A Normalized Scoring Model for Law School Competitions

Edward K. Cheng & Scott J. Farmer

1. Introduction

The structure of moot court competitions, particularly at the "round robin" stage, commonly generates concerns about consistency. Teams face different judging panels, and panel composition frequently changes based on the day and time availability of judges. In short, few teams ultimately face the same group of judges. And given that the judges are drawn from disparate populations (faculty, practitioners, students) and bring a diverse set of reference points to the competition, fairness concerns quickly enter the discussion. Did Team A win because it drew a particularly generous set of judges? Did Team B lose because it drew a particular harsh one?[1]

Ed Cheng is a Professor of Law at Vanderbilt Law School and a doctoral candidate in the Department of Statistics at Columbia University. Scott Farmer is a 2013 graduate of Vanderbilt Law School, and was a member of the 2012-13 Managing Council of the Vanderbilt Moot Court Board.

[1] One could argue that in this situation, the scoring is still "fair" in the sense that each team has the same expected score or has judges randomly drawn from the same distribution. In reality, however, participants often consider this additional random factor to be "unfair," and as a statistical matter, it nonetheless represents variation that one should seek to control or limit.

Moot court organizers conventionally address consistency concerns through detailed scoring rubrics and strong admonitions (or wishful pleas) to follow them. These rubrics, however, are unfortunately difficult and unpleasant to follow. Most judges have limited prior exposure to the scoresheets and lack the time or interest to study them. In addition, during arguments, judges are principally focused on the participants and their legal arguments. Paying additional attention to the participants' poise, decorum, citation of precedent, and other technical details poses an unenviable multitasking problem. Consequently, scores become invariably impressionistic, allowing a judge's personal scale to taint the scoring process.

Ideally, one would like some mechanism for normalizing judges' scores. Such a mechanism would take the raw scores and adjust them based on a judge's proclivities (overly generous, typical, or overly harsh). But how to perform such an adjustment is far from obvious for organizers. In extreme cases in which one judge's scores deviate wildly from all others, the competition organizers might drop the score, but such a procedure is entirely ad hoc, raising other fairness concerns.

Another possible procedure is to drop the highest and lowest scores and average the remainder, a method reminiscent of some Olympic events.[2] This method, however, is both over- and under-inclusive: It throws away substantial amounts of information at the high and low ends, potentially unnecessarily. At the same time, while it handles extreme outliers handily, it does not adjust for the possibility that a participant randomly draws an especially generous or harsh group of judges (as opposed to a lone maverick).

Although the focus in this Article is moot court scoring, one can envision many other instances of law school assessment in which such a normalization problem arises. Law review competitions also

[2] E.g., Samantha Ashworth, How to Score a Dive, Chicago Tribune, Aug. 2012, at www.chicagotribune.com/media/acrobat/2012-08/71557861-02205441.pdf (describing Olympic diving scoring procedure). Although the "drop-highest-and-lowest" procedure addresses concerns about nationalistic bias, Olympic competitions do not always face the same consistency problem seen in moot court competitions, since competitors in a given round still face the same set of judges.

involve different sets of graders, whose subjective determinations must be reasonably commensurate to make fair comparisons. Even more intriguing, although presenting a more complicated problem, law school grades suffer the same normalization concern. Courses feature material with different degrees of difficulty, attract different pools of students, and are taught by different instructors. Yet, class rank and graduation honors are ultimately calculated under the assumption that all grades are commensurate.

In this Article, we propose a statistical method for normalizing scores in the moot court context. By making some reasonable assumptions, we can model each judge's propensities, allowing us to ultimately rank participants on a standardized scale. The method is quite general, accounting for not only judges, but also the procedural posture of the case – for example, whether the participant argued for petitioner or respondent, or whether the participant was on-brief or off-brief. Part 2 provides some basic background on the Vanderbilt Moot Court competition for which the model was developed. Part 3 elaborates on the normalization problem. Part 4 introduces and details the statistical model. Part 5 applies the model and reports the results. Part 6 discusses the model's implications and possible future extensions. The Article then briefly concludes.

2. BACKGROUND

Like most law schools, Vanderbilt offers an intramural moot court competition ("Intramural Competition"), which is run by its Moot Court Board. The Intramural Competition is a voluntary activity open to all second- and third-year students at Vanderbilt. In addition to awards and cash prizes for the top teams, the competition also determines successor members of the Moot Court Board.[3] Historically, approximately 75 percent of the second-year class par-

[3] Thirty second-year students are chosen annually to become members of the Moot Court Board. Members from the eight teams who make the quarterfinals automatically qualify for membership. The final fourteen spots are given to the remaining individuals with the highest individual scores. The individual score combines the team's brief score with the individual's oral scores. See *infra*.

ticipates in the competition.[4] Each moot court team comprises two students, who submit a joint brief and then divide up the case for oral argument.

All competitors participate in a preliminary "round robin" competition. During these preliminary rounds, teams argue the two sides of the case against different opponents. Teams are then ranked based on their brief and oral scores, which are weighted equally. The top 16 teams advance to a single elimination tournament, ultimately leading to the championship. Because the tournament is head-to-head, it does not raise the score normalization problem, and thus we will not discuss it further. Our focus will be on the preliminary round scoring only.

Three-judge panels score the oral arguments. Typically, panels consist of one faculty member, one local practitioner, and one member of the Moot Court Board. Each judge gives each competitor a score out of 100, which is broken down into ten categories worth between five and 20 points. The judges are provided with general scoring instructions along with scoring guidelines for each category.[5]

Brief scoring is divided into "subjective" and "objective" components. The subjective score rates the brief on its explanation and analysis of the problem, its overall persuasiveness, and its use of precedent, facts, and policy arguments. The objective score rates the brief on largely technical items, such as spelling, citation format, tone, stylistic issues, and its adherence to competition rules. Briefs are scored three times on the subjective criteria and twice on the objective criteria by Moot Court Board members. Because of the large number of briefs, however, the task is spread among many Board members, meaning that briefs will invariably be assessed by different sets of scorers.

---

[4] Students are allowed to compete in the Intramural Competition once, either during their second or third year of law school. However, because third-year students cannot advance beyond the preliminary rounds, a vast majority of students participate during their second year.

[5] The finals deviate from this structure. For example, the 2012 Finals were judged by three federal appellate judges, who were free to determine the winner however they wished.

3. THE PROBLEM OF NORMALIZATION

From the outset, governing members of the 2012 Moot Court Board had concerns about scoring consistency in the Intramural Competition, particularly with respect to oral argument scores. The competition typically involves large numbers of participants and judges interacting over the course of two preliminary argument rounds (one on-brief and one off-brief) held over a two-week period. As a consequence, the composition of judging panels varies considerably.

Traditionally, the Board facilitates consistency by providing judges with detailed information and structuring their scoring. Judges are given a comprehensive Bench Brief to introduce them to the issues and possible arguments, as well as the relevant case law. Then, scoring is broken down into many specific categories, and judges receive guidelines defining what a particular score means in a particular category. For example, a competitor's "road map" during oral argument is given a maximum of five points, as seen in Figure 1.

FIGURE 1: EXCERPT OF A MOOT COURT SCORING RUBRIC

> 0-1: Offers a limited introduction. Does not mention party's position and/or basic arguments. **First oralist on team only:** Does not identify self, co-counsel, or client.
>
> 2-3: **Average.** Provides adequate introduction but relies on notes and/or leaves out an element of the "road map."
>
> 4-5: Begins with "May it please the court . . ." **First oralist on team only:** Identifies self and co-counsel, identifies client, quickly summarized the party's basic position. **Both oralists on team:** Gives a preview of party's basic arguments with very limited reliance on notes.

In theory, such a strategy cabins discretion, minimizing impressionistic scoring and promoting consistency. In practice, however, many judges found the detailed rubric frustrating, confusing, and otherwise unmanageable. One of the authors (Cheng) found it nearly impossible to keep track of all the scoring considerations while

also listening to the substance of the argument and engaging in a thoughtful dialogue with the participants. At least one alumnus refused to follow the rubric at all, opting to provide only a total score based on his overall impression. Still other judges expressed frustration over how the rubric's complexity made it difficult for them to tell which team "won" the round.

Even for judges who navigate the multitasking challenge in good faith, the conventional strategies offer limited success. Judges simply disagree on the definition of "average," a problem exacerbated by the fact that many judges do not see sufficient competitors to calibrate their impressions. Scoring also tends to be comparative among the four participants immediately before the judge in the given round, as opposed to absolute. Worse yet, within this comparative framework, some judges distinguish competitors using a small difference in points, while others are more dramatic. Finally, some judges, whether because of their practice area (for practitioners) or their involvement in the construction of the problem (for Board members), simply know more about the issues involved in the case, leading to a more critical assessment of a competitor's knowledge.

Rubrics are therefore only a partial solution to the scoring consistency problem. To address the problem fully, adjustments must be made to the scores themselves. Unfortunately, the most natural procedures, such as dropping the highest and lowest scores, or adjusting scores based on a judge's mean, have some serious drawbacks. Dropping the highest or lowest scores discards scoring data in a context in which such data is limited and labor-intensive. Adjusting scores purely based on mean judge scores requires that judges score a large pool of participants; otherwise, chance variations in whom the judge observes can result in overcorrections. For example, if a judge happens to observe three participants from the bottom 25%, the adjustment procedure will overinflate those scores.

What we would like is to normalize the scores while minimizing these drawbacks. First, the model should take advantage of all the information available, meaning that while the procedure can weight or adjust the scores, it should not drop scores entirely from consideration. Second, the model should take advantage of the redundancy

in moot court scoring, in which participants are judged multiple times for multiple performances. Ideally, the model should use these additional observations to calibrate its adjustment of scores. In the next Section, we propose a statistical model that accomplishes both of these goals.

4. A HIERARCHICAL MODEL FOR SCORING

In rough, abstract terms, participant scores in a moot court competition are the result of two factors: One is the participant's speaking ability and the other is the judge's generosity. Accordingly, we can sketch the following model for moot court scoring.

$$Score_{ij} = \alpha + Ability_i + Generosity_j + \varepsilon_{ij}$$

where $Score_{ij}$ represents a score given by Judge j to Participant i, $Ability_i$ is a measure of Participant i's ability, and $Generosity_j$ is a measure of Judge j's generosity. α is a constant which represents the "grand mean" of all scores given in the competition. It merely centers the model so that all participant ability and judge generosity measures are centered around zero. ε_{ij} is an error term, which represents the random "error" between the model predictions and the observed scores.

We can further refine the model to account for additional characteristics that may affect participant performance. For instance, we can add an explanatory variable $Petitioner_{ij}$ that is a "1" when the participant argues for petitioner (and "0" for respondent). Similarly, an explanatory variable $OnBrief_{ij}$ can account for whether the oralist is arguing in the same posture as the brief his/her team wrote for the competition. These two explanatory variables can affect the score in different ways, so to weigh them appropriately, we multiply them by coefficients, β_P and β_B respectively. The augmented model thus becomes:

$$Score_{ij} = \alpha + Ability_i + Generosity_j + \beta_P * Petitioner_{ij} + \beta_B * OnBrief_{ij} + \varepsilon_{ij}$$

Before moving forward, let us be clear about which values are observed and which need to be estimated from the data using the model. From the observations, we have the actual scores ($Score_{ij}$), and we know whether the participant argued as petitioner ($Petitioner_{ij}$) and/or on brief ($OnBrief_{ij}$) during that particular argument. We also implicitly know the participant (i) and the judge (j). All other values are "parameters" that we have to estimate using the model.

Following a Bayesian approach, we can place prior distributions on the unknown parameters. These priors are initial probabilistic guesses as to how the parameter values are distributed prior to seeing the data. Given that both participant ability and judge generosity are population characteristics (like height, weight, etc.), we can reasonably model them as coming from normal distributions with mean zero and unknown variances σ_A^2 and σ_G^2 respectively. In other words, we expect the ability and generosity estimates to be roughly normal – e.g., a small number of participants will be remarkably excellent or poor, a small number of judges will be very harsh or lax, but the vast majority in both categories will cluster in the center. The means of these priors are zero because the grand mean term (α) effectively normalizes them to zero. The variances are additional unknown parameters that we will handle in a moment. We can also model the error terms (ε_{ij}) as being normally distributed, here with mean zero and variance σ_S^2.

We thus have the following prior distributions on the parameters:

$$Ability_i \sim N(0, \sigma_A^2)$$
$$Generosity_j \sim N(0, \sigma_G^2)$$
$$\varepsilon_{ij} \sim N(0, \sigma_S^2)$$

We have no prior beliefs about any of the other parameters in the model. Accordingly, we impose "flat" priors on the remaining parameters. These flat priors model all values as equally likely as an initial matter.[6]

[6] For technical reasons, we gave α, β_P, and β_B a flat normal prior (N(0, 1E6)), and the variance parameters, σ_A^2, σ_G^2, and σ_S^2, a flat gamma prior (Γ(1E-3, 1E-3)).

5. APPLICATION

5.1. 2012 Competition Data

The 2012 Vanderbilt Intramural Competition involved 160 participants paired into 80 teams during the preliminary rounds. For subjective brief scoring, the 80 briefs were each scored three times using a pool of 20 judges for a total of 240 subjective brief scores. For objective brief scoring, the briefs were each scored twice using a pool of 10 judges for a total of 160 objective brief scores. Oral arguments featured all 160 participants arguing twice (once on-brief and once off-brief), each before a three-judge panel drawn from a pool of 73 judges.

While we used the proposed model (suitably adapted) to normalize the subjective brief, objective brief, and oral argument scores, for brevity we will focus solely on oral argument scores here. Figure 2 plots histograms for the oral argument scores, as well as the mean scores awarded to individual participants and mean scores awarded by individual judges.

5.2 Results

Using the data described in Subsection 5.1, we estimated the parameters in the (hierarchical) model described in Section 4. Estimation was done using *JAGS: Just Another Gibbs Sampler*, in combination with *R*. *JAGS* is an open-source program written by Martyn Plummer for analyzing Bayesian hierarchical models.[7] *R* is a standard, open-source statistical software package.

Because our methods are Bayesian, the software produces estimated posterior probability distributions for each of the parameters. These distributions estimate each parameter's range of possible values, and how probable those values are. "Posterior" (as opposed to prior) denotes the fact that the estimated distributions have been informed by the data observed. So for example, a participant's ability might have the posterior probability distribution seen in Figure 3.

[7] See mcmc-jags.sourceforge.net/.

FIGURE 2: HISTOGRAMS OF SCORING DATA

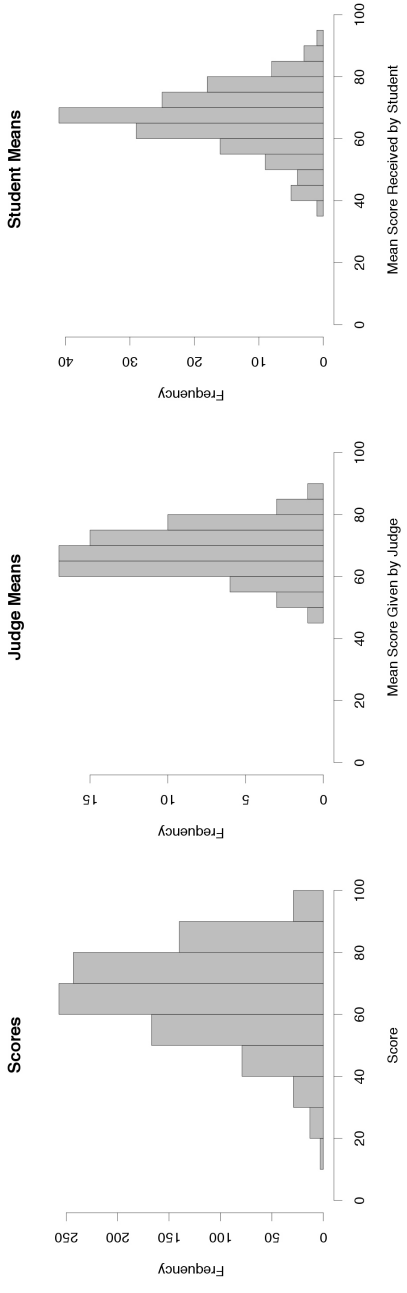
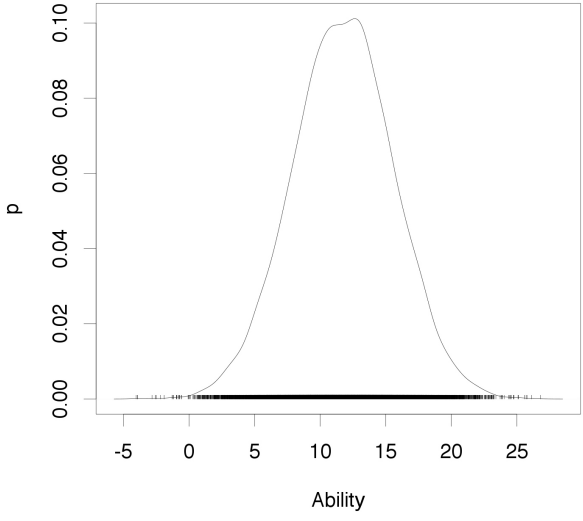


FIGURE 3: EXAMPLE POSTERIOR DISTRIBUTION FOR A PARTICIPANT'S ABILITY SCORE



For scoring/ranking purposes, we of course do not want distributions; we want single scores. Thus, we use the median of the posterior distribution as the score. We also calculated 68% (i.e., one standard error) credibility intervals to express the amount of uncertainty in the normalized score.[8]

The normalized scores produced from our proposed procedure are listed in Table 1 and graphically displayed with 68% credibility intervals in Figure 4. For legibility and to protect anonymity, we have replaced participant names with numbers, and have also displayed only a randomly selected group of 36 participants, as opposed to the full 160. The results, however, come from estimating the model on the full dataset.

Some readers may wonder whether the parameters for arguing as petitioner (β_P) and being on brief (β_B) showed any effect. Indeed, being the petitioner seemed to confer a disadvantage of approximately -0.5 points, a relatively minor effect when viewed in the context

[8] Credibility intervals are often described as the Bayesian equivalent of classical confidence intervals. Though they are not the same thing, they do similarly express information about uncertainty.

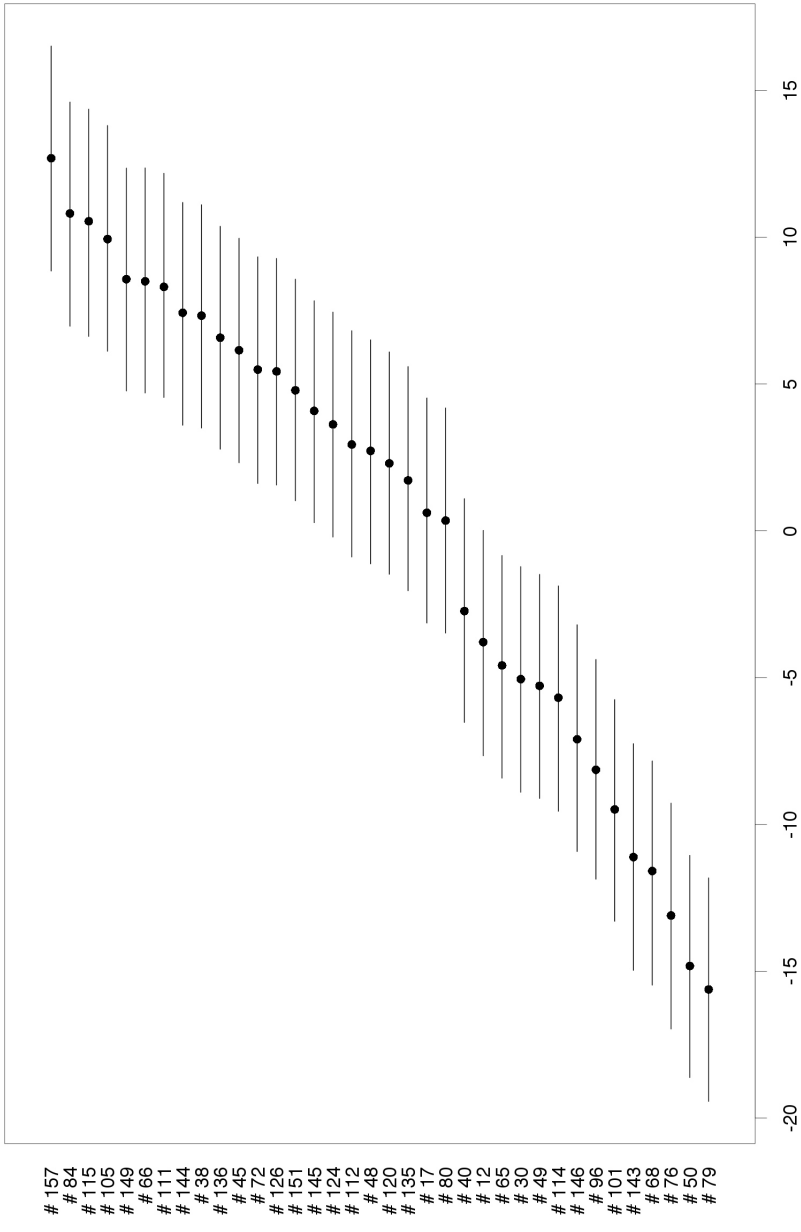
TABLE 1: (EXCERPTED) NORMALIZED SCORES FROM COMPETITION

Rank	Student	Normalized Score
1	# 157	12.71
2	# 84	10.88
3	# 115	10.44
4	# 105	10.00
5	# 149	8.58
6	# 66	8.54
7	# 111	8.38
8	# 144	7.45
9	# 38	7.27
10	# 136	6.53
11	# 45	6.18
12	# 72	5.49
13	# 126	5.42
14	# 151	4.82
15	# 145	4.10
16	# 124	3.55
17	# 112	2.95
18	# 48	2.71
19	# 120	2.34
20	# 135	1.71
21	# 17	0.70
22	# 3	0.39
23	# 40	-2.67
24	# 12	-3.84
25	# 65	-4.65
26	# 30	-5.00
27	# 49	-5.26
28	# 114	-5.68
29	# 102	-7.15
30	# 96	-8.18
31	# 101	-9.50
32	# 143	-11.07
33	# 68	-11.64
34	# 76	-13.08
35	# 50	-14.74
36	# 79	-15.66

of the participant ability scores seen in Table 1. As might be expected, arguing on brief gave a considerably larger advantage of 2.2 points. For the preliminary rounds, these advantages/disadvantages are arguably of no consequence, since all teams argue both sides. The data does suggest, however, that in the later elimination rounds, the serendipity of arguing one side versus another can make a difference, particularly when teams are evenly matched.

The model also provides generosity estimates for all of the judges, and the sheer magnitude of the judge generosity estimates is quite eye-opening. The mean of the generosity estimates is of course approximately 0 as specified by the model, but the standard error was a sizable 5.6, and generosity values ranged from -14.9 to 10.8. Given that the participant ability scores have similar magnitudes, one can easily imagine that the failure to control for judge proclivities can potentially affect outcomes, particularly at the extremes of the distribution.

FIGURE 4: PLOT OF NORMALIZED SCORES
(WITH 68% CREDIBILITY INTERVALS)



6. DISCUSSION

6.1. Limitations

While the proposed model provides a method for normalizing scores across judges and procedural postures, it does carry some cautions or limitations. These limitations were debated at length prior to implementing the model in the competition, and the organizers concluded that the benefits of normalization were well worth the costs. We discuss some of them here to help would-be adopters decide whether our approach makes sense for them.

The primary limitation is modest instability in the estimated parameters, the most important of which are the normalized participant ability scores. The standard way of estimating the model used here, a Bayesian hierarchical model, involves a simulation method called Markov Chain Monte Carlo (MCMC). Since estimation occurs via simulation, estimates necessarily fluctuate from one run to the next, meaning that participants' ranks can change, particularly if they are closely clustered.

In our competition, the variation seen between simulation runs was relatively minor. Typically, ranks changed at most a few places (if at all), with the most movement seen in the middle of the distribution, where participants are more closely clustered. At the tails, the normalized scores for competitors were sufficiently separated that random variation between runs had little effect on the rank order. Since the scores of consequence – those that result in advancement to later competition rounds or to invitations to the Moot Court Board – reside in the upper tail, we deemed the variation acceptable.

As a practical matter, decisions of course must be made on the basis of some specific score. We thus pre-selected a single random seed[9] for the simulation, and decisions were based on the results that followed.

[9] Random draws on a computer are not truly "random," but are pseudo-random in the sense that the numbers generated are completely deterministic (although appearing random) given a starting value. One can therefore reproduce "random" simulations by using the same starting value. This starting value is called the random seed.

A related drawback is the acknowledgment of uncertainty or error in the scoring model. Traditional competition scoring carries a certain flawless determinism. Judges score the participants, and the participants accept those raw scores as the "truth," the unassailable competition result. In contrast, our normalization procedure introduces uncertainty into the outcomes. Between simulation runs, random variations can change the results. Even within a simulation run, the presence of posterior distributions with credibility intervals suggests that things could have turned out differently for the participants.

In actuality, to believe that raw scores lack uncertainty is merely to fool ourselves. The judges' raw scores may appear deterministic, but they are fundamentally not – precisely for the reasons that motivated this project from the start. Moot court competitions inevitably involve some degree of serendipity and random variation. What set of judges a participant faces, what side of the dispute he or she argues, and what side's brief he or she wrote – all are a matter of chance. Thus, while the actual observed scores may be certain, as a measure of appellate advocacy ability they are at least as uncertain as the normalized scores, and arguably far more. The critical question thus becomes whether the normalized scores reduce more uncertainty than they create. Given that the judge generosity estimates are relatively large, it appears that the normalized scores are a net benefit.

Finally, hierarchical models have the considerable drawback of being difficult to understand and implement. Gone are the intuitive adjustments such as dropping the highest and lowest scores. In their place is a full statistical model that requires not only sophisticated software packages and heavy computation, but also an understanding of the model and its assumptions. From a participant standpoint, we doubt that the model's complexity makes much of a difference. Scoring rubrics and weighting schemes in moot court and other competitions are often highly opaque, and so whether the normalization occurs using simple mechanisms or a full statistical model will be largely lost on participants.

From an organizer standpoint, however, our model's complexity poses a significant obstacle to broader adoption, since it currently requires a statistician to implement. To that end, in future work, we

hope to produce easy-to-use software for inputting competition scores, running the requisite computations, and displaying results.

6.2. Extensions

The proposal presented only scratches the surface of what hierarchical models can do in normalizing scores. For example, within the moot court context, presuming sufficient data, one can theoretically control not only for judges and procedural posture, but also for opponent. To the extent that judges score arguments largely on a comparative basis, then perhaps the model should account for opponents (if not teammates) as well. Those familiar with college football computer rankings will undoubtedly recognize the "strength-of-schedule" theme.

The idea of score normalization is also not confined to the moot court context. As mentioned in the Introduction, journal competitions face similar issues, as do many large-scale competitions in which participants cannot practically be scored by a single judging panel. Indeed, scoring consistency is a frequent worry among organizers of science fairs and other youth competitions,[10] as well as faculty members running large lecture courses with multiple graders or teaching assistants.[11]

Furthermore, arguably all academic grading – or more specifically, grade point averages – would benefit from a normalization model. Instructors exhibit different levels of generosity in grading, and

---

[10] E.g., Discussion at ask.metafilter.com/151205/science-fair-judging (seeking advice for normalizing science fair scores); Fred Rose, What the Heck do My Scores Mean?, June 9, 2006, at fll-sw.sourceforge.net/ScoreExplaination.pdf (discussing scoring in the Minnesota First Lego League competition).

[11] See Norman Jacobson (2001), A Method for Normalizing Students' Scores When Employing Multiple Graders, SIGCSE Bulletin 33(4):35-38; Julianne Dalcanton, Normalizing Grades Across TA Sections, Discover Magazine, Cosmic Variance Blog, Dec. 14, 2007, at blogs.discovermagazine.com/cosmicvariance/2007/12/14/normalizing-grades-across-ta-sections; Ben Krop, Justin Meyer, and Nipa Patel, Grading and Evaluation Procedures, at www.ideals.illinois.edu/bitstream/handle/2142/1894/Rough%20Draft.pdf (reporting survey results at the University of Illinois showing such concerns).

courses have varying degrees of difficulty. Even in contexts with mandatory grading curves, the pool of students in a given class may vary, making an A in one class more precious than in another. Instructors and students alike intuitively understand these nuances, but GPAs are entirely blind to them. A hierarchical model to normalize GPAs would be more complex because of the additional factors, but in theory, nothing prevents it from working – at least assuming sufficient data.

7. CONCLUSION

This Article has proposed and demonstrated the use of a Bayesian hierarchical model to normalize moot court competition scores. Specifically, the model accounts for differing levels of judge generosity as well as the postures from which a participant argues. Future work will include developing software to help other moot court organizations implement this model in their competitions.

GB