Analyses of the Adjudication of Writing Essays For the Pennsylvania System of Student Assessment

Richard Hill
The National Center for the Improvement of Educational Assessment, Inc.

May 7, 2001
Revised May 16, 2001

Background

After students write essays as part of a writing test, those essays are scored by raters. Since human judgment is fallible, adjudication, the systematic rereading of an essay when two raters do not agree on a score, is an important element of the scoring process.

There are several possible rules that could be implemented into an adjudication process. In Pennsylvania, for example, each essay is scored on five domains. One adjudication rule might be to reread any paper on which two scorers did not agree on the score for any domain. Another rule might be to reread only those papers when the sum of scores by the two raters is different by, say, three points or more. Tighter rules would result in more accurate scoring (at least in theory), but would also require more scoring, which would add to the cost and time needed to complete the scoring. The purpose of this study was to look at a variety of possible adjudication rules and to learn what each rule provided in terms of (1) improved scoring accuracy and (2) increased costs for scoring. The ideal outcome would be recommendations that provide sufficient accuracy for a minimal increase in the amount of additional scoring that needs to be done.

As part of its on-going research efforts, the Pennsylvania Department of Education recently conducted a "generalizability study." That study will provide information about the magnitude of the various sources of error involved in the assessment of writing, and permit calculation of the reliability of the writing test under several possible designs. To provide the data for this study, over 1,500 students throughout the state each wrote four essays. Six scorers (two teams of three scorers) scored each of the essays. While this paper will use the data produced for that study, computing generalizability coefficients is not the focus of this paper; the results of the generalizability study will be provided in another paper.

These data were ideal for an adjudication study because every paper already had been "adjudicated" when scored by the third scorer. Thus, any time an adjudication rule might call for a paper to be scored by a third person, we already had the desired data. Since two independent teams scored every paper, we could see how consistent the decision about the student would have been if that adjudication rule had been applied.

Data Analyses

As a starting point, we looked at two extreme adjudication rules: (1) adjudicate nothing, and (2) adjudicate a paper any time there was any disagreement between the two scorers on any of the five domains. Table 1 shows the results for those two rules.

TABLE 1

Comparison of Results for Two Most Extreme Adjudication Rules

Statistic	Adjudicate Nothing	Adjudicate Any Disagreement on Any Domain
Percentage of additional scorings	0	79
Correlation of total scores between two teams	.85	.85
Average difference between total scores of two teams	2.6	2.5
Percentage of agreements within 2	56	68
Percentage passed by one team but failed by the other	14.0	13.3

Perhaps the most startling result of this first analysis was how little accuracy improved through adjudication, despite the fact that 79 percent of the papers were adjudicated. The only statistic for which there was a significant gain was the percentage of agreements within 2. However, after some reflection, we realized that the adjudication rules virtually assured that every domain score, when summed across two scorers, would be an even value. To make the comparison fairer, we made a rule that every paper in the "adjudicate nothing" group would be raised to the next even value if the initial result were an odd number. When we did this, 64 percent of the papers had an agreement within 2—only 4 percent less than the "adjudicate any disagreement" rule. Under the conditions of this study, at least, one could not recommend adjudication as a general process; the additional cost was considerable, and the improvement in accuracy was minor.

As a result, we changed our focus from adjudication in general to an adjudication process that would be applied only if a paper was failing, but very close to passing. For purposes of this study, we defined passing as an average of 2.5 or more on the scale of 1-4 used to score each domain. When standard setting is completed, another passing score might be established; in that case, these analyses would need to be redone using that cut score.

The initial analyses were conducted on individual papers (rather than on the collection of papers a student had written). For these analyses, three possible adjudication rules were compared:

1. Adjudicate nothing
2. Adjudicate any paper that is passed by one scorer but failed by a second; in that case, use the score of the third scorer
3. Adjudicate any paper that is within 4 points of passing (a total of 25 points out of 40 was needed to pass, given the five domain scores assigned by two scorers); in that case, take the sum of the two highest scores.

The results of these analyses are shown in Table 2.

TABLE 2

Comparison of Pass/Fail Results for Four Possible Adjudication Rules

Statistic	Adjudicate Nothing—Pass if Both Scorers Rate as Passing	Adjudicate Nothing—Pass if Either Scorer Rates as Passing	Adjudicate When One Scorer Passes and Other Fails—Pass if Third Scorer Passes	Adjudicate When Total Is 21-24; Pass if Total of Two Best ≥ 25
Percentage of Papers Rescored	0	0	18	19
Percentage Failing Both Teams	33	18	26	18
Percentage Passing One Team but Failing Other	18	14	13	12

Again, adjudication seemed to have a small effect, but because the adjudication is focused, the additional amount of adjudication is modest. In any case, the large percentage of papers that passed one team but failed the other is of concern. Regardless of the adjudication process chosen, there was at least one paper for which the teams disagreed on a pass/fail decision for every two papers that the teams agreed were failing.

However, decisions about students will not be made on the basis of one paper. Each student will be writing three, and the pass/fail decision will be based on a total score across the three papers. Fortunately, each student wrote four essays for the generalizability study, so it was easy enough to select three papers for each student and look at the scoring accuracy across those three papers.

There are several adjudication rules that could be employed across the three papers:

1. Adjudicate nothing.
2. Adjudicate any paper that is passed by one scorer but failed by a second; in that case, take the sum of the two highest scores.
3. Adjudicate any paper that is within 4 points of passing (a total of 25 points out of 40 was needed to pass, given the five domain scores assigned by two scorers); in that case, take the sum of the two highest scores.
4. Adjudicate any individual paper that is rated as close to passing by either scorer (a total of 11 or 12 points, with 12.5 needed to pass); take the sum of the two highest scores.
5. Adjudicate any total of 71 to 74 points (a total of 75 is needed to pass). In that case, rescore all three papers and take the sum of the two highest scores on each paper.

Table 3 shows the results of applying these five rules.

TABLE 3

Comparison of Pass/Fail Results across Three Papers for Five Possible Adjudication Rules

Statistic	Adjudicate Nothing	Adjudicate When One Scorer Passes and Other Fails	Adjudicate When Total Is 21-24	Adjudicate When Either Score is 11 or 12	Adjudicate When Total is 71-74
Percentage of Papers Rescored	0	18	15	28	7
Percentage Failing Both Teams	21	19	20	19	18
Percentage Passing One Team but Failing Other	9	9	8	9	8

One obvious point worth noting is that the reliability of the process is substantially higher when judgments are made about students on the basis of three papers rather than one. The percentage of times that the teams disagreed about whether a student should pass or fail was reduced from 12-18 percent (depending on the adjudication rule) for one paper to 8-9 percent. Still, it is a point of concern that 8-9 percent of the students would pass if one team scored their papers, but would fail if scored by a second team.

For the most reliable adjudication rule, the teams disagreed on 146 of 1,719 students; for the least reliable, 150. So, the differences between the rules, at least in terms of agreement between the teams, were minor. The last rule—looking at how close a student was to passing when scores are summed over all three papers—required substantially less adjudication than any of the other rules, and therefore, might be considered the most cost effective. However, this rule could not be applied practically in a paper-based scoring system; it would be too costly to locate the three papers to have them rescored. However, if scoring were done from computer images, this adjudication rule could be implemented quite straightforwardly.