Unfortunately, under this objective the scoring rules are not strictly proper. Lichtendahl and Winkler (2007) investigated this problem in the case of two assessors under quadratic scoring. They showed that competition between assessors should lead assessors to provide more extreme (closer to categorical—i.e., 0 or 1) forecasts. Lichtendahl and Winkler (2007) did not analyze the relative sensitivity of each scoring rule to this effect, and this appears to be an open research question. However, as we show in §4, our data do not suggest that competition among students is a significant issue.

2.2. Scoring Rule Decomposition

Any strictly proper scoring rule can be decomposed into the sum of two components: one measuring "care" or sharpness and the other "honesty" or calibration (DeGroot and Fienberg 1982, Winkler 1996). In general, Equation (1) can be decomposed as (Winkler 1996)

$$\bar{R}(\mathbf{r}\mid\mathbf{p}) = \underbrace{\bar{R}(\mathbf{p}\mid\mathbf{p})}_{\text{Care}} + \underbrace{C(R,\mathbf{r},\mathbf{p})}_{\text{Honesty}}; \tag{6}$$

C is a penalty function, which is maximized at zero when $\mathbf{r}$ equals $\mathbf{p}$. Under L scoring, this decomposition yields

$$\begin{aligned}\bar{L}(\mathbf{r}\mid\mathbf{p}) &= \sum_i p_i \ln p_i - \sum_i p_i \ln\frac{p_i}{r_i}\\ &= -H(\mathbf{p}) - \mathrm{KL}(\mathbf{p}\,\|\,\mathbf{r}),\end{aligned} \tag{7}$$

where $H(\mathbf{p})$ is the entropy of $\mathbf{p}$ (Shannon 1948, Cover and Thomas 1991) and $\mathrm{KL}(\mathbf{p}\,\|\,\mathbf{r})$ is the Kullback–Leibler (KL) divergence between $\mathbf{p}$ and $\mathbf{r}$ (Kullback and Leibler 1951, Cover and Thomas 1991).[4] The KL divergence is minimized at zero when $\mathbf{r}=\mathbf{p}$. If the student assigns her true beliefs, then her expected score is simply the negentropy of $\mathbf{p}$ (i.e., the negative of the entropy of $\mathbf{p}$), which measures the sharpness of the student's assignment. A categorical assignment of zero or one has zero entropy, and uniform assignment has maximum entropy (equal to $\ln n$). Thus, the calibrated student can maximize her score by reducing the entropy of $\mathbf{p}$, which implies that she must have greater knowledge. Thus, the use of L scoring has the property that calibrated students can only increase their score by improving their knowledge of the test material.

3. Classroom Implementation

Based on the properties discussed above, we decided to use logarithmic scoring for the exams discussed in this paper. Specifically, students were scored based on the following rule:

$$\begin{aligned} L_i(\mathbf{r}) &= a + b\ln(r_i),\\ a &= 100/N\\ b &= a/\ln(n),\end{aligned} \tag{8}$$

where N is the total number of questions on the exam, and n is the number of possible answers for each question. For the exams discussed in this paper, $N=15$ and $n=4$. As discussed above, this rule is strictly proper. The constants a and b, although arbitrary, have been selected such that the maximum score on a 15-question exam is 100, and a uniform assignment of $(1/n, 1/n, 1/n, 1/n)$ will earn a score of zero. Under this normalization, a negative score implies that the student did worse than if she had no basis for favoring one answer over another.

On the first day of class we explain that the midterm and weekly take-home quizzes will be graded using the probabilistic scoring method. Students are told that a negative infinity on the midterm or any quiz will be treated as such; these students will either need to drop the class or will *earn* an F. The midterm is generally worth 20% of the final grade and the quizzes are worth 10%. We explain the grading system in detail during the first lecture so that a decision to take the class implies acceptance of this grading scheme. We have observed that some students choose to drop the class at this point, but do not know their underlying reasons for doing so.

All probability assignments are normalized to 1.0, in the event that the student's assessments violate this constraint. For example, an assignment of $(0.8, 0.1, 0.1, 0.1)$ would be normalized to $(8/11, 1/11, 1/11, 1/11)$, whereas an assignment of $(0.2, 0.1, 0.5, 0.1)$ would become $(2/9, 1/9, 5/9, 1/9)$. If a student leaves an answer blank, then any remaining probability is equally distributed among the

[4] Kerridge (1961) has referred to Equation (7) as a measure of "inaccuracy."

blank answers. For example, an assignment of $(-,-,-,-)$ would become $(0.25, 0.25, 0.25, 0.25)$. However, an assignment of $(-, 0.1, 0.1, 0.9)$ would become $(0, 1/11, 1/11, 9/11)$ because the student did not have any additional probability to distribute to the blank answer.

We do not formally prove that logarithmic is strictly proper, but instead assign a homework problem where students determine their optimal response for a binary question ($n = 2$). Specifically, we ask them to plot (in Excel®) their expected score as a function of their response on the answer they think is most likely to be correct. This reveals that their expected score is maximized when they set their response equal to their beliefs. We then note that questions with more than two possible answers can be thought of as a series of binary questions. For example, in the case of $n = 3$, we first assign probabilities to (a) and the union of (b) and (c). We know the best strategy in this binary setting. We then consider (b) and (c) and divide our remaining certainty among these two answers, which is again binary. For those students that want a proof, we refer them to Bickel (2007), which contains the proofs for all three rules.

The scoring system discussed here is wholly consistent with a course in decision making under uncertainty. The students' assignment $\mathbf{r}$ is a decision that requires careful consideration. Once they understand the scoring system and that their response should equal their beliefs (i.e., $\mathbf{r} = \mathbf{p}$), the exam becomes an exercise in probability assessment with students needing to assess $\mathbf{p}$. Because there is no notion of long-run frequencies, this assessment highlights the view taken in the course that probability is a statement of belief.

3.1. The Negative Infinity "Problem"

Some students worry about the possibility of earning a negative infinity—as if this is a random event that is not under their control. We ask the students, "Who *decides* what probably you assign to each answer? ... You do! So, if you are afraid of earning a negative infinity you can simply decide not to assign a zero." This being said, we have found it helpful to minimize the chance that students "accidently" assign a zero probability. The quizzes and midterm have a safety mechanism that the students may choose to employ. This is called the "safe harbor statement." This statement allows students to specify that any probability assignment of *zero* should be replaced by q, where q is set by the student. For example, a student may elect to set q equal to 0.001. In this case the assignment $(0.3, 0.3, 0.4, 0)$ would become $(0.3/1.001, 0.3/1.001, 0.4/1.001, 0.001)$. The student may alternatively have her probability assignments taken at face value. An analogy to rock climbing seems fitting; the student may choose to climb with or without a rope. Because the safe harbor statement only applies to probability assignments of zero, nonzero probability assignments of less than q are *not* replaced by q (i.e., students are not specifying a minimum probability assignment). This is again in the spirit of decision making. The student must still think carefully about her assignments. Continuing our climbing analogy, the safety keeps them only from killing themselves, not from getting severely injured. We believe the realization that decisions can have negative consequences should be a part of the class—before students are released into the real world. If a civil engineer designs a walkway (a series of decisions) that later collapses and kills over 100 people, she will face consequences significantly more painful than the prospect of failing a graduate course. Decisions have consequences, sometimes tragic ones.

Sometimes students will counter that if they "truly" believe a particular answer is impossible, they should assign zero because the optimal policy is to set $\mathbf{r} = \mathbf{p}$. To this challenge, we simply ask a series of questions that encourage self-reflection. Have they have ever been sure of something and then later proven wrong? Have they ever thought they "aced" an exam, but were later disappointed with the result? Have they ever transposed answers on a multiple-choice exam? Do they think that the wrong answers on the exam are generated by the instructor at random or are drawn from common and seductive mistakes? Perhaps more important for Bayesians is the concept of strict coherence, or Cromwell's rule (Dawid 1982, Lindley 1982), which states that a probability of zero should not be assigned to any possibility. This is important because in Bayesian analysis, the posterior distribution is proportional to the product of the prior and likelihood. If one assigns a categorical prior (a probability of 1 to one of the possibilities and 0 to the others), then no amount of evidence could ever change one's mind.

We suggest a degree of humility and encourage students to allow for the possibility that within the context of a timed exam, they could be making a mistake. We further suggest that they carry this perspective into their personal and professional lives.

3.2. The Issue of Risk Aversion

As discussed in §2, Q, S, and L are only strictly proper if students seek to maximize their expected score (i.e., they are risk neutral over course points). If students are instead risk sensitive, then they should seek to maximize some utility function over total course points. Bickel (2007) demonstrated that even in this case, students' assignments should be nearly proper as long as too much weight is not placed on any one question. Specifically, assume a particular student's utility function can be modeled as being exponential such that $u(P) = -\text{Exp}[-P/R]$, where R is the student's risk tolerance (measured in course points), and P is the total points he or she earns in the class. A positive risk tolerance implies that the student is risk averse. Recall from Equation (8) that $b = (100/15)/\ln(4) \approx 4.81$. As long as b/R is less than 7.5%, the student should not reduce her assessment more than 0.03 in an effort to hedge (Bickel 2007). We believe that 0.03 is a reasonable threshold, because students probably cannot assess their beliefs any closer than this.

The parameter b is under the instructor's control, whereas R is a characteristic of the student. Assessing R is difficult, but consider the following: Suppose at the end of class, a student has earned a total of 70 points out of a possible 100. We now offer this student a gamble where with probability p we will change their score to 100 and with probability $1-p$ we will reduce their score to 0. What probability p would make the student indifferent to accepting the gamble compared to his or her current score of 70? If the student, who we assume has an exponential utility function, replies 0.8, then his or her risk tolerance is about 100 points. If the student replies 0.99, then his or her risk tolerance is 15.75 points. To be conservative, assume that students' risk tolerances are 15.75, and therefore b must be less than 1.18 (0.075×15.75). For the 15-question, four-answer midterm, we use $b \approx 4.81$. However, this assumes that the midterm is worth 100% of the final grade. To hold b below 1.18 (in terms of total course points), we should not place more than about 25% (1.18/4.81) of the student's total score on the midterm, or no more than about 1.5% of his or her total points on any one question. As we discuss in §4, our results do not suggest that risk aversion is a significant issue.

4. Insights for Instructors

The issues discussed thus far turn a simple exam into an opportunity to teach fundamental concepts about decision making. In addition to this benefit for students, the grading scheme provides the instructor with a richer understanding of the students' mastery of course material. We will illustrate this by discussing the results for a single midterm exam involving 166 Stanford University graduate students.

The exam consisted of 15 questions with four possible answers each. The average probability assignment on the correct answer for each question is displayed in Figure 1.

Figure 1 shows that students had trouble with Problems 3, 5, 9, 10, 12, 14, and 15. In fact, the average assignment to the correct answer on Problem 3 was below 0.25, which would have earned a zero; the class' performance on this problem was worse than someone with no knowledge. The students would have been better off skipping this problem, which is what they might have done if they faced it the first day of class. Their learning (or their instruction) was negative!

Under an assumption that students are responding honestly, their response is equal to their belief, and their expected score is their negentropy (see §2.2).

Figure 1 **Average Probability Assignment to the Correct Answer**

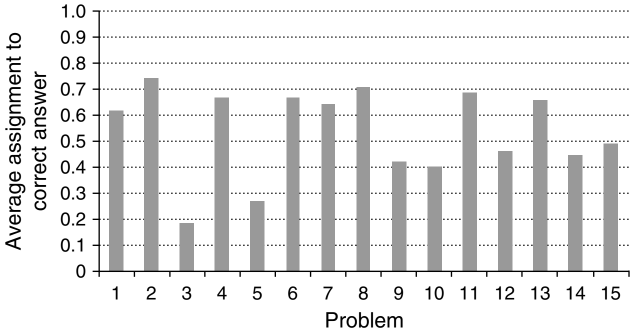
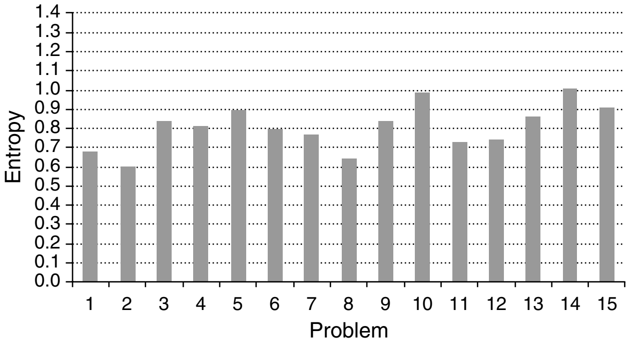


Figure 2 Average Entropy by Problem



Therefore, we can calculate the entropy of each student's probability assignment on each question. Figure 2 plots the average entropy of each problem (averaged over all students). This provides insight into how certain or uncertain the class was about a particular problem. Lower entropies imply that the class was more certain of a particular answer, but not necessarily the correct answer. The maximum possible entropy is $\ln 4 \approx 1.38$.

Figure 2 indicates that students were the most uncertain of the concepts covered in Problems 10 and 14. Although Problem 3 created uncertainty, it was not as high as one might expect based on the students' low assignment to the correct answer. This implies that students were more certain of a wrong answer, which can be seen in Figure 3. The darkest bar, d, was the correct answer, yet the class as a whole thought c was over twice as likely to be correct. At this point in our review of test results, we would discuss the specific concepts involved in Problem 3 and surface what students found attractive about c.

Figure 3 Probability Assignment to Each Answer for Problem 3

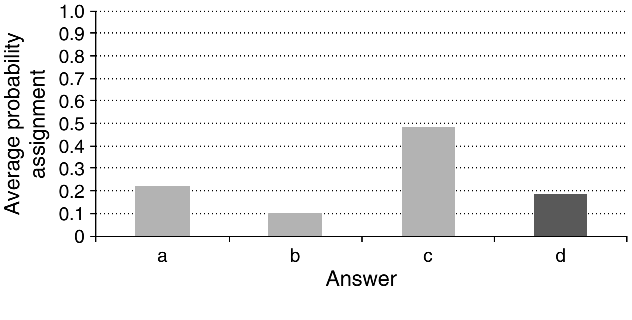
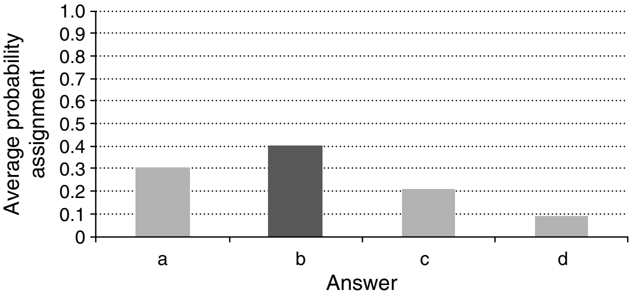


Figure 4 Probability Assignment to Each Answer for Problem 10

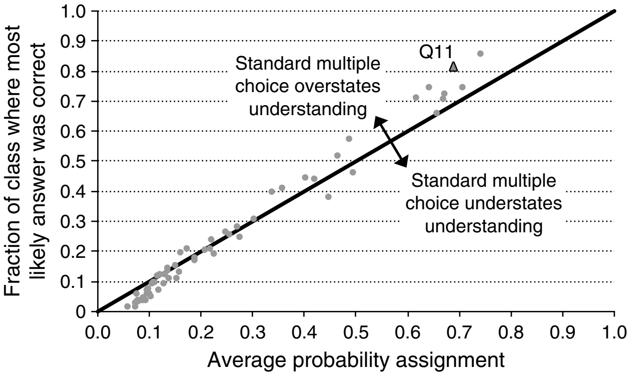


4.1. Comparison to Standard Multiple Choice Exams

Problem 10 had the second highest entropy, and its average probability assignment is shown in Figure 4. For this problem, whose answer was b, the class' average assignment was quite dispersed, with answers a and c attracting some attention. This insight may not surface with a traditional multiple-choice exam. For example, suppose all students held the beliefs shown in Figure 4. In this case, they would have all marked b, and the instructor would have no idea how poor their understanding really was.

To investigate this phenomenon more fully, we compared the fraction of students that would have marked the correct answer in a standard multiple-choice exam, assuming each marks the answer that he or she thinks is the most likely, to the average probability assignment on that answer. This is plotted in Figure 5. Although there is a correlation of 0.76, the discrepancy is instructive. For example, consider Question 11 (Q11), identified with the triangle

Figure 5 Comparison of Standard Multiple-Choice Results to Probabilistic Scoring Results



in Figure 5, in which answer b was correct. If students were simply asked to mark which answer they thought was the most likely, then 81% would have selected this answer, and the instructor might have believed that the class had mastered the underlying concept. However, the average probability assignment on this answer was only 69%, which implies a lesser degree of understanding. We see this behavior in the other questions as well. In fact, a standard multiple-choice exam would have overestimated student understanding over a wide range of beliefs.

4.2. Additional Insights for Students and Instructors

A more complete understanding of student test results is aided by two other decision analysis topics: probability assessment and the combination of expert forecasts.

4.2.1. Probability Assessment. Figure 6 compares five semesters of students' midterm scores (1,030 students) to the average entropy of their responses (averaged over 15 questions). Based on Equation (7), students that knew the material and assessed their state of knowledge well should have low entropy and receive a high score. The solid line is the maximum achievable midterm score given a particular entropy. This will obtain when a student believes one answer to be the most likely and the other answers to be equally unlikely. For example, a student that assigned 0.7 to one answer and 0.1 to the other three answers would have an entropy of $0.7 \ln 0.7 + 0.3 \ln 0.1 = 0.94$ and a maximum possible score of $100/15 + (100/15/\ln 4) \ln 0.7 = 4.95$. If they did this on each of the 15 problems, their average entropy would be 0.94, and their maximum possible score would be 74.

Some students had very low entropies and assessed their state of knowledge well. Even students with entropies around 0.9 (equivalent to a maximum assignment of about 0.72) still earned some of the highest marks because they assessed their more limited state of knowledge well. On the other hand, the student with the lowest average entropy scored only a 60 on the exam because he overestimated his knowledge on one or more problems. The lowest score (−50) was by a student who was very confident of his knowledge. As Mark Twain wrote, "What gets us into trouble is not what we don't know. It's what we know for sure that just ain't so."

Given the subjective view of probability taken in the class, how can we say someone is good at assessing probability? Although this is difficult to address for a single assessment, it can be partially addressed if one has access to many probability assessments, as we do in this case. The concept we use is referred to as *calibration*. If a probability assessor is well-calibrated, then a probability assignment of p should occur $p \times 100\%$ of the time. In the case of the midterm, we have analyzed the calibration and the students' probability assessments over five semesters, which includes 1,030 students and 61,800 probability assignments (1,030 students × 15 questions × 4 possible answers). The results are presented in Figure 7. We used a bin size of 0.05 to group probability assignments and treated all assignments between $p - 0.05$

Figure 6 Midterm Score vs. Average Entropy for 1,030 Students Over Five Semesters

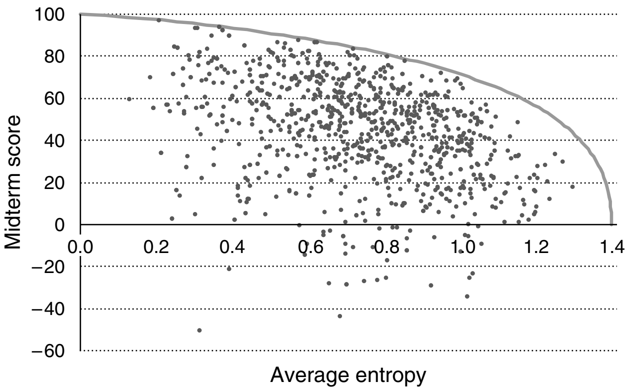
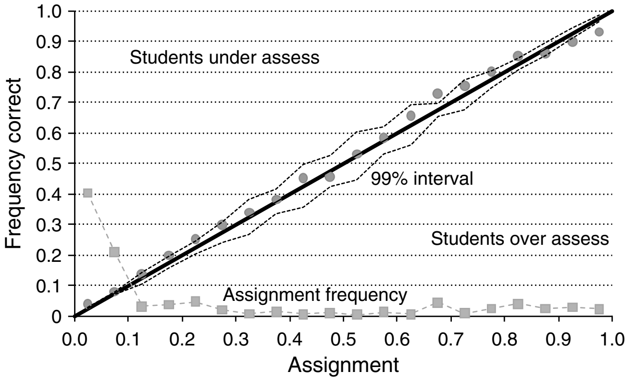


Figure 7 Calibration of Student Probability Assignments (61,800 Assessments)



and p ($0.05 \leq p \leq 1$) as an assignment of $p - 0.025$. For example, we treat all assignments between 0 and 0.05 as an assignment of 0.025. The results are presented in Figure 7. We added to this figure the frequency with which different assessments were given. For example, assessments between 0 and 0.05, represented as 0.025, were given almost 40% of the time. We notice local peaks at 0.675 and 0.825. Thus, students' tended to give assessments between 0.65–0.70 and 0.80–0.85 more frequently than other nearby probabilities.

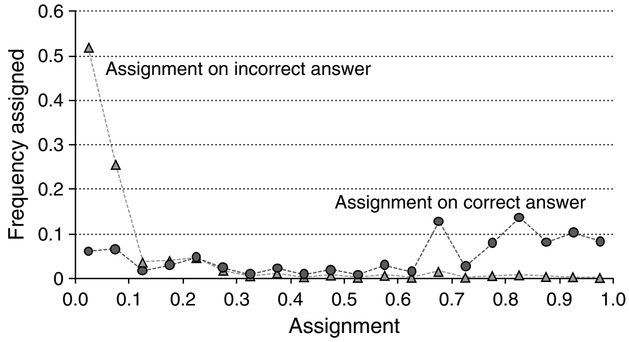
Based on the normal approximation to the binomial distribution, we establish a 99% probability interval around the line of perfect calibration. There is a 1% chance the observed relative frequency will lie outside this interval (0.5% chance of being above and 0.5% chance of being below). For example, if the probability an answer is correct is truly f, then there is a 99% chance that the actual relative frequency of correct answers will be within

$$f \pm \Phi^{-1}(0.995)\sqrt{f(1-f)B^{-1}}, \tag{9}$$

where Φ^{-1} is the inverse of the standard normal cumulative ($\Phi^{-1}(0.995) = 2.576$) and B is the number of probability assessments. The interval is not smooth because the number of assessments, B, at any particular probability is not constant. We see that the majority of the results are within this interval, with only assessments of 0.025, 0.675, 0.925, and 0.975 lying clearly outside. This performance is encouraging and demonstrates that students can reliably provide calibrated probability assessments.

The fact that the students' probability assessments are well calibrated serves to allay concerns regarding the affect of risk aversion and competition among students. However, we cannot completely rule out these concerns. Risk aversion would serve to move students' assessments closer to uniform, as they try to hedge (Bickel 2007). On the other hand, competition would serve to push the students towards categorical forecasts (Lichtendahl and Winkler 2007). Thus, these effects could be somewhat offsetting. However, as discussed in §3.2, we design the exam such that risk aversion should not (in a normative sense), be a significant factor. Whether or not Lichtendahl and Winkler's (2007) normative model of assessment competition is a good descriptive model is an open question.

Figure 8 **Student Likelihood Functions (61,800 Assessments)**



It is also interesting to investigate the students' likelihood function. In Figure 8 we present the assignment students made to the correct and incorrect answers. We see that students are very good at identifying wrong answers (they assign a low probability to them). Their ability to identify correct answers is not as strong, but still impressive. We again notice their predilection for assigning probabilities in particular ranges.

4.2.2. Combination of Expert Assessments. Another important topic in decision analysis is how to combine probabilistic assessments from multiple experts. We use the probabilistic scoring exercise to demonstrate this concept as well.

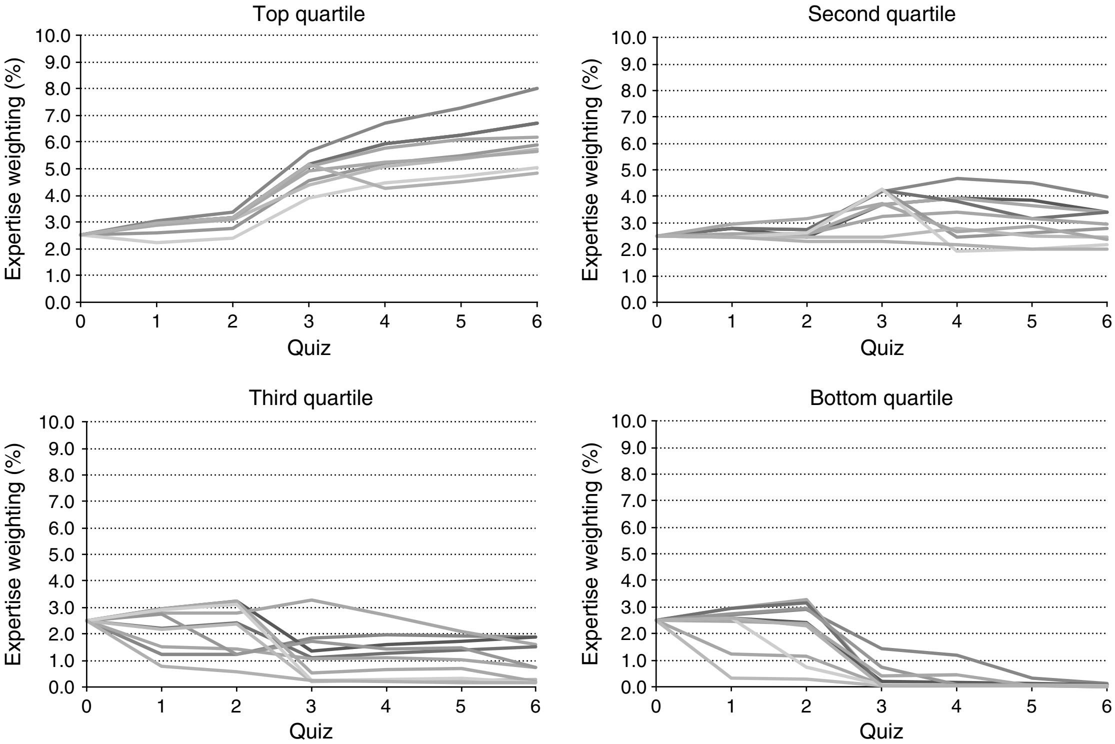
As mentioned in §3, we give students a weekly quiz containing a single problem that is graded in the manner discussed here. We begin by assigning student i a weight w_i that represents his or her degree of prior expertise. At the start of the semester, $w_i = 1/M$, where M is the number of students. If we interpret w_i as the probability that student i's probability assignment is the "truth," then we can use Bayes' rule to update the weights after each quiz (Roberts 1965).

Formally, let p_i be the probability that student i assigned to the *correct* answer. The posterior weight for student i is then

$$(w_i \mid p_i) = \frac{w_i \cdot p_i}{\sum_j w_j \cdot p_j}. \tag{10}$$

The denominator is the probability the instructor would assign to the correct answer based upon the prior weights. As is true of Bayesian analysis, the posterior depends only upon the probability assigned to the event that actually occurred (the likelihood) and

Figure 9 Dynamics of Student Expertise Ratings (40 Students Shown in Quartiles)



not events (or data) that might have been observed but were not (the likelihood principle). Thus, students seeking a good rating should seek to maximize their likelihood or the probability they assign to the correct answer. Because logarithmic scoring depends only upon the probability assigned to the correct answer, it is consistent with this strategy and may be used both to incentivize students to respond truthfully and to evaluate their performance though the use of likelihoods. As discussed in §2.1, the logarithmic scoring rule, being the only local scoring rule, is the only strictly proper scoring rule that satisfies these criteria (Winkler 1969).

After each week's quiz, we update the expertise rating for all students. The results of the first six quizzes, for a smaller class (recently taught at Texas A&M University), are shown in Figure 9.

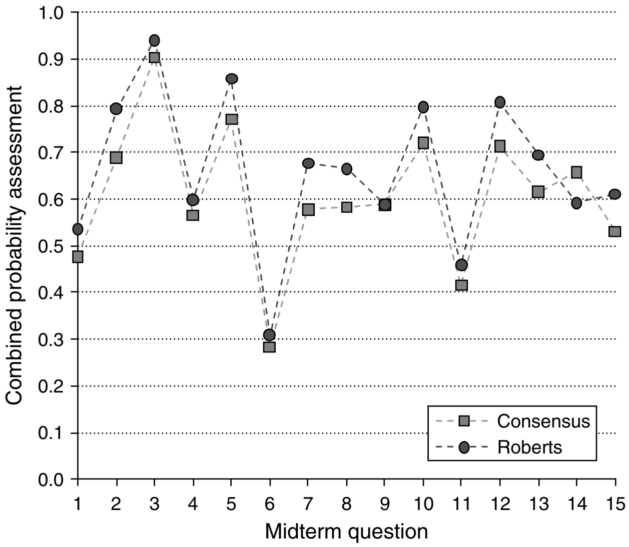
The range of expertise is quite broad, with one student's expertise weight increasing from 2.5% (1/40) to slightly above 8%. The worst-performing student's expertise weight dropped to 0.013%.

The intent of combining expert assessments is to arrive at a better forecast. The midterm takes place after the sixth quiz, at which point we discuss applying their expert weighting to develop an assignment on each of the midterm questions; that is, we use the Roberts model (Equation (10)) and multiply each student's expertise weighting going into the midterm by their probability assessment and sum over all students. The Roberts method is a linear opinion pool. Although we do not discuss linear opinion pools and other methods to combine expert judgments, other instructors could use this analysis as a jumping off point for these concepts.

We compare the Roberts weighting to a simple average of their assignments, referred to as the *consensus* assessment. The combined probabilities assigned to the correct answer for each midterm question are shown in Figure 10.

Figure 10 indicates that the Roberts forecast beat the simple consensus forecast in all but one question (Question 15) and tied in one case (Question 9). The average score on this midterm was 35. The consensus forecast would have earned 61, whereas the Roberts weighting would have earned a 68. A strategy of

Figure 10 **Comparison of Consensus and Roberts Assessments**



updating the weights after each midterm question would have earned a 74.

5. Conclusion and Suggestions for Future Research

Strictly proper scoring rules offer the opportunity to turn testing situations into rich learning opportunities. In this paper, we described the insights that can be obtained by students and instructors. These learnings reinforce essential decision analysis topics such as decision making, the meaning of probability, probability assessment, risk aversion, entropy, calibration, and combination of expert forecasts. In addition, probabilistic assessments provide instructors with a much richer understanding of class needs. Our work has also surfaced the following research needs:

- **Performance under different scoring rules.** Despite over 40 years of scoring rule use, we still lack a descriptive study that carefully explores how assessors respond to different scoring rules. For example, how are calibration results affected by the use of Q, S, or L scoring? We have assumed that because each of these rules is strictly proper, they all provide the same incentive to respond truthfully. This normative conclusion should be compared to descriptive reality.
- **Sensitivity to distance in practice.** As discussed in §2.1, there has been a resurgence of interest in designing scoring rules that are sensitive to distance or that can measure performance relative to a baseline distribution (e.g., see Jose et al. 2009). To our knowledge, research regarding how individuals respond to these rules in practice has not been published. For example, a paper that detailed the use of a distance-sensitive rule in an educational setting, where the instructor specifies the degree of correctness, would be most interesting. How do these results affect the quality of the assessments? Are the calibration results better than what is achieved under nonlocal rules?
- **Descriptive importance of competition among forecasters.** Competition among forecasters holds the potential of undermining the use of scoring rules, because in this case they are no longer strictly proper (Lichtendahl and Winkler 2007). Yet, our results suggest that competition is not a significant issue, even though students realize that they are in a competitive environment. How general is this conclusion? Have we just managed to avoid this concern because we stress the properness of logarithmic scoring and do not mention competition? Would our results be different if we made it clear to students that they are competing? To what degree does risk aversion offset the effects of competition?
- **Normative implications of competition under different scoring rules.** The research discussed above is focused on the descriptive implications of competition (i.e., does it impact the quality of the assessments). There is also additional normative work to be done regarding competition. For example, Lichtendahl and Winkler (2007) examined the normative implications of competition between two assessors under quadratic scoring. What are the implications of other rules? Are certain rules less sensitive to competition? For example, Bickel (2007) found that log scoring is the least sensitive to issues of risk aversion. Likewise, how does the inclusion of risk aversion alter the normative implications of competition? It should ameliorate the incentive to provide extreme forecasts, but to what degree?
- **Educational use of other scoring rules.** As mentioned at the outset, scoring rules are being used in several decision analysis educational programs. In some cases, the selected scoring rule is nonlogarithmic. Publishing the experience with these rules would deepen our understanding of the implications of different rules and serve as a useful balance to this paper, which is heavily focused on the use of log scoring.

In closing, our experience with the use of scoring rules over the last 15 years has been quite positive. Although students are initially worried about the new grading scheme, we find that with practice they overcome their fear and some even enjoy it. We hope that other instructors will adopt this scoring system and publish their results regarding its use.

Acknowledgments

The author thanks the reviewers, the associate editor, and the editor, L. Robin Keller, for very helpful comments and suggestions regarding a draft of this paper. The author also thanks Thomas Seyller for providing three years of Stanford scoring data.

References

Bernardo, J. M., A. F. M. Smith. 2000. *Bayesian Theory*. John Wiley & Sons, West Sussex, UK.

Bickel, J. E. 2007. Some comparisons between quadratic, spherical, and logarithmic scoring rules. *Decision Anal.* **4**(2) 49–65.

Bickel, J. E. 2009. Experiential learning and strictly proper scoring rules. *ASEE Annual Conf.: Paper AC 2009-2249, Austin, TX*, June 17.

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

Dawid, A. P. 1982. The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77**(379) 605–610.

DeGroot, M. H., S. E. Fienberg. 1982. Assessing probability assessors: Calibration and refinement. S. S. Gupta, J. O. Berger, eds. *Statistical Decision Theory and Related Topics III*. Academic Press, New York, 291–314.

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

Jose, V. R. R., R. F. Nau, R. L. Winkler. 2009. Sensitivity to distance and baseline distributions in forecast evaluation. *Management Sci.* **55**(4) 582–590.

Kerridge, D. F. 1961. Inaccuracy and inference. *J. Roy. Statist. Soc.* **23**(1) 184–194.

Kullback, S., R. A. Leibler. 1951. On information sufficiency. *Ann. Math. Statist.* **22**(1) 79–86.

Lichtendahl, K. C., Jr., R. L. Winkler. 2007. Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.* **53**(11) 1745–1755.

Lindley, D. V. 1982. The Bayesian approach to statistics. J. Tiago de Oliveira, B. Epstein, eds. *Some Recent Advances in Statistics*. Academic Press, New York, 65–87.

Roberts, H. V. 1965. Probabilistic prediction. *J. Amer. Statist. Assoc.* **60**(309) 50–62.

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

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Tech. J.* **37**(3) 379–423.

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. A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteorology* **9**(3) 360–364.

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

Winkler, R. L. 1968. "Good" probability assessors. *J. Appl. Meteorology* **7**(5) 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. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1) 1–60.