Strictly Proper Scoring Rules in an Absolute Grading Environment – Preliminary Findings

Robert Dees, MS, United States Military Academy
Ken S. Gilliam, MS, United States Military Academy
Michael J. Kwinn, Jr. PhD, United States Military Academy

ABSTRACT: Strictly proper scoring rules are used to elicit a person's true probability beliefs about an uncertain outcome. The application of strictly proper scoring rules to grading in an academic environment is not new and is typically restricted to classes centered on Decision Analysis. For the purpose of explanation, a typical application of strictly proper scoring rules in academic grading would be as follows: assume that a multiple choice question with four possible answers has correct answer "D" and is worth one point. The traditional technique requires students to select one right answer, so if a student answers "D", the student receives a 1 or a 0 for all other answers. Conversely, a strictly proper scoring rule requires the student assign probabilities that each possible answer is correct, say A=0.1, B=0.2, C=0.05, D=0.65. The student's score depends on the scoring rule applied. Under the logarithmic scoring rule, the student would receive ln(0.65) points or -0.43. The scores are obviously bounded by $(-\infty, 0]$. Usually, the instructor rank orders students' scores and then assigns final grades. This situation can be extremely punitive for students who assign a low probability to a correct answer, and only slightly rewarding for those who submit their true understanding of the problem. Alternatively, the quadratic scoring rule allows a range of scores for the "correct" answer but is bounded between -1 and 1 allowing the instructor to similarly rank the scores. We discuss a modification of the quadratic rule applied at the United States Military Academy in our Decision Analysis course. In our approach, we are restricted to an absolute grading requirement - the grade a student earns is not curved in any way. We explore the trade off between information gained about the students' true beliefs and points awarded. We examine initial student feedback and compare probabilistic grades to the hypothetical traditional multiple choice grades. Finally, we explore options for integrating strictly proper scoring rules into other engineering courses.

Introduction

The mission of the United States Military Academy is "To educate, train, and inspire the Corps of Cadets so that each graduate is a commissioned leader of character committed to the values of Duty, Honor, Country and prepared for a career of professional excellence and service to the Nation as an officer in the United States Army".[1] During their four years of education at West Point, cadets learn the value of being bold, decisive leaders who are committed to action. What is often not as well learned however is the risk assessment associated with committing to the wrong course of action and the consequences therein. Quite naturally, cadets tend to apply the decisive action – and minimal risk assessment – they learn in a field training environment to their academic requirements. For most of these students, the real world will quickly manifest itself as a hostile environment in which a new platoon leader must weigh life or death situations laced with multiple levels of uncertainty. In our Decision Analysis course for Systems Engineering cadets, we aspire to make our students better assessors of probability and risk, and thereby better decision-makers in the face of uncertainty, through a series of challenging and thought provoking "probabilistic multiple choice" problem sets. Secondly, we aspire to gain more information about the state of our students' information regarding course material by having them respond to questions in a way that has more distinction than a binary response.

In an effort to make our students better assessors of probability, we have introduced the concept of probabilistic scoring rules, also known as Strictly Proper Scoring, in the Decision Analysis course. Essentially, this approach requires each student to solve a group of multiple choice problems and then assign a probability that each of the given multiple choice answers is correct. This method allows a student who is not confident in her answer to assign her true beliefs about the answer to the problem.

Although there is a correct answer, this scoring method allows students to trade a small portion of points to avoid losing all credit for a particular question.

In this paper, we begin by discussing the background of probabilistic scoring rules and then discuss the technical aspects of the approach. We then transition to our application, our assessment of the study to this point and then conclude with a discussion of the future directions of our study.

Background

Probabilistic scoring rules are used in a variety of ways. In the late 1960s, probabilistic scoring rules were introduced as a means for evaluating meteorologists' probability assessments on the weather.[2,3] Probabilistic scoring is used in the field of medicine to evaluate diagnosis of disease. Probabilistic scoring is used in the world of finance to evaluate market analysts' predictions. Recently, probabilistic scoring is used in the development of speech recognition software.

Probabilistic scoring rules applied in an academic environment are not new. Shuford, Albert, and Massengill began the discussion of probabilistic scoring in education in 1966.[4] Decision Analysis courses at Stanford and Texas A&M currently apply strictly proper scoring rules to many of their graded assignments. Most programs use the logarithmic scoring rule which allows a student to earn an infinitely negative score on any question, and theoretically fail an entire course over the smallest question. These other programs have the ability to rank order and subsequently assign a grade for the course. This ranking and grade assignment is counter to the guidance established by the US Military Academy's Dean of the Academic Board and as such, our application has been modified from this more drastic approach which we explain in greater detail later in this paper.[10] Regardless of the approach, the mathematical manipulations may seem unnecessarily complex for grading a simple homework. We explain these rules below and then follow with the explanation of the payoff in educational value for the increased calculation burden.

Probabilistic Scoring Rules

Consider an individual X who assesses a probability distribution over $n > 1$ mutually exclusive and collectively exhaustive events. Let $\mathbf{b} = (b_1,\ldots,b_n)$ be a vector of X's "true probability beliefs," where b_i is the probability that event i will occur. Let $\mathbf{r} = (r_1,\ldots,r_n)$ be a vector of X's "actual probability report," where r_i is the probability that event i will occur. In that the n events are mutually exclusive and collectively exhaustive, the sum of the probabilities $(b_1,\ldots,b_n)$ and $(r_1,\ldots,r_n)$ are both equal to 1. A scoring function S is **strictly proper** iff X's expected score is strictly maximized by setting $\mathbf{r} = \mathbf{b}$; that is, X's score is strictly maximized by reporting his or her true probability beliefs.[3,4,5,6] We note that assigning a uniform distribution over the n events equates to an admission of no information (or insight); under strictly proper scoring rules, it is better to admit that you have no information than to guess. This is a large departure from traditional multiple-choice scoring.

Many scoring rules have been developed, but three of the most popular (for n multiple choice questions) are:

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

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

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

where r_i is the probability assigned to the correct answer (i=1…n).[7]

As discussed by Bickel, the first thing to notice is that scoring rule L (equation 2, above) is defined as *local*, or that the assessor's score only depends on the probability assigned to the correct answer; a higher probability assigned to the correct answer will always result in a higher score. Locality is considered desirable by some because it should be easier for evaluated individuals to understand and it generates consistent rank orderings among assessors for the same assessments. Conversely, scoring rules Q (equation 1, above) and S (equation 3, above) are *global*, as the scores depend on both the probability assigned to the correct answer and the probabilities assigned to the remaining incorrect answers. Under these global rules, a reward is given to the probability assessed to the correct answer and a cost is deducted for probabilities assigned to incorrect answers. This implies that one assessor may assign a *higher* probability than another assessor to the correct answer but receive a *lower* score. This means that if individuals X and Y assigned [0.7,0.1,0.1,0.1] and [0.7,0.3,0,0] vectors respectively on an n=4 exercise with the first answer being true upon revelation, then X would receive a higher score even though they assigned identical probabilities to the correct answer. Individuals X and Y are equally rewarded for their assignment to the correct answer, but Y receives a larger penalty due to a concentration of probability assigned to a particular incorrect answer. In both cases, X and Y may have assigned their true probability beliefs. We believe that locality is desirable in situations where rank ordering results are important, and also recognize that an argument that the inclusion of probabilities assessed to both correct and incorrect answers with a global scoring rule also has merit. On a contextual level, the evaluator must decide whether to evaluate assessors locally or globally.[7]

Another consideration is whether or not the scoring rule is *bounded*. If an individual assigns a probability of 0 to the correct answer under scoring rule L, then the result is an infinitely negative score, from which the assessor can not recover. This essentially results in an expected value of $-\infty$ which increases the assessor's risk aversion. As scoring rule L results in only non-positive values, the evaluator must rank order the scores to assign positive scores (or the students would never do their homework at all!) This rank ordering and then "curving" or "shifting" the scores for grading may be less appealing to the evaluator who wants to score assessors according to an ex ante standard rather than an ex post rank. Finally, if an evaluator concludes that a negative score on any given assessment exercise is not acceptable, then L will not work without being truncated. However, If L is truncated (vice being unbounded below), then the scoring rule is no longer considered strictly proper. In contrast, both Q and S are bounded, and can easily be linearly transformed to any desired scale. We note that by definition a linear transformation of a strictly proper scoring rule is still strictly proper.[5]

USMA Approach

The primary objectives of the Decision Analysis course at the United States Military Academy are for the cadets to cover both single and multiple objective decision analysis as well as risk attitudes. We began early in the semester to train the students to be better assessors of probability through integration of a modified quadratic scoring rule. Our goals for using this system rather than a traditional multiple-choice method are: 1) Train students to be better decision-makers through probability assessment and 2) Provide the instructors with more information about each student's true level of understanding of the material.

We use a linear transformation of the quadratic scoring rule (*global, bounded*) which allows scores on individual questions to be between 0 and 5 points. There are three problem sets valued at 25 points each – so each problem set includes 5 questions, and each question has four possible answers. An example question is provided below in Figure 1.

3. The four elements of a decision situation are:

   a. Decision Trees, Influence Diagrams, Subjective Probabilities, and Risk Profiles
   b. Expected Values, Decisions, Uncertain Events, and Consequences
   c. Complexities, Uncertain Events, Multiple Objectives, and Differing Perspectives
   d. Values and Objectives, Decisions, Uncertain Events, and Consequences

Figure 1: Sample Problem Set Question

The score for any particular question is calculated by using the formula in equation 2, above or more specifically, equation 4 below.

$$2.5 + 2.5 \times Q_i(\mathbf{r}) \qquad (4)$$

where **r** is the vector of reported probability assessments, and r_i is the probability assessed to the correct answer. Note that $Q_i(\mathbf{r})$ (from equation 2, above) returns a score on the interval [-1,1]; using equation 4, we have linearly transformed this rule to return a score on the interval [0, 5]. Once again, a linear transformation of a strictly proper scoring rule is still strictly proper.[5] Similarly, the original interval can be transformed to the interval [0, 100] and interpreted as a percentage and any number of points can then be assigned to various questions.

Figure 3 depicts the possible score ranges for differing assessments on the correct answer. The fact that Figure 3 displays ranges of possible scores given the probability assigned to the correct answer is a result of the *global* property. It visually depicts how student *X* scores better than student *Y* even though they both assigned the same probability to the correct answer. Student *Y* incurs a larger cost for the distribution of a larger probability on a single incorrect answer in accordance with his or her true beliefs. To attain the absolute maximum score, the student must assign a probability of 1.0 to the correct answer, and conversely, to attain the absolute minimum score, the student would assign a probability of 1.0 to any of the incorrect answers. Both of these techniques equate to approaching the problem set as a traditional multiple choice exercise when a student can choose only one right answer.

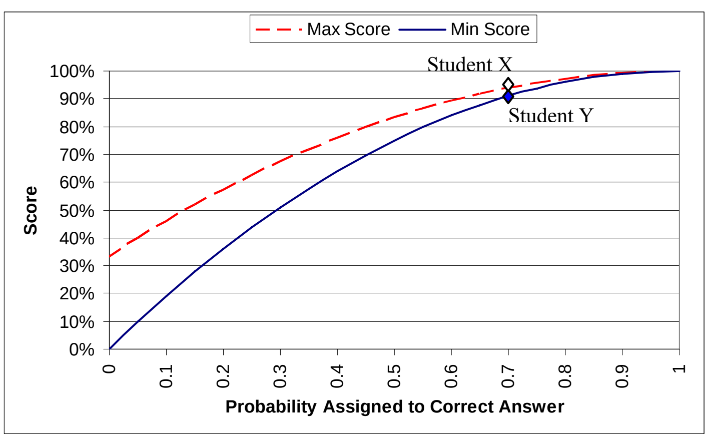

Figure 3: Possible Scores for Student *X* and Student *Y*

This approach has several features that we find desirable. First, it allows us to establish, publish, and score against an ex ante standard rather than using a student's ex post rank to determine grades. This means that a student knows where they stand in the course as soon as they receive the solutions and scores rather than waiting until the end of the course to see their ranking. In line with current research on effective teaching, we have avoided a grading system that puts students in competition with their classmates and we keep students informed of their progress throughout the term.[8]

Second, if a student is uncomfortable or ignorant about this grading system, they can still use a multiple-choice approach by answering with nothing other than 1s and 0s. In our in-class explanations and demonstrations, we advise them that this does not maximize their expected score; we use this to advocate assigning their true probability beliefs. We also show them how this allows the student to receive partial credit on a multiple choice type of question.

Third, this methodology does not produce negative scores. We believe that the possibility of a negative score on any particular problem increases the level of risk aversion in some students. We want to foster a risk neutral attitude in our students' approach to our problem sets. In doing so, we recognize that some will actually act in a risk seeking manner, but we have found that it is harder to convince our students out of risk aversion than it is to convince them out of risk seeking behavior. This discussion also reinforces the fact that the best strategy is to assign true probability beliefs.

Finally, we reward an admission of ignorance with a 62.5% score; this equates to a "high F" on our scale. A student attains this score by assigning equal probabilities to all possible answers. This reinforces the principle that it is better to admit ignorance than to feign understanding. We, the instructors, get more information about our students' state of information as it relates to course material.

Assessment

We have used two tools to assess our approach: the student scores and a brief student survey. The student scores provide a means for hypothetical comparisons between different scoring rules and the opportunity to explore the advantages and disadvantages for students under each rule. The student survey provides insight into student awareness, motivation, and risk attitudes concerning the first problem set administered. (Note that previous editions of this course did not have similar problems sets and thereby making direct comparisons impossible.)

Scores on the initial problem set averaged 75%. As a part of their submission, students were required to also submit their "total commitment" answer – that is, the student had to pick one and only one correct answer. This was used to calculate a hypothetical score under traditional multiple choice conditions. If scored in the traditional multiple choice manner, the course average would have decreased to 70%. More interesting yet, only 16 of the 74 students chose to answer every question as if it were a traditional multiple choice environment, and only 3 of the 16 achieved 100%. In comparison, 41 students realized an improvement in their grade for the assignment over a traditional multiple choice environment, and only 15 experienced a reduction in their score. This includes 2 students who received zero credit for problems on which they assigned probabilities whose sum exceeded 1. If we remove the students whose all-in answers do not match their assigned probabilities, then the maximum points lost on a 25 point problem set was 0.9625, or 3.85% of the assignment.

We collected student feedback after the first problem set but before any student had received their grade for the event. 67 of the 74 students completed the 10 question survey which attempted to assess the students' attitude towards the scoring rule, their perception of their grade, and some brief questions to assist with future measurements of risk attitudes.

There were 16 students that scored the same when comparing traditional multiple choice scoring and our scoring methods. Of those, only 11 indicated that they believe their scores would be the same. This shows a misunderstanding or ignorance of how the scoring rule is calculated. Of the 67 respondents, 63 (94%) predicted their scores would be within +/- 10% of traditional multiple choice scoring rules, but only 30 of 67 (45%) were accurate in predicting how the probabilistic scoring rule would affect their grades. Additionally, 59 of the 67 (88%) respondents indicated that they are indifferent or prefer probabilistic multiple-choice over traditional multiple-choice. Also of note, 63 of 67 (94%) respondents stated they spent the same amount of time or longer on this assignment than they would have if it were scored in a traditional multiple-choice manner.

Our most interesting findings concern the information gained by the instructors. When students choose to answer with anything other than assigning a probability of one to an answer, the instructor gains some piece of information about the student. Since the points we are willing to give cost us no more than the computing power necessary to accurately calculate a score, there is virtually no investment on the instructors' part. For that minimal investment instructors can learn about each student as long as each answers with their true beliefs. The probabilities assigned to both correct and incorrect answers give us a better fidelity about the current state of our students' information. This reveals where the students as a whole could use improvement or review of material. We aim to gather more data before we quantify the level of information gained relative to traditional multiple-choice scoring.

Future research

We believe that our scoring rule has a valid application in our Decision Analysis course. It can also be leveraged in other engineering courses to elicit the true level of understanding of students. Initial student feedback is positive, with some skepticism mixed in as well. The students continue to improve their ability to assess their own understanding of probability and the uncertainties they face. We believe this understanding of probability and uncertainty is applicable in all areas of engineering education.

Possible future research will focus on several areas. Our ultimate goal is to improve each student's ability to assess uncertainty and apply that improved ability to the decision situations in their everyday lives. We intend to continue soliciting feedback from students in several areas and looking for significant relationships that may improve the quality of instruction over the next several years. We plan to evaluate the relationships between learning styles, risk attitudes, and probabilistic scoring rules. We also will assess students' performance based on course objectives and their approach to probabilistic scoring rules. We will also continue to pursue opportunities to include probabilistic scoring rules in other courses at West Point. We believe there is merit in exploring the possibility of finding a strictly proper scoring rule that is both local and bounded. We also hope to compare the accuracy of multiple probability assessors with various states of information as compared to an individual expert assessor.

Conclusion

Every decision situation requires the decision maker to consider four elements: the decision to be made, uncertain events, possible consequences, and values and objectives.[9] We have explicitly focused this paper on the uncertain events, but have encouraged the incorporation of the other three elements by allowing an infinite spectrum of possible outcomes and requiring each student to weigh their values and objectives against those uncertain events and consequences. By doing so, we hope to build a cohort of future leaders more aware of the uncertainties affecting their decisions and the ramifications of their bold commitment to action. We do not attempt to strip away the bold and decisive nature; rather we strive to augment the deft commitment to action with an ability to recognize the uncertain nature of future events and mitigate the risk of bad outcomes.

1. United States Military Academy, *USMA Mission*; available from http://www.usma.edu/mission.asp; Internet; accessed 24 February 2008.
2. Murphy, A. H. 1969. On the "Ranked Probability Score." *J. Appl. Meteorology* **8** 988-989
3. Winkler, R. L. 1968. "Good" probability assessors. *J. Appl. Meteorology* **7** 751-758.
4. Shuford, E. H. Jr., A. Albert, H. E. Massengill. 1966. Admissible probability measurement procedures. *Psychometrika*. **31**(2) 125-145.
5. 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.
6. Roby, T. B. 1965. Belief states: A preliminary empirical study. *Behavioral Sci.* **10**(3) 255-270.
7. Bickel, J. E. 2007. Some comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules. *Decision Analysis* **4**(2) 49-65.
8. Davis, B. G. 1993. *Tools for Teaching.* Jossey-Bass publishers, San Fransisco.
9. Clemen, Riley, *Making Hard Decisions*, 21.
10. United States Military Academy, "Grading Philosophy," *Academic Program*; available from http://www.dean.usma.edu/sebpublic/curriccat/static/AcademicProgram.htm#GradingPhilosophy; Internet; accessed 24 February 2008.

Biographical Information

Rob Dees is an Instructor and Analyst in the Department of Systems Engineering at the United States Military Academy. He is an instructor for two courses, *Statistics for Engineers* and *Decision Analysis.* Rob received his BS in Engineering Management from USMA in 1998 and his MS in Industrial and Systems Engineering from Texas A&M University in 2005.

Ken S. Gilliam is an Instructor and Analyst in the Department of Systems Engineering at the United States Military Academy. He is the course director for two courses, *Statistics for Engineers* and *Decision Analysis.* Ken received his BS in Environmental Engineering from USMA in 1994 and his MS in Operations Research from the Georgia Institute of Technology in 2003.

LTC Michael J. Kwinn, Jr., Ph.D. is an Associate Professor of Systems Engineering at the United States Military Academy at West Point where he currently directs the Systems Engineering and Operations Research programs. Mike graduated from the United States Military Academy at West Point in 1984 and was commissioned as a Second Lieutenant in the Air Defense Artillery. He has been stationed at Fort Bliss, Texas, Germany, Fort Carson, Colorado and Camp Stanley, Korea. He has received a Master of Science from the University of Arizona, a Master of Arts in National Security and Strategic Studies (with Distinction) from the Naval War College, and a PhD in Management Science from the University of Texas at Austin. He has worked on systems engineering projects for over 10 years and recently served as the Director of the Operations Research Center (ORCEN) at the United States Military Academy. Some of his recent work is in the areas of acquisition simulation analysis, military recruiting process management, condition-based maintenance implementation and assessment systems development for combat operations. He is President-elect for the Military Operations Research Society (MORS), is a member of the International Council on Systems Engineering (INCOSE), the Institute for Operations Research and Management Science (INFORMS) and has served as an advisory member for the Army Science Board. LTC Kwinn is married to LTC(R) Brigitte Kwinn (USMA '84) and has four daughters: Cheryl (21) who is a junior at the Massachusetts Institute of Technology, Jasmine (5), Jade (4) and Emerald (2). His son Michael passed away five years ago and Mike is the President of The Friends4Michael Foundation named in memory of his son.