AC 2009-2249: EXPERIENTIAL LEARNING AND STRICTLY-PROPER SCORING RULES

J. Eric Bickel, The University of Texas, Austin

Experiential Learning and Strictly Proper Scoring Rules

Abstract

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.

Background

In several graduate industrial engineering / operations research programs (e.g., Stanford University, The University of Illinois at Urbana-Champaign, and The University of Texas at Austin), students face a unique grading system on their midterm exams, which are multiple choice. This grading scheme is also used in an undergraduate decision analysis course at Stanford University. 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. In theory, such an exam should better reveal the students' mastery of the subject, but how should the instructor assign scores in this situation?

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 to 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 different objectives in such a situation, ranging from learning the material to getting a good grade. We assume that letter grades are a strictly increasing function of the total points earned on the exam and that students seek to maximize their points. This simplification seems reasonable, particularly in programs that fractionalize 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 $\bar{R}(\mathbf{r} \mid \mathbf{p}) = \sum_i p_i R_i(\mathbf{r})$, 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

$$\mathbf{r}^* = \arg\max_{\mathbf{r}} \bar{R}(\mathbf{r} \mid \mathbf{p}) . \qquad (1)$$

If the scoring rule is linear such that $R_i(\mathbf{r}) \propto r_i$, then the optimal response is to assign 1.0 to the answer that student believes is most likely, which is the best strategy in traditional multiple-choice exams. Thus, students believing **p** = (0.85,0.15) and **p** = (0.51,0.49) would both assign **r** = (1,0) and receive the same score. Likewise, students believing **p** = (0.51,0.49) and **p** = (0.49,0.51) would assign **r** = (1,0) and **r** = (0,1), respectively, and receive very different scores-even though their knowledge is almost identical.

The insensitivity of students' scores to their knowledge is a major limitation of standard multiple-choice exams. These exams are not incentive compatible, in that they do not encourage students' responses that reflect their beliefs. A set of scoring rules is needed that encourages students to set their responses equal to their beliefs. Such a class of scoring rules does exist and is discussed next.

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 **r** = **r*** = **p**; that is, $\overline{\mathrm{T}}(\mathbf{r} \mid \mathbf{p}) < \overline{\mathrm{T}}(\mathbf{p} \mid \mathbf{p})$ for all $\mathbf{r} \neq \mathbf{p}$ and $\overline{\mathrm{T}}(\mathbf{r}^* \mid \mathbf{p}) = \overline{\mathrm{T}}(\mathbf{p} \mid \mathbf{p})$ when **r*** = **p**. [1-4] Many strictly proper scoring rules have been developed. Three of the most popular are given below.

Quadratic (Q): $\mathrm{Q}_i(\mathbf{r}) = 2r_i - \mathbf{r} \cdot \mathbf{r} \in [-1,1]$ (2)

Spherical (S): $\mathrm{S}_i(\mathbf{r}) = r_i / (\mathbf{r} \cdot \mathbf{r})^{1/2} \in [0,1]$ (3)

Logarithmic (L): $\mathrm{L}_i(\mathbf{r}) = \ln(r_i) \in (-\infty, 0]$ (4)

The range of possible scores differs considerably. For example, logarithmic scoring holds the possibility of an infinitely negative score. While 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. [1] In meteorological settings, the Brier score, which is a linear transformation of Q, is used extensively—making Q the most popular scoring rule.

Deciding among Scoring Rules

Shuford et al. [3] 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 and has two important implications. First, such a rule should be easier for students to understand. For example, a two-dimensional chart can be provided that details their score for any set of *n* assignments, which is only possible for other rules in special circumstances (e.g., the student assigns r_i to the correct answer and $(1-r_i)(n-1)^{-1}$ to each of the remaining statements). Second, because L is local, it will always assign a higher score where the student has assigned a higher probability 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. Students are likely to perceive such a result as being unfair. A related implication is that different scoring rules may generate different rank orderings among students for the same set of assessments. Bickel[5] closely examined the rank order properties of Q, S, and L in actual testing situations and found that Q and S performed poorly. L scoring will always reward the assignment of a higher probability to the correct answer.

In addition, 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[5] demonstrated that L is the least affected by this, which was surprising given that logarithmic scoring introduces the possibility of an infinitely negative score. We will discuss the issue of risk aversion and the negative infinity "problem" later in the paper.

Additional Properties of Log Scoring

Under L scoring, a student's expected score is

$$\bar{L}(\mathbf{r} \mid \mathbf{p}) = \sum_i p_i \ln r_i = \sum_i p_i \ln p_i - \sum_i p_i \ln \frac{p_i}{r_i} = -H(\mathbf{p}) - KL(\mathbf{p} \| \mathbf{r}), \tag{5}$$

where $H(\mathbf{p})$ is the entropy of $\mathbf{p}$ [6, 7] and $KL(\mathbf{p} \| \mathbf{r})$ is the Kullback-Leibler (KL) divergence between $\mathbf{p}$ and $\mathbf{r}$.[7, 8] The KL divergence will equal zero when $\mathbf{r} = \mathbf{p}$ and will otherwise be positive, reducing the student's expected score. This term measures the student's calibration (her ability to set $\mathbf{r}$ equal to $\mathbf{p}$) and encourages honesty because the expected score will be maximized when $KL = 0$. If the student does assign her true beliefs, then her expected score is simply the negative of the entropy of $\mathbf{p}$, or the student's negentropy. A categorical assignment of 0 or 1 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. The entropy term is said to measure the sharpness of the student's assignment. Thus, the use of L scoring has the property that students can only increase their score by improving their knowledge of the test material.

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:

$$\begin{aligned} L_i(\mathbf{r}) &= a + b \ln(r_i) \\ a &= 100 / N \\ b &= a / \ln(n). \end{aligned} \tag{6}$$

N is the total number of questions on the exam, and n is the number of possible answers. For the exams discussed in this paper, $N = 15$ and $n = 4$. As discussed above, this rule is strictly proper (see the appendix for a proof), which means that the students should assign their true beliefs such that $\mathbf{r} = \mathbf{p}$. 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 (0.25, 0.25, 0.25, 0.25) will earn a score of 0. Under this normalization, a negative score implies that the student did worse than if she had no basis for favoring one answer over another.

The graduate programs mentioned in the Background section all use the Logarithmic scoring rule. The undergraduate decision analysis course at Stanford University uses the Quadratic scoring rule. As mentioned above, and detailed by Bickel[5], we believe that log scoring is superior.

The grading scheme is explained on the first day of class, so that a decision to take the class implies acceptance of this scheme. We have observed that some students choose to drop this class at this point, but do not know the underlying reasons for doing so. To provide the students with practice and to build comfort with the grading system, students are assigned weekly take-home quizzes consisting of a single problem that are scored in the same way (thus $N = 12$ because the students are given 12 single-problem quizzes over the semester). 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 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.

Although the scoring system discussed here may not be appropriate for a course in pre-Renaissance European history, for example, it is wholly consistent with a course in decision making under uncertainty. The students' assignment **r** is a decision that requires careful consideration. Once they understand the scoring system and that their response should equal their beliefs, the exam becomes an exercise in probability assessment with students needing to assess **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.

The Negative Infinity "Problem"

In an effort to avoid any scores of negative infinity, 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 0 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 0, non-zero 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 significantly negative consequences should be a part of 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.

If a student "truly" believes that a particular answer is impossible, she should assign 0 because the optimal policy is to set **r = 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 been proved 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,[9, 10] which states that a probability of 0 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, 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.

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 much 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.

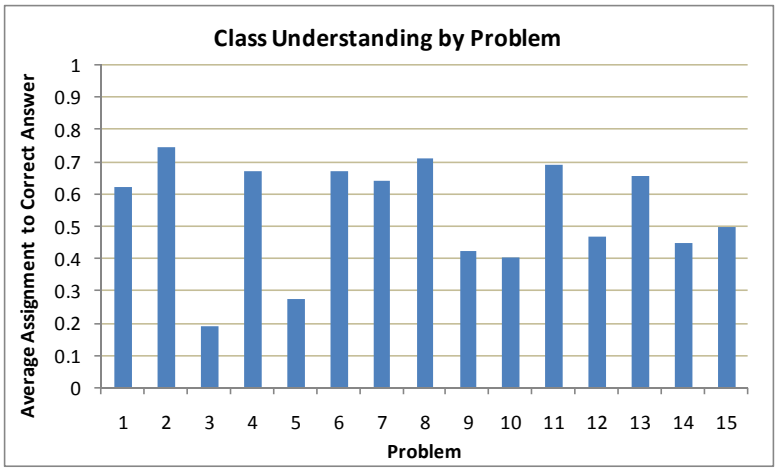

Figure 1: Average Probability Assignment to the Correct Answer

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 0. The class's performance on this problem was worse than if someone had no knowledge. The students would have been better off skipping this problem, which is what they might have done if they faced the problem on the first day of class.

As discussed above, 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). Lower entropies imply that the class was more certain of a particular answer, but not necessarily the correct answer. The maximum possible entropy is $\ln 4 \approx 1.38$.

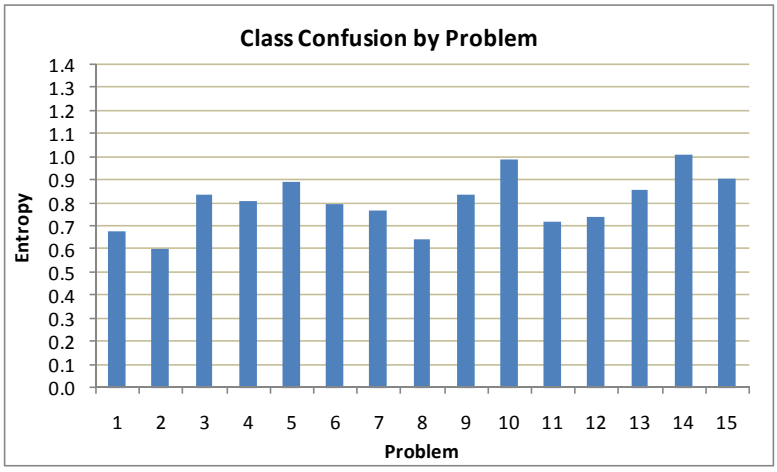

Figure 2: Average Entropy by Problem

Figure 2 indicates that students were the most confused by problems 10 and 14. While 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, as 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*.

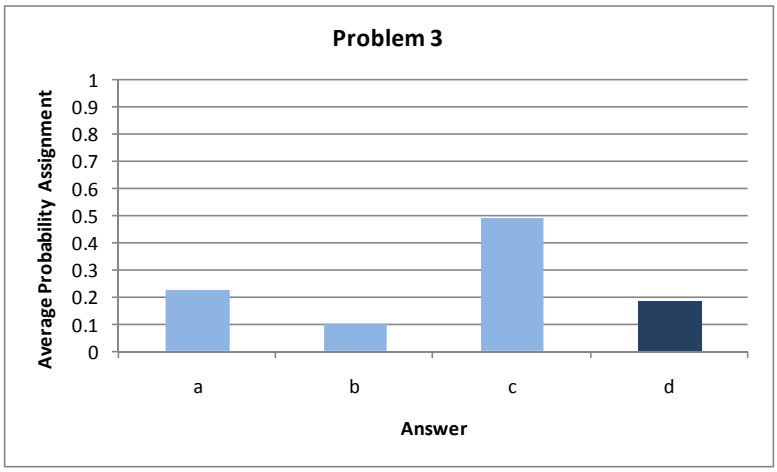

Figure 3: Assignment to each Answer for Problem 3

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's average assignment was

quite dispersed, with answers *a* and *c* attracting some attention. This type of 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.

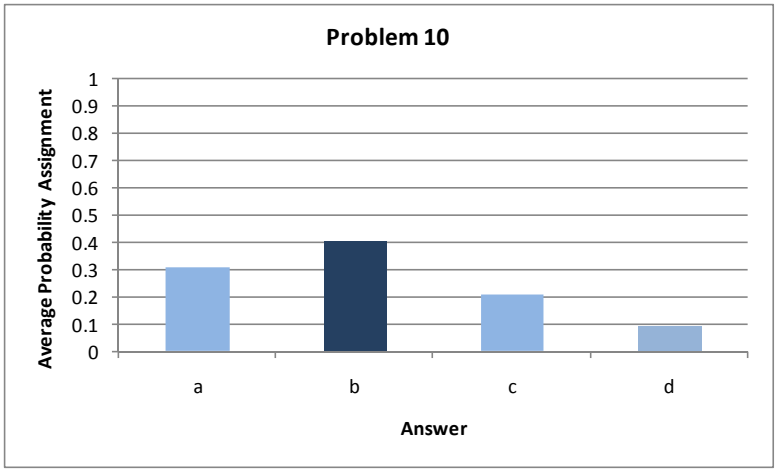

Figure 4: Assignment to each Answer for Problem 10

In order to investigate this phenomenon more fully, we compared the fraction of students that would have gotten the answer correct 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 high correlation of 0.76, the discrepancy between variables can be instructive. For example, consider problem 11, identified with the triangle 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 believe the class has mastered the underlying concept. However, the average probability assignment on this answer was only 69%, which implies a lesser degree of understanding.

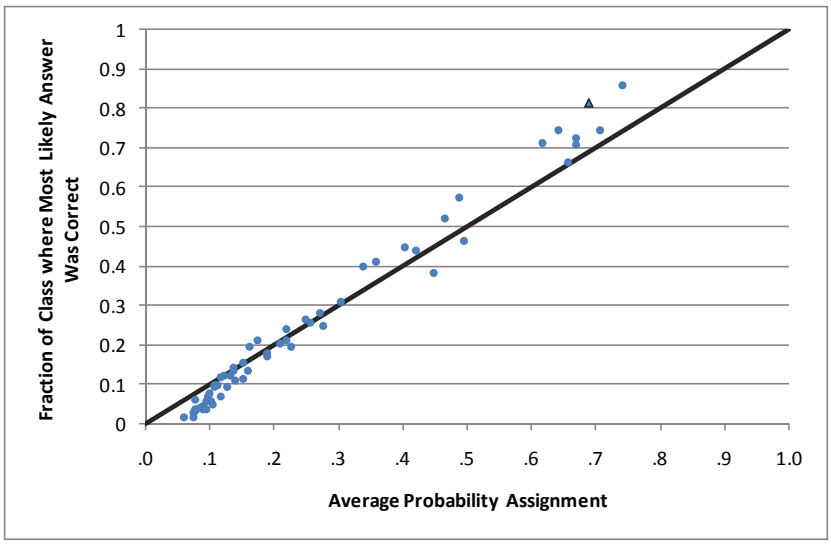

Figure 5: Comparison of Standard Multiple-Choice Results to Probabilistic Scoring Results

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.

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 (5), students that knew the material and assessed their state of knowledge well should have low entropy and receive a high score. The blue 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 $100/15 + \left(100/15/\ln 4\right)\ln 0.7 = 4.95$. 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 only scored a 60 on the exam because he overestimated his knowledge on one or more problems. The lowest score (-50) was by a student that was very confident of his or her 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."

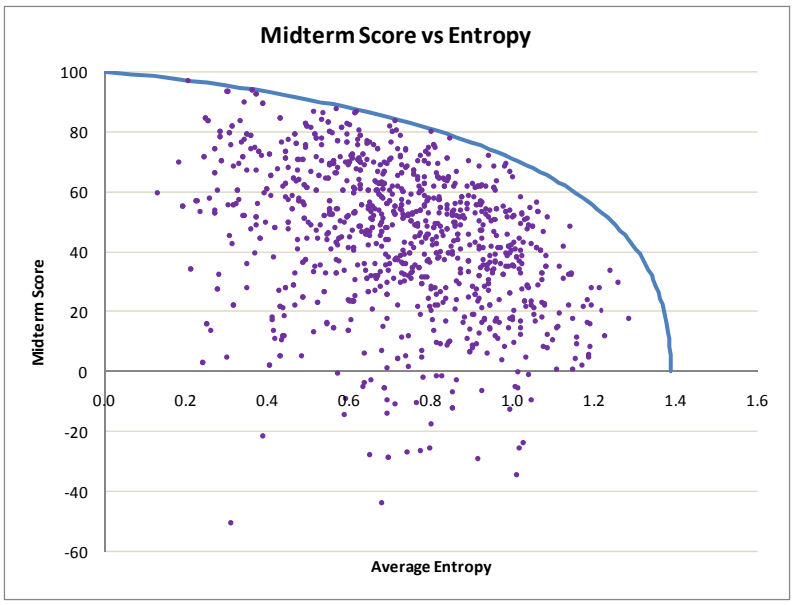

Figure 6: Midterm Score versus Average Entropy for 1,030 Students Over 5 Semesters

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 x 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 x 15 questions x 4 possible answers). The results are presented in Figure 7. We have used a bin size of 0.05 to group probability assignments and have treated all assignments between p and $p - 0.05$ ($.05 <= p <= 1$) as an assignment of p. The results are well calibrated, with only assessments above 0.9 being overaggressive. This performance is encouraging and demonstrates that students can reliably provide calibrated probability assessments.

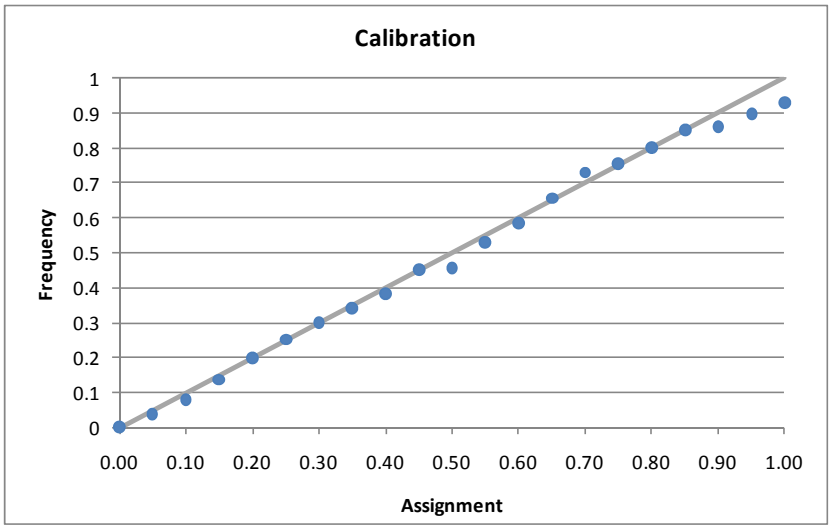

Figure 7: Calibration of Student Probability Assignments (61,800 assessments)

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 the Introduction, we give students a weekly quiz containing a single problem that is graded in the manner discussed here. We begin by assigning each student i a weight w_i that represents his or her degree of 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.[11]

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

$$(w_i \mid p_i) = \frac{w_i \cdot p_i}{\sum_j w_j \cdot p_j} . \tag{7}$$

The numerator 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 not events (or data) that might have been observed but weren't (a.k.a. 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. The logarithmic scoring rule, being the only local scoring rule, is the only strictly proper scoring rule that satisfies these criteria.[12]

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 8.

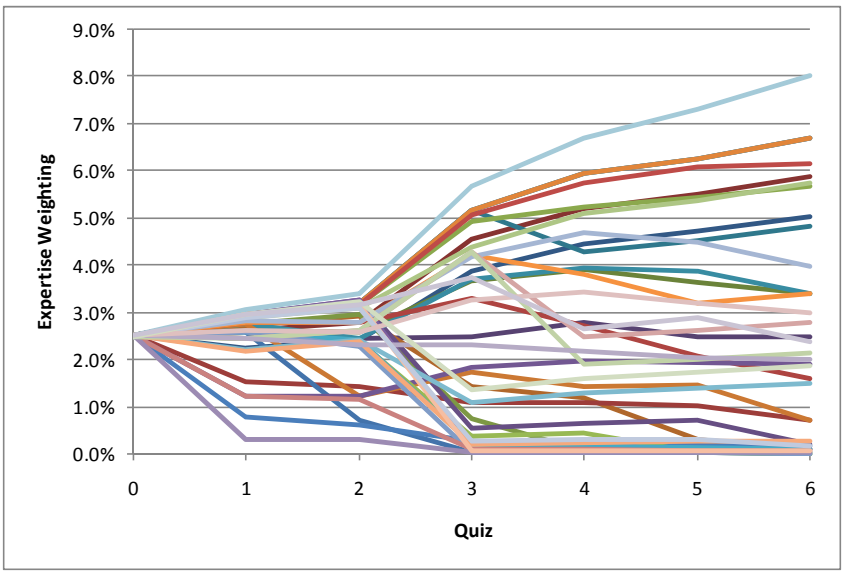

Figure 8: Dynamics of Student Expertise Ratings (40 Students)

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 has dropped to 0.013%.

The intent of combining expert assessments is to arrive at a better forecast. The midterm takes place after Quiz 6, at which point we discuss applying their expert weighting to develop a Bayesian assignment on each of the midterm questions. That is, we multiply each student's expertise weighting going into the midterm by their probability assessment and sum over all students. We call this the *static Bayesian* assessment because we are not further updating the weights based on the results of each midterm question. We compare this 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 9.

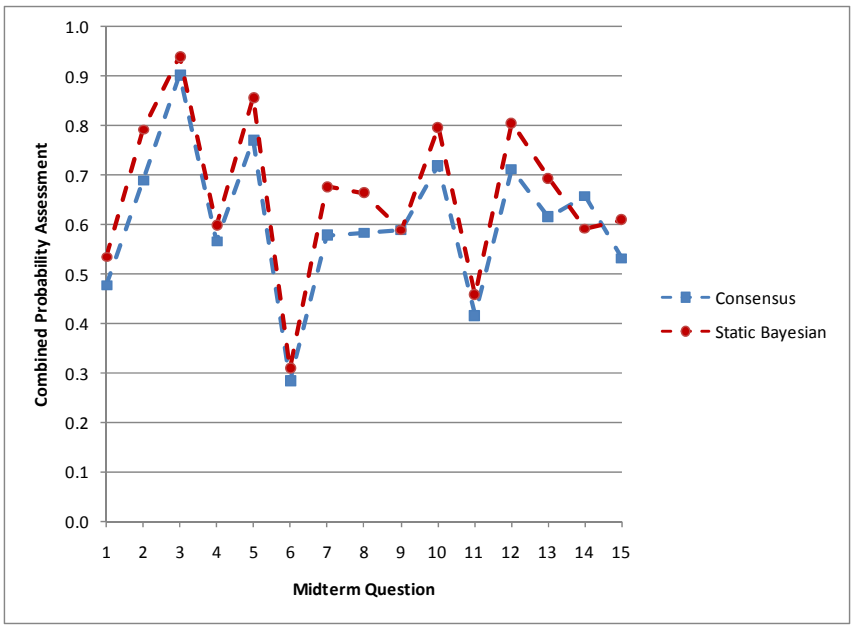

Figure 9: Comparison of Consensus and Static Bayesian Assessments

Figure 9 indicates that the static Bayesian forecast beat the simple consensus forecast in all but one question (question 15) and tied in one case (question 10). The average score on this midterm was 35. The consensus forecast would have earned 61, whereas the static Bayesian would have earned 68. A strategy of updating the weights after each midterm question, which we call the *dynamic Bayesian*, would have earned 74.

The Issue of Risk Aversion

As discussed in the section introducing strictly proper scoring rules, Q, S, and L are only strictly proper if students seek to maximize their expected score (i.e., they are risk neutral). If students are instead risk sensitive, then they should seek to maximize some utility function over total course points. Bickel [5] demonstrated that even in this case, students' assignments should be nearly proper as long as the instructor does not place too much weight on any one question. Specifically, assume a particular student's utility function can be modeled as being exponential such that $u(P) = -Exp[-P/R]$, where R is the student's risk tolerance and P is the total points he or she earns in the class. Recall from Equation (6) 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 to more than 0.03 in an effort to hedge. We believe that 0.03 is a good threshold, as students probably cannot assess their beliefs any closer than this.

b is under the instructor's control, whereas R is a characteristic of the student. Assessing R is difficult. 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 the 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 replies 0.8, then his or her risk tolerance is about 100. If the student replies 0.99, then his or her risk tolerance is 15.75. To be conservative, let us assume that students' risk

tolerances are 15.75 and therefore b must be less than 1.18 (0.075 x 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. In order 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. The results in Figure 7 suggest that risk aversion is not materially affecting our results.

Conclusion

Strictly proper scoring rules offer the opportunity to turn testing situations into rich learning opportunities. In this paper, we have 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 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.

Appendix

Proof that Logarithmic Scoring is Strictly Proper

The student seeks to maximize his or her expected score and solves the following program

$$\max_{r_i} \sum_{i=1}^{n} p_i (a + b \ln r_i)$$
$$st \quad \sum_{i=1}^{n} r_i = 1.$$

The Lagrangian is

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

The first order conditions are

$$\frac{\partial L}{\partial r_i} = p_i \frac{b}{r_i^*} - \lambda = 0 \Rightarrow r_i^* = p_i \frac{b}{\lambda}$$
$$\frac{\partial L}{\partial \lambda} = -\sum_{i=1}^{n} r_i + 1 = 0.$$

Substituting r_i^* into the second equation, we have

$$\sum_{i=1}^{n} p_i \frac{b}{\lambda} = 1$$
$$\lambda = b$$
$$r_i^* = p_i.$$

Bibliography

1. Toda, M., *Measurement of Subjective Probability Distributions*. 1963, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force: L. G. Hanscom Field, Bedford, MA.
2. Roby, T.B., *Belief States: A Preliminary Empirical Study.* Behavioral Science, 1965. **10**(3): p. 255-270.
3. Shuford, J., H. Emir, A. Albert, and H.E. Massengill, *Admissible Probability Measurement Procedures.* Psychometrika, 1966. **31**(2): p. 125-145.
4. Winkler, R.L., *"Good" probability assessors.* Journal of Applied Meteorology, 1968. **7**: p. 751-758.
5. Bickel, J.E., *Some Comparisons between Quadratic, Spherical, and Logarithmic Scoring Rules.* Decision Analysis, 2007. **4**(2): p. 49-65.
6. Shannon, C.E., *A Mathematical Theory of Communication.* The Bell System Technical Journal, 1948. **37**(3): p. 379-423,623-656.
7. Cover, T.M. and J.A. Thomas, *Elements of Information Theory*. Telecommunications. 1991, New York, NY: John Wiley & Sons.
8. Kullback, S. and R.A. Leibler, *On Information Sufficiency.* The Annals of Mathematical Statistics, 1951. **22**(1): p. 79-86.
9. Dawid, A.P., *The Well-Calibrated Bayesian.* Journal of the American Statistical Association, 1982. **77**(379): p. 605-610.
10. Lindley, D.V., *The Bayesian Approach to Statistics*, in *Some Recent Advances in Statistics*, J.T.d. Oliveira and B. Epstein, Editors. 1982, Academic Press, Inc.: New York, NY. p. 65-87.
11. Roberts, H.V., *Probabilistic Prediction.* Journal of the American Statistical Association, 1965. **60**(309): p. 50-62.
12. Winkler, R.L., *Scoring Rules and the Evaluation of Probability Assessors.* Journal of the American Statistical Association, 1969. **64**(327): p. 1073-1078.