Chapter XV: WRITING ASSESSMENT

Who takes it?
Student writing samples can be obtained from every level.

When is it administered?
Samples are obtained in the fall and spring semesters and are evaluated in the summer and/or subsequent semesters.

How long does it take for a student to complete the assessment?
No time—the writing samples require no extra work by the students. Submissions are obtained through writing students have already completed through their coursework.

What office administers it?
No office administers it. The Analytical Assessment is embedded in the Portfolio Project. The Collegial Review will be supervised by The Writing Across the University Committee.

Who originates the assessment?
The Writing Across the University Committee maintains oversight of the analytical and qualitative assessments.

When are results typically available?
Some assessment data is currently available about the Analytical Assessment. The Collegial Review is still being refined, although some information exists concerning the pilot project.

What type of information is sought?
The Analytical Assessment proposes to identify evidence demonstrating students' knowledge and skills in writing described in the Learning Outcomes for Writing-Enhanced courses. The Writing Across the University Committee will also review appropriate survey data (NSSE, GSQ, CSEQ) to assess students' writing habits, attitudes, etc. Additionally, the collegial review of faculty-selected writing samples from students in their classes will provide more qualitative data for faculty and other constituencies.

From whom are the results available?
The results will eventually be available through the Writing Across the University Committee.

To whom are the results regularly distributed?
The results will eventually be regularly distributed to the campus community through reports, presentations, and the *Assessment Almanac*.

Are results available by division or discipline?
No specific divisional or discipline results are currently available at this time.

Are results comparable to data of other universities?
No results are currently comparable to data of other universities.

In this chapter:

A. Analytical Writing Assessment Pilot
Summer 2004
Report and Recommendations
Natalie Alexander and Sue Pieper

The analytical writing assessment, a branch of a newly designed three-branch writing assessment initiative at Truman State University, was piloted in the summer of 2004. The goals of this new writing assessment are 1) to collect evidence of students' demonstration of the knowledge and skills in writing described in the Learning Outcomes for Writing-enhanced (WE) courses; 2) to inform faculty understanding of students' strengths and weaknesses in writing across the curriculum; 3) to promote continuing faculty reflection on what constitutes good writing and to provide opportunities for the further development of effective methods of teaching writing; and 4) to provide accountability to a variety of stakeholders, including students, faculty, staff, parents, and lawmakers.

Designed to be incorporated within the Liberal Arts and Sciences (LAS) Portfolio, the analytical writing assessment provides an efficient and effective means for assessing student writing across the University. In AY 2004-2005, students were asked, as part of the LAS Portfolio, to submit an example of "your best writing that demonstrates critical thinking skills." The complete prompt is included in Appendix A. For the purposes of the pilot, however, faculty read student LAS Portfolio selections that were submitted by students in AY 2003-2004 in response to a prompt asking them to submit a work that "demonstrates your best critical thinking from your academic career." Over the years, students have frequently responded to the critical thinking prompt by submitting a piece of their best writing, so it was anticipated that this particular portfolio submission would be appropriate to use for the writing assessment pilot.

Members of the Writing Assessment Committee worked collaboratively with the Director of the Portfolio Project to design a scoring rubric that assessed our WE student learning outcomes in four areas: critical thinking, organization, style, and mechanics. The scoring rubric is included in Appendix B. For every rubric item, students were assessed on a four-point scale ranging from zero to three. Every item was behaviorally anchored. In other words, for every rubric item scoring point, a description of behaviors typically associated with writing at that scoring point was provided. We expected that the behaviorally anchored rubric would both enhance reliability among readers as well as provide specific, helpful information regarding students' writing strengths and weaknesses to both faculty and students.

Reader training was provided by one of the members of the Assessment Committee with assistance from the Assessment Specialist and the Director of the Portfolio Project. The pilot writing assessment was conducted during one day of each of the three weeks of the LAS Portfolio reading with different foci each of the three weeks. The first week focused on discussion and fine-tuning of the rubric and identification of new rangefinders to be used in weeks two and three. Week two focused on a generalizability study to assess the reliability of the writing assessment design and make recommendations for future improvements. The third week focused on the use of the refined rubric and new rangefinder papers for reader training. Over the three week period, a total of 973 critical thinking/writing papers were read and assessed by LAS Portfolio readers, all Truman faculty members.

In response to recommendations from the Analytical Writing Assessment Pilot Summer 2004 Report as well as feedback from the Portfolio readers, small modifications were made to the writing assessment design.

This report has two purposes: 1) To report on the results of the generalizability study conducted in week two, and 2) To make suggestions for future improvements to the writing assessment based on the generalizability study and the overall pilot experience. The results of student performance for all three weeks of the writing assessment can be found in the 2003-2004 Assessment Almanac, Volume II.

G-Study Results

The results of the generalizability (G) study begin with a report of student performance results. The mean scores for each of the rubric items across all of the students whose papers were assessed in the second week of the analytical writing assessment are tabled and described. Following the student performance results, the item and rater means are listed along with the grand mean across items and raters. The phi reliability coefficient is reported in this text because the goal of the analytical writing assessment is to estimate the writing proficiency level of Truman students using a criterion-referenced rubric rather than to rank order students. In subsequent sections of the report, the dependability of the original design using 3-rater teams and 4 rubric items is investigated and the reliability projected when varying the number of raters, items, or both.

Student Performance Results

A total of 120 graduating students' papers were read for the generalizability study. Table 1 displays the item means or average rating for each rubric item across 100 students. The additional 20 students' papers were read by a team who lost one reader midway through the day, and those results are be reported in Appendix C.

Table 1

Item Means Across All Students (0-3 Scale)

Rubric Item	Mean Score
1. Critical Thinking	1.8
2. Organization	2.0
3. Style	2.0
4. Mechanics	2.1

As shown by Table 1, students performed fairly consistently in the upper middle of the zero to three score range. Students performed best at mechanics, with a mean rating of 2.1. A 2 rating for mechanics indicates that students "demonstrated adequate command of mechanical conventions" and "errors were minimally distracting to readers." Performance in mechanics was closely followed by performance in style and organization. Students scored a mean of 2.0 in each of these areas. A 2 rating in style indicates that students "showed audience awareness" and "used appropriate words, sentence structure, and stylistic conventions" but that their writing "contained occasional lapses in tone or voice." A 2 rating in organization indicates that students "included an adequate introduction and conclusion," "displayed an adequate controlling idea," and "exhibited adequate clarity and logical structure." Students performed slightly less well in critical thinking, with a mean of 1.8. A 2 rating in critical thinking is described as writing that "develops ideas with some consistency and depth and adequate support," "makes some good connections between ideas," "shows some analysis, or synthesis, or interpretation," and "displays some skill at integrating ideas to make meaning."

G-study Results

To investigate the dependability of the measurement design, a fully crossed, two-facet random effects model was used. In other words, two facets (faculty raters and rubric items) represent a larger domain and influence the object of measurement (students). A set of rubric items representative of a universe of items that sample the writing domain guide the rating of student writing performance. Similarly, the

raters of student performance are representative of a universe of acceptable raters of writing proficiency. Because the pairing of any rater with any item is accepted as meaningful, the design is described as being "crossed." This interaction can be estimated because raters assigned to a team evaluated the same papers.

In the assessment of student writing, individual differences exist among students' writing proficiency, as we would expect. However, other differences exist in our writing assessment that we want to study and need to control, including rubric item difficulty, rater severity, students' interactions with specific rubric items, rater's ratings of specific students, raters' ratings of specific rubric items, and a "residual" of unspecified and random error. These components all contribute to imprecision in our students' scores, and they are called the "variance components" of a G-study. Using generalizability theory, we are able to estimate the variance components for each source of measurement error we have identified (faculty raters and rubric items), all interactions between the facets, and the remaining error along with the variance component for the object of measurement (students). Variance that is not attributable to actual student writing proficiency, therefore, is identified by its measurement source and defined as error.

Generalizability theory provides two phases of study: a generalizability (G) study and a decision (D) study (Shavelson & Webb, 1991). The purpose of a G-study is to develop the measurement procedure. The G-study estimates each of the variance components and allows for comparisons to be made to determine which sources of error are most troublesome. The purpose of a D-study is to assist in actual decision-making and to improve the measurement design for future applications. The D-study allows us to estimate the reliability of ratings given the current design, as well as the predicted reliability that would result with modifications to the number of raters, items, of both. Two coefficients are produced that summarize the G-study results: a generalizability (g) coefficient and a phi coefficient. A g coefficient is selected when making relative decisions where the interest is in rank ordering students. Phi coefficients are used for making absolute decisions where the interest is in students attaining a particular level of mastery or competency.

Using the 4-item rubric, each writing sample was assessed by a 3-rater team. Each 3-rater team read 20 papers, for a total of 100 papers read by five teams. Table 2 provides means for the teams by rubric items, rater means across rubric items, the grand mean, and the coefficient indicating the dependability of the design.

Table 2

Item and Rater Means by Teams (0-3 Scale)

Rubric Item Number	Team 1	Team 2	Team 4	Team 5	Team 6
1. Critical Thinking	1.8	1.5	1.6	2.0	2.1
2. Organization	2.1	1.6	1.8	2.0	2.2
3. Style	2.1	1.7	1.9	2.1	2.2
4. Mechanics	2.2	1.8	2.1	2.2	2.3
Rater 1 Mean	2.4	1.3	1.8	2.2	2.2
Rater 2 Mean	1.8	1.7	2.0	1.9	2.2
Rater 3 Mean	2.1	1.9	1.8	2.1	2.2
Grand Mean	2.1	1.6	1.9	2.1	2.2
Dependability (Phi Φ)	.68	.73	.84	.77	.66

The range of item means used by the rater teams is small, consistent with the fact that very low rater variance is reported in the G-study (Table 4) for three out of five of the teams. This finding indicates that, for the most part, the raters used the scales consistently. In other words, most of the teams rated the writing criteria similarly. A rating pattern is easily detectable: ratings for rubric item 4 (mechanics) are consistently higher than for the other rubric items.

The phi coefficients reported for the five teams are promising, with three of the five teams (teams two, four, and five) achieving phi coefficients approaching or exceeding the .75 level of desirability for making group proficiency decisions based on criteria. An examination of the error variance associated with the measurement facets will further illuminate the reported phi coefficients and what contributed to a lower than desired phi for teams one and six.

Table 3 presents the percentage of variability accounted for by each of the variance components included in the design. The variance in the objects of measurement, persons (or, in this case, students), is desirable because it reflects the amount of variability in the writing sample that is due to actual individual differences in student writing proficiency. Variance that is attributable to raters, items, and the interactions between these sources of error, however, is not desirable.

Table 3

G-Study Percent of Error Variance Associated with Facets of Measurement

	Percent Variance				
Source of Variance	Team 1	Team 2	Team 4	Team 5	Team 6
Persons	28	37	52	38	31
Raters	12	10	1	1	0
Items	1	1	5	0	0
Persons x Raters	18	19	16	18	39
Persons x Items	0	5	3	12	4
Items x Raters	9	2	0	0	2
Error (Persons x Raters x Items)	32	26	23	32	24
Total*	100	100	100	101	100

* The percentages do not sum exactly to 100 due to rounding.

As can be seen in Table 3, relatively little variance can be attributed to raters, items, persons by items, or items by raters. The largest variance component is the interaction among persons, raters, and items. Also included in this component are sources of variability arising out of "randomness, other systematic but unidentified or unknown sources of variability, or both" (Shavelson & Webb, 1991, p.6). Because these sources of variability cannot be disentangled, it is difficult to interpret this variance component.

The second largest variance component is persons by raters, accounting for anywhere from 16% of the variance for team four to 39% of the variance in team six. These findings indicate that the raters may have differed in their ratings for different individual students. The problem may be linked to lack of systematic rubric use by the raters, but is more likely the result of the selection of papers submitted in response to last year's critical thinking prompt. As was mentioned earlier, for the writing assessment pilot we used papers students submitted in response to a critical thinking prompt rather than a writing prompt. Some of those papers might not have been a good "fit" for the writing assessment and consequently may have posed challenges to the raters in terms of evaluating them.

Given that person by item variance was the second largest source of error for all teams, why is the phi coefficient noticeably lower for teams one and six? Two factors contributed to lower reliability for these two teams: lower person variance and higher error variance for both teams. As indicated above, person variance is desirable because it reflects individual differences in student writing. The lower person variance suggests that the students evaluated by teams one and six were treated as being more alike in proficiency than the students evaluated by teams two, four, and five. This could mean that raters on teams one and six were inappropriately using less of the scale. Raters on teams one and six, for example, might be judging students more by their own personal expectations than by the collaboratively developed rubric. On the other hand, the student writing samples for teams one and six might legitimately be more alike.

Lower person variance in teams one and six resulted in higher error variance for both teams. For team one, variance was split between persons by raters and raters. In other words, not only did team raters differ in scoring severity, but they also differed in their scoring of different individual students. For team six, nearly all of the error variance not accounted for by the interaction among persons, raters, and items was persons by raters' variance. In fact, the error variance for persons by raters for team six is more than twice the error variance for teams two, four, and five.

D-study Results

Table 4 presents the phi coefficients associated with varying both the number of raters and the number of items. As shown in Table 4, the phi coefficient increases very gradually with the addition of either raters or items. As expected, teams two, four, and five gain very little with the addition of raters and items, having already achieved acceptable phis. Teams one and six, however, would need to add one rater and one item in order to attain reliability similar to teams two, four, and five.

Table 4

Phi Coefficient Estimates Based on Alternative Designs for Each Team

	Team 1			Team 2			Team 4			Team 5			Team 6		
	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi
Current Design	3	4	.68	3	4	.73	3	4	.84	3	4	.77	3	4	.66
Alternative Designs	2	3	.56	2	3	.64	2	3	.77	2	3	.68	2	3	.55
	2	4	.58	2	4	.66	2	4	.79	2	4	.71	2	4	.57
	2	5	.60	2	5	.67	2	5	.80	2	5	.72	2	5	.58
	3	3	.66	3	3	.72	3	3	.82	3	3	.74	3	3	.65
	3	4	.68	3	4	.73	3	4	.84	3	4	.77	3	4	.66
	3	5	.69	3	5	.74	3	5	.85	3	5	.79	3	5	.67
	4	3	.72	4	3	.76	4	3	.85	4	3	.78	4	3	.70
	4	4	.74	4	4	.78	4	4	.87	4	4	.80	4	4	.72
	4	5	.75	4	5	.79	4	5	.88	4	5	.82	4	5	.73

Recommendations

In summary, students performed fairly consistently in the upper middle of the four-point score range on the analytical writing assessment. Students performed best at mechanics and less well in critical thinking. The rubric items seem to be functioning well with very little error variance associated with rubric items. The effectiveness of the rubric may be attributed to a number of factors including the careful crafting of the rubric by the Writing Assessment Committee, the tying of rubric items to already identified WE outcomes for writing across the University, and the continued discussion of the rubric over the portfolio reading weeks. The process of using the first week of portfolio reading for identification of rangefinders, the second week for the g-study, and the third week for use of the refined rubric and new rangefinder papers worked well. Also effective was the collaboration between the Writing Assessment Committee, the Assessment Specialist, and the Portfolio Director in the administration of the analytical writing assessment as part of the LAS Portfolio.

Although the current writing assessment appears to have several strengths, a few concerns exist. The large person by rater error variance across all teams indicates that at least some portfolio submissions were problematic for readers. It is reasonable to assume that these problematic submissions were a result of students responding to the AY 2003-2004 prompt asking for work that demonstrated critical thinking rather than writing. Two teams out of five also had high rater variance, meaning the raters differed in the severity of their ratings. Because care was taken to select team members that mixed experienced and inexperienced portfolio readers, this problematic rater variance can most likely be attributed to a new rater training process and new rangefinder papers. However, we can most likely increase reliability by ensuring that raters bring both a range of experience and disciplinary perspectives to each team.

Overall, the results of the G-study indicate that we should continue to use the same administrative strategy over the three weeks of portfolio reading and continue to use the current rubric with minor modifications as needed. At the same time, we will need to continue to refine the rater training process, including the use of effective rangefinder papers. Perhaps most importantly, we will want to assess the effectiveness of the new writing prompt that is designed to elicit writing that demonstrates critical thinking, as well as solid organization, style, and mechanics

Although the D-study indicates that adding one rater and one rubric item to the current design would bring teams one and six up to a desirable level of reliability, adding a new rater to each team would be costly in terms of time and money. Furthermore, the current rubric appears to be working well for faculty raters. Therefore, it is recommended that the current design be continued and that another G-study be conducted in AY 2004-2005 to determine if changes in the writing prompt have a positive impact on the dependability of the writing assessment.

Appendix A

Critical Thinking and Writing

(Your submissions in this category may be used for other categories.)

Name _____ Major _____

Banner ID _____

Please complete and send this document along with your submission.

Please include an example of your best writing that demonstrates your critical thinking skills. As stated in Truman's LSP outcomes, good writing is a reflection of good thinking. Thus, as a result of an intellectual process that communicates meaning to a reader, good writing integrates ideas through analysis, evaluation, and the synthesis of ideas and concepts. Good writing also exhibits skill in language usage and clarity of expression through good organization.

Faculty readers will evaluate your writing sample with attention to four areas:

1. **Thinking (developing ideas, making connections between ideas, integrating ideas to make meaning) For further information regarding the nature of critical thinking, review the prompt entitled "Critical Thinking Definitions".**
2. **Organization (communicating a purpose, writing clearly, making strong arguments, drawing conclusions)**
3. **Style (employing appropriate voice and tone, having an audience in mind, choosing appropriate words, using appropriate sentence structures)**
4. **Mechanics (adhering to the accepted conventions of grammar and punctuation, spelling words correctly)**

As you consider this category, you may find that a submission from another category demonstrates strong critical thinking and writing. If so, feel free to use that item for this category as well.

NOTE: Writing samples from ENG 190 ("Writing as Critical Thinking") are generally NOT the best examples of critical thinking.

♦Source of this entry?

Course number and name: _____

Which best describes this course? ____LSP ____Major ____Minor ____Elective

OR other source: (describe specifically) _____

_____

♦In which year did you originally produce this work? ____ Fr. ____So. ____Ju. ____Sr.

♦Was this work the result of collaboration? ____Yes ____No

♦Please describe the instructor's assignment. If the work was not generated by an assignment, please describe your purpose and process in using this kind of thinking.

Please reflect on the kinds of thinking you engage in with this work.

Also comment on how you have grown in critical thinking skills since arriving at Truman.

This self-assessment is as important to us as the work you submit, and we will read it with care.

Appendix B

Rubric for Analytical Writing Assessment

	0	1	2	3
Critical Thinking	displays no real development of ideas lacks convincing support exhibits no attempt to make connections between ideas includes no real analysis, or synthesis, or interpretation, or … demonstrates no real integration of ideas (the author's or those of others) to make meaning	develops ideas superficially or inconsistently provides weak support begins to make connections between ideas begins to analyze, or synthesize, or interpret, or … begins to integrate ideas (the author's or those of others) to make meaning	develops ideas with some consistency and depth develops adequate support makes some good connections between ideas shows some analysis, or synthesis, or interpretation, or … displays some skill at integrating ideas (the author's or those of others) to make meaning	displays insight and thorough development of ideas develops consistently strong support reveals mature and thoughtful connections between ideas shows sophistication in analysis, or synthesis, or interpretation, or … is adept at integrating ideas (the authors or those of others) to make meaning
Organization	lacks introduction lacks controlling idea lacks clarity lacks logical structure lacks conclusion	includes weak introduction displays controlling idea exhibits weak clarity exhibits weak logical structure includes weak conclusion	includes adequate introduction displays adequately developed controlling idea exhibits adequate clarity exhibits adequate logical structure includes adequate conclusion	includes strong introduction displays clear, well-developed controlling idea exhibits excellent clarity exhibits strong logical structure includes well-supported conclusion
Style	tone or voice is off-putting seems to have no audience in mind frequently chooses inappropriate words exhibits frequent inappropriate sentence structure uses no appropriate stylistic conventions	contains inconsistent tone or voice shows little audience awareness sometimes chooses inappropriate words exhibits occasional inappropriate sentence structure uses few appropriate stylistic conventions	contains occasional lapses in tone or voice shows audience awareness chooses appropriate words exhibits appropriate sentence structure uses appropriate stylistic conventions	maintains a consistent tone and voice shows consistent audience awareness exhibits skill in word choice exhibits sophisticated sentence structure skillfully uses appropriate stylistic conventions
Mechanics	lacks command of mechanical conventions: grammar, punctuation, or spelling errors present major distraction to readers	Demonstrates weak command of mechanical conventions: grammar, punctuation, or spelling errors are occasionally distracting to readers	demonstrates adequate command of mechanical conventions: grammar, punctuation, or spelling errors are minimally distracting to readers	demonstrates excellent command of mechanical conventions: grammar, punctuation, and spelling small errors do not distract readers

Appendix C

Item and Rater Means for Team Three (0-3 Scale)

Rubric Item Number	Team 3
1. Critical Thinking	1.7
2. Organization	2.1
3. Style	2.1
4. Mechanics	2.1
Rater 1 Mean	1.8
Rater 2 Mean	2.3
Grand Mean	2.0
Dependability (Phi Φ)	.30

G-Study Percent of Error Variance Associated with Facets of Measurement for Team Three

	Percent Variance
Source of Variance	Team 3
Persons	.12
Raters	.19
Items	.04
Persons x Raters	.24
Persons x Items	.07
Items x Raters	.02
Error (Persons x Raters x Items)	.32
Total	100

Phi Coefficient Estimates Based on Alternative Designs for Team Three

	Team 3		
	Raters	Items	Phi
Current Design	2	4	.30
Alternative Designs	2	3	.28
	2	4	.30
	2	5	.31
	3	3	.36
	3	4	.38
	3	5	.39
	4	3	.41
	4	4	.44
	4	5	.45

Analytical Writing Assessment
Summer 2005
Report and Recommendations
Sue Pieper, Assessment Specialist

The analytical writing assessment, one branch of the new three-branch writing assessment at Truman State University, was implemented in the summer of 2005 following a pilot administration in the summer of 2004. The goals of this new writing assessment are 1) to collect evidence of students' demonstration of the knowledge and skills in writing described in the Learning Outcomes for Writing-enhanced (WE) courses; 2) to inform faculty understanding of students' strengths and weaknesses in writing across the curriculum; 3) to promote continuing faculty reflection on what constitutes good writing and to provide opportunities for the further development of effective methods of teaching writing; and 4) to provide accountability to a variety of stakeholders, including students, faculty, staff, parents, and lawmakers. Designed to be incorporated within the Liberal Arts and Sciences (LAS) Portfolio, the analytical writing assessment provides an efficient and effective means for assessing student writing across the University.

For the 2004 pilot administration of the analytical writing assessment, faculty read LAS Portfolio selections that were submitted by students in AY 2003-2004 in response to a prompt asking them to submit a work that "demonstrates your best critical thinking from your academic career." In AY 2004-2005, however, students were given a prompt that more closely aligned with the analytical writing assessment evaluation criteria. Students were asked, as part of the LAS Portfolio, to submit an example of "your best writing that demonstrates critical thinking skills." Additionally, students were given the list of criteria on which their critical thinking/writing submission would be evaluated. The complete prompt is included in Appendix A. It was anticipated that the new prompt would ease some of the difficulties raters experienced last year with evaluating student papers. Results of a generalizability (G) study conducted during the 2004 pilot showed that some portfolio submissions were problematic for readers. It is reasonable to assume that these problematic papers were at least partially the result of students responding to the AY 2003-2004 prompt asking for work that demonstrated critical thinking, but not necessarily writing. The new prompt, designed to elicit writing that demonstrates critical thinking as well as solid organization, style, and mechanics, was expected to produce more appropriate portfolio submissions.

The scoring rubric, refined during the first week of the 2004 pilot administration based on reader feedback, assesses our WE student learning outcomes in four areas: critical thinking, organization, style, and mechanics. Based on the results of the 2004 G-study and reader feedback indicating that the rubric worked well for raters and students, the rubric was used in 2005 without further modifications. For every rubric item, students were assessed on a four-point scale ranging from zero to three. Every item was behaviorally anchored. In other words, for every rubric item scoring point, a description of behaviors typically associated with writing at that scoring point was provided. The scoring rubric is included in Appendix B. We expected that the behaviorally anchored rubric would both enhance reliability among readers as well as provide specific, helpful information regarding students' writing strengths and weaknesses to both faculty and students.

Reader training was once again provided by one of the members of the Assessment Committee with assistance from the Assessment Specialist and the Director of the Portfolio Project. Based on the findings of the 2004 G-study as well as the feedback from readers, the rater training protocol was not changed, including the continuous effort to use student papers for rater training that demonstrated the entire range of student behaviors described by the rubric in order to enhance reliability.

Similar to the 2004 pilot, the 2005 writing assessment was conducted during one day of each of the three weeks of the LAS Portfolio reading. In response to 2004 reader feedback, however, the analytical writing

assessment was moved from later to earlier in the Portfolio reading week in order to alleviate reader fatigue. Once again, a generalizability study was conducted in week two to assess the reliability of the writing assessment design and make recommendations for future improvements. Over the three week period, a total of 1,040 critical thinking/writing papers were read and assessed by LAS Portfolio readers, all Truman faculty members.

This report has two purposes: 1) To report on the results of the generalizability study conducted in week two of the analytical writing assessment, and 2) To make suggestions for future improvements to the writing assessment based on the generalizability study and the overall writing assessment experience. The results of student performance for all three weeks of the writing assessment can be found in the 2004-2005 Assessment Almanac, Volume II.

G-Study Results

The results of the generalizability (G) study begin with a report of student performance results. The mean scores for each of the rubric items across all of the students whose papers were assessed in the second week of the analytical writing assessment are tabled and described. Following the student performance results, the item and rater means are listed along with the grand mean across items and raters. The phi reliability coefficient is reported in this text because the goal of the analytical writing assessment is to estimate the writing proficiency level of Truman students using a criterion-referenced rubric rather than to rank order students. In subsequent sections of the report, the dependability of the original design using three-rater teams and four rubric items is investigated and the reliability projected when varying the number of raters, items, or both.

Student Performance Results

A total of 156 graduating students' papers were read for the generalizability study. Table 1 displays the item means or average rating for each rubric item across students.

Table 1

Item Means Across All Students (0-3 Scale)

Rubric Item	Mean Score
2. Critical Thinking	1.8
2. Organization	2.0
3. Style	2.1
4. Mechanics	2.1

As shown by Table 1, students performed fairly consistently in the upper middle of the zero to three score range. Students performed best at style and mechanics, with a mean rating of 2.1 in each of these areas. A 2 rating in style indicates that students "showed audience awareness" and "used appropriate words, sentence structure, and stylistic conventions" but that their writing "contained occasional lapses in tone or voice." A 2 rating for mechanics indicates that students "demonstrated adequate command of mechanical conventions" and "errors were minimally distracting to readers." Students' performance on style and mechanics was closely followed by organization. Students scored a mean of 2.0 in organization. A 2 rating in organization indicates that students "included an adequate introduction and conclusion," "displayed an adequate controlling idea," and "exhibited adequate clarity and logical structure." Students performed slightly less well in critical thinking, with a mean of 1.8. A 2 rating in critical thinking is described as writing that "develops ideas with some consistency and depth and adequate support," "makes

some good connections between ideas," "shows some analysis, or synthesis, or interpretation," and "displays some skill at integrating ideas to make meaning." The performance results are nearly identical to the results from the 2004 G-study, with the exception of style. In 2004, the mean score for style was 2.0.

G-study Results

To investigate the dependability of the measurement design, a fully crossed, two-facet random effects model was used. In other words, two facets (faculty raters and rubric items) represent a larger domain and influence the object of measurement (students). A set of rubric items representative of a universe of items that sample the writing domain guide the rating of student writing performance. Similarly, the raters of student performance are representative of a universe of acceptable raters of writing proficiency. Because the pairing of any rater with any item is accepted as meaningful, the design is described as being "crossed." This interaction can be estimated because raters assigned to a team evaluated the same papers.

In the assessment of student writing, individual differences exist among students' writing proficiency, as we would expect. However, other differences exist in our writing assessment that we want to study and need to control, including rubric item difficulty, rater severity, students' interactions with specific rubric items, rater's ratings of specific students, raters' ratings of specific rubric items, and a "residual" of unspecified and random error. These components all contribute to imprecision in our students' scores, and they are called the "variance components" of a G-study. Using generalizability theory, we are able to estimate the variance components for each source of measurement error we have identified (faculty raters and rubric items), all interactions between the facets, and the remaining error along with the variance component for the object of measurement (students). Variance that is not attributable to actual student writing proficiency, therefore, is identified by its measurement source and defined as error.

Generalizability theory provides two phases of study: a generalizability (G) study and a decision (D) study (Shavelson & Webb, 1991). The purpose of a G-study is to develop the measurement procedure. The G-study estimates each of the variance components and allows for comparisons to be made to determine which sources of error are most troublesome. The purpose of a D-study is to assist in actual decision-making and to improve the measurement design for future applications. The D-study allows us to estimate the reliability of ratings given the current design, as well as the predicted reliability that would result with modifications to the number of raters, items, of both. Two coefficients are produced that summarize the G-study results: a generalizability (g) coefficient and a phi coefficient. A g coefficient is selected when making relative decisions where the interest is in rank ordering students. Phi coefficients are used for making absolute decisions where the interest is in students attaining a particular level of mastery or competency.

LAS Portfolio readers were grouped into six teams of three raters each. In response to recommendations from the 2004 G-study to enhance team "diversity" in order to improve reliability, an effort was made to mix raters from all three major groups (Arts/Humanities, Science/Math, Professional) within each team along with experienced and inexperienced Portfolio readers, male and female readers, and fast and slow readers. All three raters on a given "virtual team" assessed the same set of papers on-line. A total of 156 papers were read by six teams. Table 2 provides means for the teams by rubric items, rater means across rubric items, the grand mean, and the coefficient indicating the dependability of the design.

Table 2

Item and Rater Means by Teams (0-3 Scale)

Rubric Item Number	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6
1. Critical Thinking	1.5	2.0	1.6	2.0	1.9	1.5
2. Organization	1.8	2.2	1.8	2.5	2.2	1.8
3. Style	1.8	2.3	2.0	2.4	2.1	1.7
4. Mechanics	1.9	2.3	2.0	2.3	2.1	1.7
Rater 1 Mean	1.7	2.0	2.0	2.2	2.0	1.9
Rater 2 Mean	1.8	2.6	1.6	2.4	2.3	2.1
Rater 3 Mean	1.8	2.0	1.9	2.4	1.9	1.0
Grand Mean	1.8	2.2	1.9	2.3	2.1	1.7
Dependability (Phi Φ)	.57	.38	.75	.72	.63	.52

The range of item means used by the rater teams is small, consistent with the fact that very low rater variance is reported in the G-study (Table 4) for four out of six teams. This finding indicates that, for the most part, the raters used the scales consistently. In other words, most of the teams rated the writing criteria similarly. A rating pattern is easily detectable: ratings for rubric item 1 (critical thinking) are consistently lower than for the other rubric items.

The phi coefficients reported for the six teams are not as promising as those reported in the 2004 G-study. Only two of the six teams (teams three and four) approached or met the .75 level of desirability for making group proficiency decisions based on criteria. An examination of the error variance associated with the measurement facets will further illuminate the reported phi coefficients and what contributed to a lower than desired phi for teams one, two, five, and six.

Table 3 presents the percentage of variability accounted for by each of the variance components included in the design for each team. The variance in the objects of measurement, persons (or, in this case, students), is desirable because it reflects the amount of variability in the writing sample that is due to actual individual differences in student writing proficiency. Variance that is attributable to raters, items, and the interactions between these sources of error, however, is not desirable.

Table 3

G-Study Percent of Error Variance Associated with Facets of Measurement

	Percent Variance					
Source of Variance	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6
Persons	**22**	**11**	**39**	**27**	**25**	**22**
Raters	0	21	5	0	4	37
Items	3	1	3	4	1	0
Persons x Raters	31	21	23	9	25	16
Persons x Items	7	0	6	7	6	3
Items x Raters	5	5	1	7	0	5
Error (Persons x Raters x Items)	32	40	24	45	39	17
Total*	100	99	101	99	100	100

* The percentages do not sum exactly to 100 due to rounding.

As can be seen in Table 3, relatively little variance can be attributed to, items, persons by items, or items by raters. The largest variance component is the interaction among persons, raters, and items. Also included in this component are sources of variability arising out of "randomness, other systematic but unidentified or unknown sources of variability, or both" (Shavelson & Webb, 1991, p.6). Because these sources of variability cannot be disentangled, it is difficult to interpret this variance component.

Similar to the results of the 2004 G-study, the second largest variance component is persons by raters, accounting for anywhere from 9% of the variance for team four to 31% of the variance in team six. These findings indicate that the raters once again differed in their ratings for different individual students. The problem may be linked to lack of systematic rubric use by the raters, but is more likely the result of the selection of papers submitted in response to the prompt. Although it was anticipated that students would submit more appropriate selections this year as a result of the newly revised prompt, these results indicate that some Portfolio submissions continue to pose challenges to the raters in terms of evaluating them.

Given that person by item variance was the second largest source of error for all teams, why is the phi coefficient noticeably lower for teams one, two and six? Two factors contributed to lower reliability for these three teams: lower person variance and higher error variance for both teams. As indicated above, person variance is desirable because it reflects individual differences in student writing. The lower person variance suggests that the students evaluated by teams one, two, and six were treated as being more alike in proficiency than the students evaluated by teams three, four, and five. This could mean that raters on teams one, two, and six were inappropriately using less of the scale. Raters on teams one, two, and six, for example, might be judging students more by their own personal expectations than by the

collaboratively developed rubric. On the other hand, the student writing samples for teams one, two, and six might legitimately be more alike.

Lower person variance in teams one, two, and six resulted in higher error variance for all three teams. Team one has the greatest percentage of persons by rater variance of all of the teams with additional variance spread out among items, persons by items, and items by raters. Teams two and six have similar patterns of error variance, with both teams showing a large amount of variance split between persons by raters and raters. In other words, in addition to team raters differing in their ratings of specific students, raters also differed in the severity of their scoring.

In addition to the noticeably lower reliability reported for teams one, two, and six, team five also reported a lower than desirable phi coefficient. Team five has a large amount of person by raters error variance with additional variance spread out among raters, items, and persons by items.

D-study Results

Table 4 presents the phi coefficients associated with the current writing assessment design followed by the phi coefficients projected by varying the number of raters, the number of rubric items, or both the number of raters and the number of items. Team five approached an acceptable phi with the addition of one rater and one item. Teams one, two, and six, however, still did not increase their phis to a level of acceptability by adding one rater and one item.

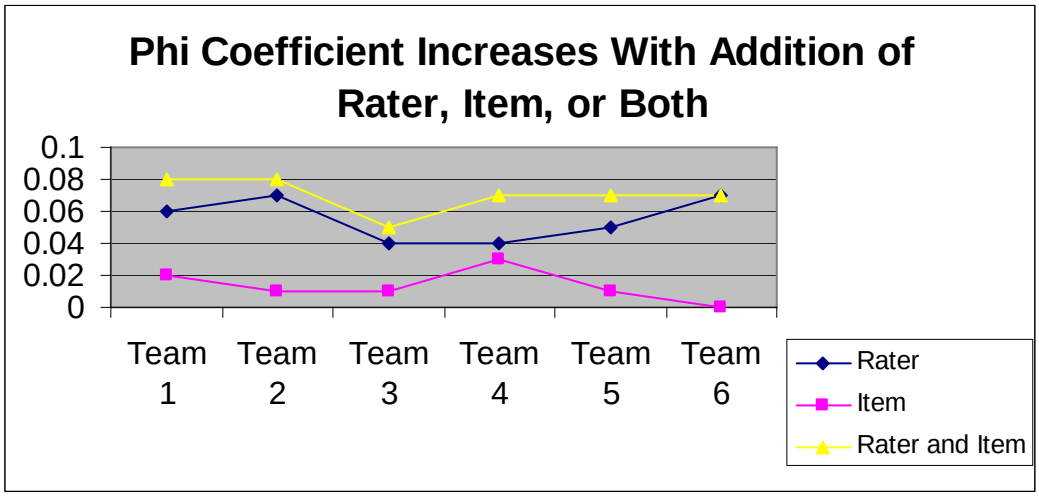
Table 4

Phi Coefficient Estimates Based on Alternative Designs for Each Team

	Team 1			Team 2			Team 3			Team 4			Team 5			Team 6		
	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi	Raters	Items	Phi
Current Design	3	4	.57	3	4	.38	3	4	.75	3	4	.72	3	4	.63	3	4	.52
Alternative Designs	2	3	.46	2	3	.28	2	3	.65	2	3	.61	2	3	.51	2	3	.41
	2	4	.49	2	4	.29	2	4	.67	2	4	.66	2	4	.54	2	4	.42
	2	5	.50	2	5	.30	2	5	.69	2	5	.69	2	5	.55	2	5	.43
	3	3	.54	3	3	.36	3	3	.73	3	3	.68	3	3	.60	3	3	.51
	3	4	.57	3	4	.38	3	4	.75	3	4	.72	3	4	.63	3	4	.52
	3	5	.59	3	5	.39	3	5	.76	3	5	.75	3	5	.64	3	5	.52
	4	3	.60	4	3	.43	4	3	.77	4	3	.72	4	3	.66	4	3	.58
	4	4	.63	4	4	.45	4	4	.79	4	4	.76	4	4	.68	4	4	.59
	4	5	.65	4	5	.46	4	5	.80	4	5	.79	4	5	.70	4	5	.59

Figure 1 shows the phi coefficient increases projected for each team with the addition of both a rater and a rubric item, as well as a rater only or an item only to the current design. Although adding both a rater and an item would achieve maximum impact, adding a rater alone would be almost as beneficial. Clearly, all teams could benefit more by adding a rater than adding an item. This finding is understandable given the large amount of error variance that is attributable to raters in the current design. Averaging over more measurement conditions should reduce error while simultaneously increasing reliability.

Figure 1



Recommendations

In summary, the results of the G-study showed that, similar to the 2004 results, students performed fairly consistently in the upper middle of the four-point score range on the analytical writing assessment. Students performed best at style and mechanics and less well in critical thinking. Feedback from the Writing Across the University committee member who conducted the rater training, the Director of the Portfolio Project, and the Assessment Specialist indicated that conducting the G-study in the second week of the Portfolio readings, similar to 2004, was successful. This schedule allowed everyone to reacquaint themselves with the rubric and the rating process during the first week and to focus more on "just reading" in the third and final week of the Portfolio readings. Additionally, rater feedback indicated that moving the analytical writing assessment to the earlier in the week helped to alleviate fatigue and allowed for raters to make a smoother transition from their own classroom evaluation methods to assessment using a behaviorally anchored rubric.

Similar to 2004, the rubric seemed to be functioning well with very little error variance associated with rubric items. The effectiveness of the rubric may be attributed to a number of factors including the careful crafting of the rubric by the Writing Assessment Committee, the tying of rubric items to already identified WE outcomes for writing across the University, and the continued discussion of the rubric over the portfolio reading weeks. Overall, the results of the G-study as well as rater feedback indicated that we should continue to use the same administrative strategy over the three weeks of portfolio reading and continue to use the current rubric with minor modifications as needed.

Although the current writing assessment appears to have several strengths, a few concerns persist. In 2005, like in 2004, large person by rater error variance across all teams indicated that at least some portfolio submissions were problematic for readers. Although we expected that the introduction of a new prompt asking students to submit their best writing that demonstrates critical thinking skills would elicit

more appropriate submissions, it is clear that some submissions continue to be difficult for raters to assess. Two teams out of six also reported high rater variance, similar to the results from 2004, meaning the raters differed in the severity of their ratings. Because care was taken to improve the diversity of the teams in order to enhance reliability, the persistence of high rater variance in these teams was puzzling. A preliminary examination of the raters in these two teams revealed, however, that although some raters had Portfolio reading experience, this experience might have been fairly minimal. Conversely, those teams with the highest reliability appeared to have raters with more substantial experience with Portfolio reading or other writing assessment reading.

Although results of the D-study indicated that adding one rater and one rubric item to the current design would enhance reliability, redesigning the writing assessment to include multiple raters for each student paper would be costly in terms of time and money. Furthermore, the current four-item rubric appeared to be working well for faculty raters. It is recommended instead that problematic person by rater and rater variance be addressed in two ways. First, the rater training process itself needs to be refined to consider a variety of different kinds of papers, including problematic papers. We have already identified one particular assignment that appears to be problematic for raters. We need to identify other problematic papers that we can use for rater training and discussion. Second, communications need to be improved with the entire campus community—students, faculty, and staff—regarding the purpose of the writing assessment, the writing/critical thinking prompt, the criteria for evaluation, and the use of results. Finally, it is recommended that another G-study be conducted again in 2006-2007 after these initiatives have been implemented. This follow-up G-study will assist us in determining if the recommended changes have had a positive impact on the dependability of the writing assessment.

Appendix A

Critical Thinking and Writing

(Your submissions in this category may be used for other categories.)

Name _____ Major _____

Banner ID _____

Please complete and send this document along with your submission.

Please include an example of your best writing that demonstrates your critical thinking skills. As stated in Truman's LSP outcomes, good writing is a reflection of good thinking. Thus, as a result of an intellectual process that communicates meaning to a reader, good writing integrates ideas through analysis, evaluation, and the synthesis of ideas and concepts. Good writing also exhibits skill in language usage and clarity of expression through good organization.

Faculty readers will evaluate your writing sample with attention to four areas:

5. **Thinking (developing ideas, making connections between ideas, integrating ideas to make meaning) For further information regarding the nature of critical thinking, review the prompt entitled "Critical Thinking Definitions".**
6. **Organization (communicating a purpose, writing clearly, making strong arguments, drawing conclusions)**
7. **Style (employing appropriate voice and tone, having an audience in mind, choosing appropriate words, using appropriate sentence structures)**
8. **Mechanics (adhering to the accepted conventions of grammar and punctuation, spelling words correctly)**

As you consider this category, you may find that a submission from another category demonstrates strong critical thinking and writing. If so, feel free to use that item for this category as well.

NOTE: Writing samples from ENG 190 ("Writing as Critical Thinking") are generally NOT the best examples of critical thinking.

♦Source of this entry?

Course number and name: _____

Which best describes this course? ____LSP ____Major ____Minor ____Elective

OR other source: (describe specifically) _____

_____

♦In which year did you originally produce this work? ____ Fr. ____So. ____Ju. ____Sr.

♦Was this work the result of collaboration? ____Yes ____No

♦Please describe the instructor's assignment. If the work was not generated by an assignment, please describe your purpose and process in using this kind of thinking.

Please reflect on the kinds of thinking you engage in with this work.

Also comment on how you have grown in critical thinking skills since arriving at Truman.

This self-assessment is as important to us as the work you submit, and we will read it with care.

Appendix B

Rubric for Analytical Writing Assessment

	0	1	2	3
Critical Thinking	displays no real development of ideas lacks convincing support exhibits no attempt to make connections between ideas includes no real analysis, or synthesis, or interpretation, or … demonstrates no real integration of ideas (the author's or those of others) to make meaning	develops ideas superficially or inconsistently provides weak support begins to make connections between ideas begins to analyze, or synthesize, or interpret, or … begins to integrate ideas (the author's or those of others) to make meaning	develops ideas with some consistency and depth develops adequate support makes some good connections between ideas shows some analysis, or synthesis, or interpretation, or … displays some skill at integrating ideas (the author's or those of others) to make meaning	displays insight and thorough development of ideas develops consistently strong support reveals mature and thoughtful connections between ideas shows sophistication in analysis, or synthesis, or interpretation, or … is adept at integrating ideas (the authors or those of others) to make meaning
Organization	lacks introduction lacks controlling idea lacks clarity lacks logical structure lacks conclusion	includes weak introduction displays controlling idea exhibits weak clarity exhibits weak logical structure includes weak conclusion	includes adequate introduction displays adequately developed controlling idea exhibits adequate clarity exhibits adequate logical structure includes adequate conclusion	includes strong introduction displays clear, well-developed controlling idea exhibits excellent clarity exhibits strong logical structure includes well-supported conclusion
Style	tone or voice is off-putting seems to have no audience in mind frequently chooses inappropriate words exhibits frequent inappropriate sentence structure uses no appropriate stylistic conventions	contains inconsistent tone or voice shows little audience awareness sometimes chooses inappropriate words exhibits occasional inappropriate sentence structure uses few appropriate stylistic conventions	contains occasional lapses in tone or voice shows audience awareness chooses appropriate words exhibits appropriate sentence structure uses appropriate stylistic conventions	maintains a consistent tone and voice shows consistent audience awareness exhibits skill in word choice exhibits sophisticated sentence structure skillfully uses appropriate stylistic conventions
Mechanics	lacks command of mechanical conventions: grammar, punctuation, or spelling errors present major distraction to readers	Demonstrates weak command of mechanical conventions: grammar, punctuation, or spelling errors are occasionally distracting to readers	demonstrates adequate command of mechanical conventions: grammar, punctuation, or spelling errors are minimally distracting to readers	demonstrates excellent command of mechanical conventions: grammar, punctuation, and spelling small errors do not distract readers

Appendix C

Item and Rater Means for Team Three (0-3 Scale)

Rubric Item Number	Team 3
1. Critical Thinking	1.7
2. Organization	2.1
3. Style	2.1
4. Mechanics	2.1
Rater 1 Mean	1.8
Rater 2 Mean	2.3
Grand Mean	2.0
Dependability (Phi Φ)	.30

G-Study Percent of Error Variance Associated with Facets of Measurement for Team Three

	Percent Variance
Source of Variance	Team 3
Persons	.12
Raters	.19
Items	.04
Persons x Raters	.24
Persons x Items	.07
Items x Raters	.02
Error (Persons x Raters x Items)	.32
Total	100

Phi Coefficient Estimates Based on Alternative Designs for Team Three

	Team 3		
	Raters	Items	Phi
Current Design	2	4	.30
Alternative Designs	2	3	.28
	2	4	.30
	2	5	.31
	3	3	.36
	3	4	.38
	3	5	.39
	4	3	.41
	4	4	.44
	4	5	.45

B. Collegial Review

Writing-Enhanced instructors were given the opportunity to submit 5-10 pieces of work generated by students in their classes, representing a range of writing ability. The faculty were then be invited to evaluate the writing samples qualitatively without using a rubric, discuss their findings, and compile a descriptive report detailing what was seen in each portfolio and what the portfolios exhibited overall. Thus, faculty members who submitted a portfolio of student work received a report detailing the traits the committee found in their portfolio, which they could compare to the list of traits the portfolios as a whole were found to exhibit. The goals of the Collegial Review are:

- Large-scale program would assess samples from Writing-Enhanced classes each semester
- Small-scale program would focus on writing generated by a particular type of Writing-enhanced course (JINS, WE in the major, WE in the LSP) and could be done on either a semesterly or yearly basis, either on Reading days spread throughout the academic year, or reading periods held in the summer or other interim periods
- Participation in the program (both the submission of student work to be evaluated and participation in the qualitative review process) would be voluntary and would be remunerated

This assessment is very similar to a ground-breaking assessment done by the Iowa Writing Project, administered by Barb Price.

C. Writing Across the University Committee

The Writing Across the University Committee was appointed in the 2004 fall semester. Members include faculty appointed by the VPAA, two students, and *ex officio* staff members: the Portfolio Project Director, the Director of Interdisciplinary Studies, the Assessment Specialist, the Director of the Center for Teaching and Learning, and the Director of the Writing Center who also chairs the committee.

The WAU Committee began meeting in earnest during the 2005 spring semester and, in addition to addressing a request by UGC for information regarding WE courses, it established its charge:

> The WAU Committee is a group of faculty, staff, and students dedicated to fostering writing on the Truman campus. As one of the three branches of writing assessment at Truman, the WAU Committee is responsible for promoting continuous improvement in the teaching and learning of writing. Specifically, the WAU Committee counsels Truman's faculty and staff on implementing the Analytical Assessment embedded in the LSP Portfolio and the Collegial Review. The WAU Committee collects, interprets, and communicates the data generated by both the Analytical Assessment and Collegial Review to faculty, students, and other interested stakeholders. The WAU Committee also serves as an advocacy group for faculty teaching or preparing to teach writing-enhanced courses. In order to support faculty, it gathers, analyzes, and distributes information about successful strategies for teaching writing-enhanced courses. By supporting writing teachers as well as advancing the assessment of writing, the WAU Committee endeavors to enhance the culture of writing at Truman which will lead to improved student writing. During the 2005 fall semester, the WAU Committee has been examining the Collegial Review in order to establish a viable qualitative assessment as proposed and approved by the UGC. This endeavor has led to the WAU Committee gathering information about the qualitative reviews sponsored by the Iowa Writing Project, interviewing participants from the 2003 Pilot Project conducted by Professor Barbara Price, and refining a Collegial Review model that will meet the objectives for qualitative assessment while taking into the consideration methods for funding.

The WAU Committee has formally collected no data and is still in the process of determining how to disseminate information in a formalized manner. Thus, although the Analytical Assessment is currently embedded as part of the Portfolio Project and has produced some data, the Collegial Review is still in the process of being implemented and the WAU Committee has yet to establish a method for formally collecting information from both assessments.