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

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

April 4, 2001
Revised--August 7, 2001

Background

Pennsylvania is in the process of revising its assessment program. Previously, the primary focus of its assessment results were at the school level; now, they will be reporting student-level results. While this change of focus has affected the design of the assessment for all content areas, changes have been most dramatic in writing.

In the past, each student took one writing prompt, matrix-sampled from a collection of nine (three different prompts from each of three modes of writing—narrative, informative and persuasive). The results across the nine prompts were aggregated to provide school-level results. Each prompt was scored once, assigned a rating on a scale from 1 to 4 for each of five domains (Focus, Content, Organization, Style and Conventions), with a small sample rescored to provide information about the reliability of the scoring process.

With the change to student-level reporting, several questions arose about how to design a test that would be sufficiently reliable at the student level. How many prompts should each student take? How many times should each prompt be scored? Is it important that students take a prompt from each mode? Answers to these questions are important, because the assessment of writing is time-consuming both to take and to score. While the easy answer to all questions of reliability is "more is better," that answer was too imprecise for this issue. Each prompt, for example, takes a full class period that otherwise could be instructional time; adding just one additional prompt to a writing assessment requires almost as much time as the assessment of an entire content area might take using a multiple-choice format. Scoring an assessment takes between 5 and 15 minutes of scorer time. If one essay were added to an assessment in Pennsylvania, the time to score the 120,000 additional writing samples that would be generated, if each were scored twice, would be between 20,000 and 60,000 scorer hours. If 100 scorers were assigned to the task, this one additional essay would add between 5 and 15 *weeks* to the gap between testing and the return of results, and $300,000 to $900,000 to the cost of the scoring.

In order to collect sufficient information about these issues, Pennsylvania is conducting 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.

The Department periodically convenes a Technical Advisory Committee of outside experts to assist in the formulation and interpretation of such studies. In the process of preparing for the generalizability study, members of the TAC noticed that there was a high intercorrelation among the scores assigned to the five domains by scorers. Each scorer scores all five domains at one time, so a series of questions arose. Is there a "halo" effect; that is, is the score a rater assigns to the other domains strongly influenced by the score assigned to the first one? Do scorers take longer to score papers when they score five domains (meaning that the scoring process would be more expensive and time-consuming than another scoring system)? If so, is that added time worth the cost? Is there an alternative means of scoring that would meet the requirement that student receive information about their performance in different domains that would be less time-consuming to score, but would be equally reliable? The purpose of this paper is to answer the first question; the others will be addressed by future papers.

There were two major sources of data used in this study. The first, from the "generalizability study," are the scores of the papers from the aforementioned study, in which over 1,500 students took four prompts each. The generalizability study actually contained six substudies; in three of them, students took four prompts in each of one mode, and in the other three, took two prompts in each of two modes. Two teams, one containing 15 scorers and the other 16, scored each of the papers. A random set of three scorers from each team scored each paper for all five domains. These data provided a wealth of information about the intercorrelations among scorers and among domains.

However, the generalizability study data did not tell us what the correlations would be if a different scorer had scored each domain; that is, what the correlations would be if there were no halo effect (if there was one). Therefore, one paper from each student for about 1,200 of the students tested in the generalizability study was selected and rescored by a subset of the scorers used in the generalizability study. For this study, called "the domain scoring study," five teams of two scorers each was formed. Each team scored just one domain; both scorers on the team scored all the papers in the study for that domain.

Results

Table 1 provides the correlations of every scorer with him/herself from the generalizability study. These are the results that raised the questions from the TAC in the first place. The intercorrelations among the first three domains are all .80 or higher, and the correlations of those three domains with the fourth are all .71 or higher. Only the correlations with the last domain (conventions) appear to be significantly lower than what we had anticipated the reliability of the scoring to be. With correlations among the domains being that high, it was likely that the scoring of each individual domain added little new information; it was necessary to look at these results more closely to see whether the amount of new information could justify the additional cost of scoring all five domains.

The results from Table 1 are based on a very large sample of scores. Each of almost 6,000 papers was scored by six scorers (from the total pool of 31), so each correlation is based on the results of 34,619 scorings.

TABLE 1

Correlations among Domains When Same Scorer Scores All Domains

Domain	1	2	3	4	5
1	1.00	.80	.82	.71	.57
2		1.00	.80	.74	.55
3			1.00	.75	.60
4				1.00	.65
5					1.00
Mean	2.79	2.67	2.79	2.81	3.08
Stan. Dev.	.74	.75	.73	.74	.78

In contrast, Table 2 provides the correlations among the domains when a second scorer scores the same paper. Since there were 6 scorers of each paper, there was a total of 30 combinations of two scorers for each of the 6,000+ papers, so each cell in the table represents the data from over 200,000 pairs. As would be expected, and is evident from the results, all the correlations are much lower when a second scorer rates the paper. However, the same patterns remains. The intercorrelations among the first three domains are quite high—almost as high as the correlations of the domains with themselves—suggesting that either (1) students' skills in the various domains are highly correlated, or (2) there is some halo effect in the scoring (or both). The correlations with the fourth domain are somewhat lower, and the correlations with conventions are substantially lower than any of the others. The within-domain correlations are highest for Content (Domain 2), with the remaining four being quite consistent with each other.

TABLE 2

Correlations among Domains with Different Scorer:
Each Scorer Scores All Domains

Domain	1	2	3	4	5
1	.67	.66	.64	.60	.49
2		.74	.66	.63	.48
3			.67	.61	.51
4				.66	.54
5					.64
Mean	2.79	2.67	2.78	2.81	3.07
Stan. Dev.	.74	.75	.73	.73	.78

All the papers scored in the domain scoring study had previously been scored in the generalizability study, and the scorers in the domain scoring study were a subset of those used in the generalizability study. As a result, scorers occasionally rescored a paper (on one domain) in the domain scoring study that he/she had previously scored (on all five domains) in the generalizability study. This allowed us to look at the correlations of a scorer with him/herself. Table 3 provides those results.

TABLE 3

Correlation among Domains When Same Scorer Scores Same Paper Twice

Domain	1	2	3	4	5
1	.62	.66	.58	.50	.49
2		.76	.64	.62	.51
3			.62	.55	.53
4				.50	.43
5					.64
Mean	3.03	2.92	2.91	3.16	3.32
Stan. Dev.	.67	.73	.67	.66	.70

In contrast to the previous tables, these results are based on considerably fewer scores and scorers. All the correlations on the diagonal (as well as the means and standard deviations) are the data from just two scorers, on a total of 260-460 papers. The off-diagonal correlations are based on four scorers, scoring 600-900 papers.

The correlations in Table 2, where a different scorer provided the second score, generally are higher than those in Table 3, where the same scorer provided the second score. One item to note in the interpretation of this result is that the papers scored in Table 3 had higher means and lower standard deviations. Smaller standard deviations could lead to lower correlations, so this issue warranted further study. One possibility was that the papers in the domain scoring study were better-written; another was that the scorers simply gave them higher scores in this study.

Table 4 provides some of the answer to that question, by showing the means and standard deviations assigned by the first scorer to papers *during the generalizability study*, separately for those used in the generalizability study only vs. those used in both studies. The standard deviation of the papers used in both studies was lower than those only in the generalizability study, probably due to the ceiling effect caused by the papers being better written. The papers of one prompt for each mode were chosen for the domain scoring study. When the choice of which prompt to use in the second study was made, the scoring contractor generally chose the prompt that was "best," meaning that students did the best job on that prompt. So some, but not all (as will be seen shortly), of the higher means of the papers in the domain scoring study came from the selection of the prompts.

4

TABLE 4

**Means and Standard Deviations Obtained During the Generalizability Study,
Of Papers Used in One Study vs. Those Used in Both**

Domain	Papers used in generalizability study only			Papers used in both studies		
	Mean	St. Dev.	N	Mean	St. Dev.	N
1	2.73	.75	5817	2.97	.71	1108
2	2.58	.74		2.88	.73	
3	2.70	.74		2.95	.71	
4	2.78	.73		3.02	.68	
5	2.97	.78		3.15	.74	

Another area of interest was the intercorrelation among the domains when a different scorer scores each domain (which was the initial focus of the domain scoring study). Those correlations are provided in Table 5. The diagonals represent the data for two scorers on 1,100 papers; the off-diagonals represent four pairs of scorers on those same 1,100 papers. A particularly unique characteristic of these data is that each scorer scored just one domain.

TABLE 5

**Correlation among Domains with a Different Scorer on the Same Paper
(Each Scorer Scoring Just One Domain)**

Domain	1	2	3	4	5
1	.51	.61	.58	.55	.45
2		.74	.67	.63	.46
3			.66	.59	.48
4				.50	.50
5					.61
Mean	3.03	2.93	2.98	3.18	3.31
Stan. Dev.	.65	.73	.63	.66	.72

Table 3 provides the correlations of scorers with themselves (scoring all five domains on one occasion), while Table 5 provides correlations on those same papers with other scorers (with both scorers scoring only one domain). Of the 15 correlations in each table, seven are higher in Table 3 and eight are lower. This is true despite the fact that the standard deviations for the papers in both tables are quite similar. Thus, when a second scoring is to be done, the correlations are pretty much the same whether the same scorer or another scorer provides the second score. This implies that these scorers were quite interchangeable—a randomly selected scorer is as likely to provide the same second score as the original scorer.

Further Results

At a meeting with staff from the Department of Education and the scoring contractor, a series of questions and suggestions arose concerning the above analyses. One suggestion was to rerun Tables 1 and 2 using only the papers that subsequently were scored in the domain scoring study; that would

eliminate one source of uncertainty in the interpretation of the subsequent tables. Another was to investigate whether, when the papers were held constant, scores were higher when scorers scored all five domains at one time or when the scored one domain at a time. Finally, a suggestion was made to consider the diagonals in each table to be reliability coefficients, and to compute was the correlation was among domains when corrected for unreliability. These suggestions were implemented and are included in the tables in this section.

Tables 6 and 7 are repeats of Tables 1 and 2, except that the papers used to compute the correlations in these tables are only those that subsequently were included in the domain scoring study. This reduced the number of observations per cell in Table 1 to a little over 6,600 (six scorers for each of over 1,100 papers), and the number of observations per cell in Table 7 to 33,225 (30 combinations of pairs of scorers for each of the papers).

The means of the papers selected for the domain scoring study were indeed somewhat higher than those for all the papers. This was already known from Table 4, and is further reflected in Tables 6 and 7. Note that the means and standard deviations in Tables 6 and 7, which are based on all six scorers, are slightly different from those in Table 4, which are based on just the first scorer. Contrary to our expectation, however, the reduced standard deviation did not produce reduced correlations. The correlations in Table 6 are all at least as high as those in Table 1; the correlations in Table 7 tend to be somewhat smaller than those in Table 2, but there was not a dramatic dropoff in the size of the correlations for this subset of papers (perhaps because they indeed were better prompts, which in turn made it easier to score them more consistently despite the reduction in variance due to the ceiling effect).

TABLE 6

**Correlations among Domains When Same Scorer Scores All Domains,
Using Only Papers Included in Both Studies**

Domain	1	2	3	4	5
1	1.00	.80	.83	.72	.60
2		1.00	.80	.76	.57
3			1.00	.76	.63
4				1.00	.66
5					1.00
Mean	2.99	2.90	2.98	3.02	3.19
Stan. Dev.	.72	.75	.72	.72	.75

TABLE 7

Correlations among Domains with Different Scorer:
Each Scorer Scores All Domains,
Using Only Papers Included in Both Studies

(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.66	.64	.63	.59	.49
2	.94	.70	.63	.60	.48
3	.95	.93	.66	.59	.51
4	.91	.90	.91	.64	.52
5	.77	.73	.80	.83	.61
Mean	2.99	2.90	2.98	3.02	3.19
Stan. Dev.	.72	.75	.72	.72	.75

Table 8 is simply Table 3 redone with corrected correlations below the diagonal. As was true with Table 3, the scorers scored all five domains on one occasion, but just one domain on the other occasion.

TABLE 8

Correlation among Domains When Same Scorer Scores Same Paper Twice
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.62	.66	.58	.50	.49
2	.96	.76	.64	.62	.51
3	.94	.93	.62	.55	.53
4	.90	1.00	.99	.50	.43
5	.78	.78	.84	.76	.64
Mean	3.03	2.92	2.91	3.16	3.32
Stan. Dev.	.67	.73	.67	.66	.70

.

Table 9 is simply Table 5 redone with corrected correlations below the diagonal. As was true with Table 5, each scorer scored just one domain. Note that the maximum reported value for the corrected correlations is 1.00, although some of the correlations between domains are higher than the mean of the diagonals.

It would seem that the clearest answer to the question of whether there is a halo effect in scoring would be to compare Table 7 to Table 9. Table 7 provides the intercorrelations among scorers when they score all five domains at one time; Table 9 provides the same information when they score just one domain at a time. Seven of the 10 correlations in the upper diagonal are larger in Table 7 than they are in Table 9, but 8 of the 10 correlations in the lower diagonal are larger in Table 9 (because of the lower correlations in the diagonals in Table 9). The results seem to indicate that there is a high correlation among the first four domains, but not due to halo effect—the intercorrelation of the skills, at least as defined for purposes of the scoring that was done, is truly that high.

TABLE 9

Correlation among Domains with a Different Scorer on the Same Paper
(Each Scorer Scoring Just One Domain)
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.51	.61	.58	.55	.45
2	.99	.74	.67	.63	.46
3	1.00	.96	.66	.59	.48
4	1.00	1.00	1.00	.50	.50
5	.81	.68	.76	.91	.61
Mean	3.03	2.93	2.98	3.18	3.31
Stan. Dev.	.65	.73	.63	.66	.72

One final analysis of the data looked into whether scorers gave higher scores to papers when they scored one domain at a time rather than all five at once. Table 10 provides the means and standard deviations for papers that were scored by the same scorer in both the generalizability study and the domain scoring study. Scorers gave slightly higher scores to the papers when each domain was scored separately rather than all being scored at one time, but the differences were small. These results include the data for just two scorers for each domain, so while the differences in Domains 4 and 5 are somewhat larger than the other three, that finding should not be overinterpreted, due to the small number of scorers in each row.

TABLE 10

Means and Standard Deviations When Papers in Table 8 Are Scored
Under Different Conditions

Domain	Scores When Scored in Generalizability Study		Scores When Scored in Domain Scoring Study		N
	Mean	St. Dev.	Mean	St. Dev.	
1	3.00	.68	3.05	.66	462
2	2.92	.73	2.93	.74	449
3	2.89	.72	2.94	.63	418
4	3.04	.65	3.29	.64	364
5	3.25	.72	3.38	.67	264

Some Additional Findings

Students wrote four papers each for the generalizability study. As a result, it was possible to look at the intercorrelation among scores from one paper to another within student. These correlations are of interest—they describe the relationships among the domains across different samples of the student's writing. If information about the domains is generalizable, the same patterns of strengths and weaknesses should hold up across papers written by the same student.

Table 11 provides the correlations when the same scorer happened to have scored both papers; Table 12 provides the same information when the two scorers were different. There were over 1,700 students in the study, with each student writing four papers, and six scorers scoring each paper. There were six unduplicated combinations of the four papers (1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4, and 3 with 4), and 36 combinations of scorers for each combination of papers (6 on the first paper times 6 one the second). Thus, the entire data file consisted of 1,700+ * 6 * 36, or over 367,000 pairs of data for each domain. Of these, approximately 1 out of 30 (the actual number was 13,230) pairs were of the same scorer; the remainder matched two different scorers.

TABLE 11

Correlation among Domains when the Same Scorer Scores
Two Different Papers by the Same Student
(Scoring All Five Domains at One Time)
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.49	.51	.49	.48	.43
2	.99	.54	.50	.50	.42
3	1.00	.97	.49	.48	.44
4	.96	.95	.96	.51	.45
5	.84	.79	.86	.87	.53

TABLE 12

Correlation among Domains when Two Different Scorers Score
Two Different Papers by the Same Student
(Scoring All Five Domains at One Time)
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.47	.49	.47	.47	.41
2	.99	.52	.48	.48	.41
3	1.00	.97	.47	.47	.42
4	.98	.95	.98	.49	.45
5	.85	.81	.88	.92	.49

Perhaps the most remarkable thing about the two tables is how closely they match each other. The diagonals are slightly, but consistently, higher in Table 11, indicating some uniqueness to each scorer. But the corrected correlations are almost identical in the two tables, which would seem to mean that those correlations describe the true relationship among the domains. And that relationship is virtually identical to that described earlier in this paper; the first four domains are highly intercorrelated, and only conventions stands out (only in relative terms) as a unique factor in the scoring system.

Interestingly, staff in the Department of Education had computed correlations similar to those in Table 1 and Table 12 for grades 6 and 9. Tables 13 and 14 provide the correlations among domains for one scorer on one paper (which is comparable to the data presented in Table 1) for grades 6 and 9,

respectively. Tables 15 and 16 provide the correlations among domains across prompts (which is comparable to the data presented in Table 12), again for grades 6 and 9 respectively.

The intercorrelations at grade 6 (Table 13) are similar to, but generally somewhat higher than, the same statistics at grade 11 (Table 1). Interestingly, at grade 9 (Table 14), the intercorrelations are the highest, and are quite high, even for Conventions. It is not clear why writing would become more unidimensional from grade 6 to grade 9, but then have at least one factor distinguish itself at grade 11.

This is especially clear when looking at the correlations across two different prompts. While there is some small uniqueness for Conventions at grade 6 (but not as much as for grade 11—compare Table 15 to Table 12), there is virtually none at grade 9 (Table 15). These results would seem to have strong implications about what the value of domain scoring is, and how that differs for various grades.

TABLE 13

Correlations among Domains When Same Scorer Scores All Domains—Grade 6

Domain	1	2	3	4	5
1	1.00	.76	.78	.73	.64
2		1.00	.81	.79	.66
3			1.00	.80	.70
4				1.00	.76
5					1.00

TABLE 14

Correlations among Domains When Same Scorer Scores All Domains—Grade 9

Domain	1	2	3	4	5
1	1.00	.85	.84	.81	.80
2		1.00	.85	.82	.79
3			1.00	.84	.81
4				1.00	.87
5					1.00

TABLE 15

**Correlation among Domains when Two Different Scorers Score
Two Different Papers by the Same Student—Grade 6
(Scoring All Five Domains at One Time)**
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.38	.39	.41	.42	.42
2	.96	.43	.43	.44	.42
3	1.00	.99	.44	.46	.45
4	.98	.97	1.00	.48	.49
5	.93	.87	92	.96	.54

TABLE 16

**Correlation among Domains when Two Different Scorers Score
Two Different Papers by the Same Student—Grade 9
(Scoring All Five Domains at One Time)**
(Raw Correlations Above Diagonals, Corrected Correlations Below)

Domain	1	2	3	4	5
1	.46	.47	.47	.47	.47
2	1.00	.48	.48	.48	.47
3	.99	.99	.49	.49	.49
4	.98	.98	.99	.50	.50
5	.97	.97	.98	.99	.51