Generalizability Studies – PA Writing Assessment
June 2001
Prepared by Suzanne Lane

Overview

Two Generalizability Studies were conducted for the 2000 11th grade PA Writing Assessment Study: 1) Student x Prompt x Rater and 2) Student x (Prompt:Mode) x Rater. The results of the first generalizability study indicate the number of prompts needed to be administered to each student and the number of raters needed to score each student response within a mode to have a certain degree of confidence in generalizing to other prompts and raters within the same mode. For each of the three modes (narrative, informational, and persuasive), a Student x Prompt x Rater generalizability study was conducted. This allowed for the estimation of seven variance components: student, prompt, rater, student x prompt, student x rater, prompt x rater, and student x prompt x rater. The results of this study indicate the extent to which raters and prompts contribute to the error in student writing assessment scores within a mode and provide information on the number of prompts and raters needed to obtain a dependable student score within a mode.

The results of the second study provide information on whether mode has an effect on the students' writing scores. For each of three combinations of modes (narrative and informational, narrative and persuasive, informational and persuasive), a Student x (Prompt:Mode) x Rater generalizability study was conducted. This design allowed for the estimation of the following variance components: student, prompt:mode, rater, mode, student x prompt:mode, student x rater, student x mode, prompt:mode x rater, mode x rater, student x mode x rater, and student x prompt:mode x rater. The results of this study provide information on whether mode has an effect on student scores.

Results for the Student x Prompt x Rater Generalizability Studies

Table 1 and 2 provide descriptive data for the Student x Prompt x Rater generalizability studies. As indicated previously, one generalizability study was conducted for each of the 3 modes, with 4 prompts at each mode, and 3 raters scoring responses to all prompts for all students.

Table 1 provides, for each of the four prompts within a mode, the number of student essays rated, and the mean total score and standard deviation across student essays. In general, within each mode, the prompt mean total scores are similar, indicating that the prompt main effect is

small. It is also interesting to note that the prompt mean total scores are also similar across modes, indicating that the difficulty of the prompts are similar across modes. For eleven of the twelve prompts, the mean total scores range from 14.30 to 14.82. However, for one of the twelve prompts, persuasive prompt 1, the mean total score is 13.91, indicating that this prompt is somewhat more difficult than the other eleven prompts.

Table 1
Descriptive Data for Student x Prompt Generalizability Studies – Mean Total Score for Prompt

	Narrative			Informational			Persuasive		
	n	Mean	s	n	Mean	s	n	Mean	s
Prompt									
1	849	14.44	3.25	873	14.45	2.90	861	13.91	2.90
2	849	14.54	3.17	870	14.43	2.86	858	14.65	2.95
3	846	14.71	3.14	867	14.38	2.92	867	14.82	3.27
4	843	14.81	3.04	864	14.30	2.80	864	14.36	3.11

Table 2 provides the number of student essays rated, and mean total score and standard deviation across student essays, for each of the 3 raters within a mode. Within each mode, the rater mean total scores are similar, ranging from 14.24 to 14.96, indicating that the rater main effect is also small.

Table 2
Descriptive Data for Student x Prompt Generalizability Studies – Mean Total Score for Rater

	Narrative			Informational			Persuasive		
	n	Mean	s	n	Mean	s	n	Mean	s
Rater									
1	1129	14.33	3.07	1158	14.38	3.13	1150	14.24	3.23
2	1129	14.58	3.07	1129	14.96	3.28	1150	14.72	2.94
3	1158	14.37	2.55	1158	14.42	2.90	1150	14.37	3.05

Table 3 provides the estimated random effects variance components for the Student x Prompt x Rater generalizability study for each mode. The VARCOMP procedure in SAS was used to estimate the variance components. For each mode, the student variance component accounts for the largest percent of the total variance, 63% for narrative, 61% for informational, and 57% for

persuasive. It is apparent that there is slightly less variability due to error for both the narrative and informational modes than for the persuasive modes.

The student x prompt variance component accounted for a substantial percent of the total variance for each mode (20%, 13%, and 22% for narrative, informational, and persuasive, respectively). The student x prompt interaction represents the differential performance of students across prompts. The student x prompt x rater variance component also accounted for a substantial percent of the total variance for each mode (14%, 23%, and 17% for narrative, informational and persuasive, respectively), indicating that the combination of raters and prompts are ordering students differently to some extent. It should also be noted that the three-way interactions between students, prompts, and raters are confounded with unmeasured sources of variation.

As indicated previously, the variance due to prompt and rater is negligible, indicating that the prompt mean total scores as well as the rater mean total scorers are very similar. The student x rater and prompt x rater variance components were negligible, indicating that the relative standing of students are similar across raters, averaging over prompts and that there is considerable consistency of raters' average ratings of students from one prompt to the next.

Table 3
Student x Prompt x Rater Generalizability Study – Random Effects Variance Components

	Narrative		Informational		Persuasive	
Source	Variance	%	Variance	%	Variance	%
s	6.28630	63	5.02767	61	5.40562	57
p	-0.01041	0	-0.03067	0	0.09497	1
r	0.07902	<1	-0.02278	0	0.02116	<1
s x p	1.96302	20	1.11867	13	2.13155	22
s x r	0.15022	2	0.19388	2	0.11248	1
p x r	0.08557	1	0.08588	1	0.16433	2
s x p x r	1.42537	14	1.87889	23	1.62347	17

Note. $n_p = 4$, $n_r = 3$
Negative variance components are most likely due to the small number of levels for a given source, and are set to zero in further analyses.

For the decision studies, a random effects design was used. The decision studies reflected 1, 2, 3, or 4 prompts, and 1 or 2 raters. The dependability coefficients, that are appropriate for absolute (criterion-referenced or standards-referenced) score interpretations, are shown in Table 4. As an example, if one rater rates student essays to one prompt in the narrative mode, the

coefficient is equal to .629, indicating that student variation accounts for approximately 63 percent of the total variability. As the number of prompts and raters increase, the coefficient increases. However, in most cases, increasing the number of prompts has a greater impact on the coefficient than increasing the number of raters. For example, if one rater rates student essays to two prompts in the narrative mode, the coefficient is .762, whereas if two raters rate student essays to one prompt in the narrative mode, the coefficient is .689. In general, the coefficients tend to be slightly higher for the narrative and informational modes than the persuasive mode. In deciding on the number of prompts and raters to include in the assessment, it is important to consider the stakes associated with the decisions to be made about individual students based on their assessment scores. Practical factors also need to be considered.

Table 4
Student x Prompt x Rater Decision Studies – Dependability Coefficients

	$n'_p = 1$ $n'_r = 1$	$n'_p = 2$ $n'_r = 1$	$n'_p = 1$ $n'_r = 2$	$n'_p = 2$ $n'_r = 2$	$n'_p = 3$ $n'_r = 1$	$n'_p = 3$ $n'_r = 2$	$n'_p = 4$ $n'_r = 1$
Narrative	.629	.762	.689	.810	.819	.860	.851
Informational	.605	.743	.696	.814	.804	.863	.839
Persuasive	.565	.716	.629	.769	.786	.830	.836

Results for the Student x (Prompt: Mode) x Rater Generalizability Studies

Table 5 provides descriptive data for the Student x (Prompt:Mode) x Rater generalizability studies. As indicated previously, one generalizability study was conducted for each of 3 combinations of modes: Narrative and Informational, Narrative and Persuasive, and Informational and Persuasive. For each study, there were 2 prompts in each mode, 2 modes, and 3 raters.

Table 5 provides the mean total score and standard deviation for each prompt across raters, for each mode across prompts and raters, and for each rater across prompts and modes. There is little variation among mean total scores for prompts, modes, and raters, indicating that the main effects for these three sources are small.

Table 5
Student x (Prompt:Mode) x Rater Generalizability Studies – Mean Total Score for Prompt, Mode, and Rater

	n	Mean	s
Prompt			
Narrative 1	1725	14.31	3.15
Narrative 2	1722	14.77	3.09
Informational 1	1737	14.01	3.13
Informational 2	1731	14.16	3.14
Persuasive 1	1770	14.04	3.18
Persuasive 2	1773	14.37	3.12
Mode			
Narrative	3447	14.54	3.13
Informational	3468	14.08	3.14
Persuasive	3543	14.21	3.15
Rater			
111	1125	14.52	3.03
112	1125	14.25	3.09
113	1125	14.18	3.01
115	1173	14.38	3.35
116	1173	14.27	2.92
117	1173	14.93	2.82
101	1188	14.07	3.12
110	1188	13.78	3.49
114	1188	14.12	3.27

Note. Raters 111, 112, and 113 rated essays for the narrative and informational combination, raters, 115, 116, and 117 rated essays for the narrative and persuasive combination, and raters 101, 110, and 114 rated essays for the informational and persuasive combination.

Table 6 provides the estimated random effects variance components for the Student x (Prompt:Mode) x Rater generalizability study for each combination of modes. The VARCOMP procedure in SAS was used to estimate the variance components. As indicated previously, one generalizability study was conducted for each of the three combinations of modes: narrative and informational, narrative and persuasive, informational and persuasive. For each study, there were 2 prompts for each mode. Eleven variance components were estimated: student, prompt:mode, rater, mode, student x mode, student x rater, mode x rater, student x prompt:mode, rater x prompt:mode, student x mode x rater, student x rater x prompt:mode.

The primary reason for presenting this table is so that the variance components that include mode can be examined. The variance due to mode is negligible, indicating that modes tend to be similar in difficulty. The student x mode variance component is also small, indicating that the relative standing of students is similar across modes, averaging over raters. The mode x rater

variance component is small, indicating that raters were consistent in their ratings across modes. The student x mode x rater variance is also negligible.

Similar to the previous studies, the student variance component accounts for the largest percent of the total variance for each mode combination. However, it is apparent that there is slightly less variability due to error for both the narrative and informational combination (36%) than for the two combinations that include the persuasive mode (42% for the narrative and persuasive combination and 43% for the informational and persuasive combination).

Table 6
Student x (Prompt:Mode) Generalizability Studies – Random Effects Variance Components

	N and I		N and P		I and P	
Source	Variance	%	Variance	%	Variance	%
s	6.01945	64	5.60568	58	6.28143	57
p:m	0.03836	<1	0.03580	<1	0.02344	<1
r	0.01934	<1	0.07588	1	0.02343	<1
m	0.03436	<1	-0.04464	0	-0.01793	0
s x m	0.09640	1	-0.17413	0	0.18822	2
s x r	0.07825	1	0.32754	3	0.08730	1
m x r	0.01249	<1	0.04809	<1	-0.02379	0
s x (p:m)	1.43585	15	1.42388	15	1.97364	18
r x (p:m)	0.02018	<1	0.08092	1	0.07915	<1
s x m x r	-0.00877	0	0.06331	1	0.04724	1
s x r x (p:m)	1.59585	17	1.93112	20	2.25075	21

Note. N is narrative, I is informational, and P is persuasive
m=2, p=2, r=3 s = 282 for N and I, 294 for N and P, 299 for I and P
Negative variance components are most likely due to the fact that there are only two levels of mode, and are treated as zero in further analyses

Also similar to the previous studies, the variance due to prompt:mode is negligible, indicating that the difficulty of the prompts, averaging across modes and raters, is similar. The variance due to rater is negligible, indicating that raters tend to be similar in terms of their leniency or stringency. The student x rater variance component is also small, indicating that the relative standing of students is similar across raters, averaging over modes. However, the student x prompt:mode variance component accounted for a substantial percent of the total variation (ranging from 15 to 18%). This interaction represents the differential performance of students across prompts. The rater x prompt:mode variance component was negligible, indicating that raters were consistent in their ratings across prompts. For each mode combination, the highest order interaction, student x prompt:mode x rater, accounted for a substantial percent of the total

variance (ranging from 17 to 21%), indicating that the combination of raters and prompts within modes are ordering students differently to some extent. This occurs to a greater extent when prompts within the persuasive mode are considered in the analyses. It should be noted that the highest order interactions are confounded with unmeasured sources of variation.

Table 7 provides the estimated variance components for the Student x (Prompt:Mode) x Rater generalizability study treating prompt and rater as random facets and mode as a fixed facet (see Brennan (1983) and Shavelson and Webb (1991) for using estimated random effects variance components for estimating variance components that include a fixed facet). Mode is treated as a fixed facet because generalization beyond the modes included in the study is not of interest. These variance components are similar to those found in the above table because the effect due to mode was small.

Table 7
Student x (Prompt:Mode) Generalizability Studies – Random Effects Variance Components for Prompt and Rater and Fixed Effects Variance Components for Mode

	N and I		N and P		I and P	
Source	Variance	%	Variance	%	Variance	%
s	6.06765	65	5.60568	59	6.37554	59
p:m*	0.03836	<1	0.03580	<1	0.02344	<1
r	0.02585	<1	0.09992	1	0.02343	<1
m*	-	-	-	-	-	-
s x m*	-	-	-	-	-	-
s x r	0.07825	1	0.35919	4	0.11092	1
m* x r	-	-	-	-	-	-
s x (p:m*)	1.43585	16	1.42388	15	1.97364	18
r x (p:m*)	0.02018	<1	0.08092	1	0.07915	<1
s x m x r	-	-	-	-	-	-
s x r x (p:m*)	1.59585	17	1.93112	20	2.25075	21

Note. * Mode is treated as a fixed effect
N is narrative, I is informational, and P is persuasive
m=2, p=2, r=3, s = 282 for N and I, 294 for N and P, 299 for I and P

For the decision studies, students, prompts, and raters were considered to be random facets, whereas mode was considered to be a fixed facet. The variance components from Table 7 were used for the decision studies. The results of the decision studies reflecting 1 2, or 3 prompts, 1 or 2 raters, and 1, 2, or 3 modes is presented in Table 8. The dependability coefficients, which are appropriate for absolute (criterion-referenced or standards-referenced) score interpretations are shown.

As indicated in the table, when the number of raters, prompts, and modes equal 1 the coefficients are .655, .587, and .588. The largest coefficient is for the narrative and informational combination and the smaller coefficients are for the two combinations that include the persuasive mode. When the number of modes equal 2, with 1 prompt in each mode, and the number of raters equal 1, the coefficient ranges from .719 to .786. As an example, if there is one prompt in each of the narrative and informational modes, and only 1 rater rates student essays, the coefficient is equal to .786. When the number of modes equal 2, with 1 prompt in each mode, and the number of raters equal 2, the coefficient ranges from .793 to .836. When the number of modes equal 3, with 1 prompt in each mode, and the number of raters equal 1, the coefficients range from .776 to .842. When the number of modes equal 3, with 1 prompt in each mode, and 2 raters rate each student's essays, the coefficients range from .850 to .881. It should be noted, however that when mode is equal to 3, the coefficients are only approximations because in each of the generalizability studies only two modes were used and mode is considered a fixed facet. Lastly, when the number of modes equal 3, with 1 prompts in each mode, and 2 raters rate each student's essay, the coefficient ranges from .850 to .881. The coefficients tend to be highest for the narrative and informational combination and lowest for the combinations that include the persuasive mode.

Table 8
Student x (Prompt:Mode) Decision Studies (Dependability Coefficients)- Prompt and Rater as Random Facets and Mode as a Fixed Facet

	N and I	N and P	I and P
n'_p =1, , n'_r =1, n'_{m*} =1	.655	.587	.588
n'_p =1, , n'_r =1, n'_{m*} =2	.786	.719	.735
n'_p =1, , n'_r =2, n'_{m*} =2	.836	.793	.794
n'_p =1, , n'_r =1, n'_{m*} =3	.842	.776	.802
n'_p =1, , n'_r =2, n'_{m*} =3	.881	.873	.850

Note. *Mode is treated as a fixed effect
N is narrative mode, I is informational, and P is persuasive

Recommendations

Based on the above analyses with the acknowledgement of each study's limitations, the following recommendations can be made:

1. For high stakes decisions at the individual student level, at least 3 prompts should be administered to each student and at least 2 raters are necessary for scoring each student's essay.

2. The s x (p:m) and s x r x (p:m) variance components tend to be larger when the persuasive mode is considered, thus resulting in less dependable scores when the persuasive mode is included. However, it would be reasonable to include all 3 modes in the assessment for curricular and instructional reasons as well as for providing content validity evidence.

References

Brennan, R. L. (1983). *Elements of Generalizability Theory*. Iowa City: American College Testing Program.

Shavelson, R. J. & Webb, N. M. (1981). *Generalizability Theory: A Primer*. London: Sage Publications.