

MEMORANDUM

To: PSC Faculty

Fr: D. Elkins, Interim Chair 

Date: September 25, 2003

Re: Student Assessment Results, Spring 2003

Overview

The results of the first student assessment have been calculated. In brief, the reviewing faculty found that over three-quarters (76.3%) of the papers assessed either met or exceeded the established expectations. However, the faculty, paired as reviewers, had a low degree of inter-coder reliability ($r = .22$). The paired faculty reviewers agreed on only two-fifths (40.0%) of the paired reviews.

Specific Findings

On the attached pages, you will find tables depicting the results of the assessment. I have divided this section into two parts. The first part relates to the student assessment, and the second part relates to the inter-coder reliability issue.

Student Assessment. The findings of the assessment indicate that the reviewers had their most serious reservations about the papers in the Critical and Analytic Criteria. As illustrated in Table 1, well over a third of the assessed papers did not meet established expectations (37.5%). The components that the reviewers found most troublesome were related to the Hypothesis Component and the Conclusions Component.

Despite the relatively low scores on the Critical and Analytic Criteria, the reviewers found that the vast majority of the assessed papers either met or exceeded established expectations for the Research Criteria and the Articulate and Communication Criteria. The Research Criteria was a particularly noteworthy result with over two-fifths of the reviewers scoring the papers in this area as exceeding established expectations, the Sources Component and the Bibliography Component led this criteria.

Inter-Coder Reliability: This was a non-trivial problem with the assessment. As Table 2 illustrates, the reviewing faculty agreed on only 40.0% assessed papers' dimensions.

The area of most frequent agreement was in the Research Criteria. The reviewers agreed over half the time. By contrast, the Critical and Analytical Criteria and the Articulate and Communicate Criteria pose had low levels of agreement. The reviewers were more likely to disagree than agree in these two criteria. However, the nature of disagreement was unique to each criterion.

The Critical and Analytic Criteria had the more difficult and troubling forms of disagreements. There were more Major Disagreements, one reviewer scores a paper's dimension as Exceeds Expectations and the other scores the paper on this dimension as Does Not Meet Expectations, in this section than in any other section of the assessment. The reviewers were as likely to disagree in the Articulate and Communicate Criteria as they were in the Critical and Analytic. However, the disagreement was almost as likely to be whether the paper exceeded expectations instead of meeting expectations.

Proposed Suggestions

My suggestions are divided into the Substantive Outcome of Assessment and Inter-coder Reliability sections. In the first, I observe that most students write case studies and that the instruction that I use in the Data Analysis class has a quantitative assumption bias. In the next section, I suggest that we modify the instrument, based on the suggestion of a colleague, to minimize hopefully the inter-coder reliability problem.

Substantive Outcome of Assessment: This is the Department's first attempt at conducting student assessment with this instrument and this process and these results are obviously preliminary. My chief observation is that most of the students write case studies and not quantitative-based papers. Given that the reviewers of the papers indicate the greatest reservations in the Critical and Analytic Criteria, I surmise that this is a weakness that should be discussed.

I can only speak for the section of PSC 251 Data Analysis that I teach, but I do not spend much time at all instructing students on proper techniques of case study analysis. Indeed, I instruct students with a quantitative assumption bias. That is to say, I instruct students presuming that they are going to *use* quantitative methods. Still, there are at least two components to this usage: consumption and production. On the one hand, the instruction in quantitative methods is important to aid students as critical consumers of quantitative research.

On the other hand, as is evident in this round of assessment, the students produce papers that are fundamentally qualitative in orientation. There are obvious overlaps between the two forms (quantitative and qualitative) of methods in the development of research questions, hypothesis formation, use and import of theory. It is my sense that in my class my emphasis on the quantitative may implicitly bias students into believing that these critical areas of overlap are exclusive to quantitative research and perhaps not relevant to qualitative research.

To the extent that my colleagues that teach Data Analysis have this problem (to varying degrees) I think it is important that we insist and demonstrate the importance of the development of research questions, hypothesis formation, use and import of theory in qualitative research as well as with quantitative research.

Inter-coder Reliability: The inter-coder reliability must be increased. There are two potential remedies. The first is a minor modification to the existing instrument's coding structure. One faculty member suggested that we create middle categories between the Exceeds Expectations and Does Not Meet Expectations. For instance,

Exceeds Expectations		Meets Expectations		Does Not Meet Expectations
5	4	3	2	1

By modifying the instrument, this may eliminate the number of Minor Disagreements and increase the number of Agreements. Two reviewers may not feel compelled to a judgment of "either-or" and settle for a "sort-of" category when scoring a dimension on a paper. The dilemma is that this may increase the number and magnitude of Major Disagreements.

The potential remedy for Major Disagreements is neither easy nor simple. One choice is to re-work and further clarify the descriptions and instructions of each dimension in the Student Assessment. The transaction cost of getting agreement among the faculty on this issue is formidable. The other option is to provide training to the faculty reviewers regarding the use of the Student Assessment. This would be time consuming and as difficult as the previous option.

My suggestion is twofold. First, adopt the coding structure modification for the Fall 2003 assessment. Analyze the Fall 2003 results and determine if a significant problem remains with

inter-coder reliability. If inter-coder reliability remains a problem, identify those substantive areas that are creating the greatest problem, most likely the Critical and Analytical Criteria, and address those problems at that time.

Attachments:

Table 1: Frequency Distribution of Scores and Average of Scores for Assessment Papers, Spring 2003;

Table 2: Frequency Distribution of Agreements and Disagreements of Paired Assessment Reviewers, Spring 2003;

Table 3: Frequency Distribution of Reviewer Agreements;

Table 4: Student Assessment Spring 2003 Evaluator's Comments

Table 1: Frequency Distribution of Scores and Average of Scores for Assessment Papers, Spring 2003

Dimension	Frequency of Scores ¹			Average ²
	Exceeds Expectations	Meets Expectations	Does Not Meet Expectations	
<i>Critical and Analytical Criteria</i>				
Thesis Component	5	18	7	1.93
Hypothesis Component	4	12	14	1.67
Evidence Component	7	15	8	1.96
Conclusions Component	3	11	16 ³	1.53
<i>Criteria Subtotal</i>	<i>15.8%</i> 19	<i>46.7%</i> 56	<i>37.5%</i> 45	<i>1.78</i>
<i>Research Criteria</i>				
Sources Component	16	13	1	2.5
Citations Component	10	15	5	2.17
Bibliography Component	14	11	5	2.3
<i>Criteria Subtotal</i>	<i>44.4%</i> 40	<i>43.3%</i> 39	<i>12.2%</i> 11	<i>2.32</i>
<i>Articulate and Communicate Criteria</i>				
Organization Component	5	18	7	1.93
Paragraphs Component	6	19	5	2.03
Sentence Structure Component	7	17	6	2.03
Diction Component	8	16	6	2
Grammar Component	8	17	5	2
<i>Criteria Subtotal</i>	<i>22.7%</i> 34	<i>58.0%</i> 87	<i>19.3%</i> 29	<i>2.03</i>
TOTALS	25.8% 93	50.5% 182	23.6% 85	2.02

1 = The frequency of scores columns represent the rankings that each faculty member gave to a paper. In this portion of the analysis the scores are treated as discrete and not paired. That is to say, though each paper had two reviewers (paired reviewers), I recorded in these columns the score that each reviewer would have given on the various dimensions. For example, two colleagues reviewed Assessment Paper #2. If the first colleague scored the Thesis Component as "Meets Expectations" and the second colleague scored it as "Does Not Meet Expectations," those scores would be represented in two separate columns in the Frequency of Scores.

N = 360 {(12 Dimensions · 15 Papers) · 2 Reviewers}

2 = The arithmetic average was derived by establishing a mean for each dimension for each paper. I then created an average of these averages.

3 = One reviewer coded a paper as "0". I recoded this to "1" corresponding to the appropriate categories.

Table 2: Frequency Distribution of Agreements and Disagreements of Paired Assessment Reviewers, Spring 2003

Dimension	Agreements ¹	Disagreements ²		
		Minor High ³	Minor Low ⁴	Major
<i>Critical and Analytical Criteria</i>				
Thesis Component	6	4	4	1
Hypothesis Component	8	1	3	3
Evidence Component	3	6	3	3
Conclusions Component	5	0	7	3
<i>Criteria Subtotal</i>	<i>36.7%</i> 22	<i>18.3%</i> 11	<i>28.3%</i> 17	<i>16.7%</i> 10
<i>Research Criteria</i>				
Sources Component	8	6	1	0
Citations Component	7	5	2	1
Bibliography Component	9	5	0	1
<i>Criteria Subtotal</i>	<i>53.3%</i> 24	<i>35.6%</i> 16	<i>6.7%</i> 3	<i>4.4%</i> 2
<i>Articulate and Communicate Criteria</i>				
Organization Component	5	5	5	0
Paragraphs Component	4	6	5	0
Sentence Structure Component	6	3	6	0
Diction Component	5	6	4	0
Grammar Component	6	4	5	0
<i>Criteria Subtotal</i>	<i>34.6%</i> 26	<i>32.0%</i> 24	<i>33.3%</i> 25	<i>0</i>
TOTAL	40.0% 72	28.3% 51	25.0% 45	6.7% 12

1 = Agreement means that the paired reviewers agree on the paper's score for a discrete dimension.

2 = There are two types of disagreements: Minor and Major. A minor disagreement means that the paired reviewers differed by one point on the score a paper for a discrete dimension. A major disagreement means that the paired reviewers disagreed by two points, specifically one reviewer indicated the paper Did Not Meet Expectations and the other indicated that the paper Exceeded Expectations.

3 = A "Minor High" Disagreement indicates that one reviewer indicated that a paper Meets Expectations and the other reviewer indicated that the paper Exceeded Expectations.

4 = A "Minor Low" Disagreement indicates that one reviewer indicated that a paper Meets Expectations and the other reviewer indicated that the paper Did Not Meet Expectations.

Table 3: Frequency Distribution of Reviewer Agreements

Dimension	Agreements		
	Exceeds Expectations	Meets Expectations	Does Not Meet Expectations
Critical and Analytical Criteria			
Thesis Component	0	5	1
Hypothesis Component	0	4	4
Evidence Component	0	2	1
Conclusions Component	0	2	3
Research Criteria			
Sources Component	5	3	0
Citations Component	2	4	1
Bibliography Component	4	3	2
Articulate and Communicate Criteria			
Organization Component	0	4	1
Paragraphs Component	0	4	0
Sentence Structure Component	2	4	0
Diction Component	1	3	1
Grammar Component	2	4	0
TOTAL	22.2% 16	58.3% 42	19.4% 14

Table 4. Student Assessment Spring 2003 Evaluator's Comments

Dimension	Evaluator's Comments
Critical and Analytical Criteria	
<i>Thesis Component</i>	<ul style="list-style-type: none"> • No political science or political economy thesis – purely descriptive (2) • Too many of them (5) • No purpose apparent (11) • Application of a theory-based theory thesis or hypothesis is not made clear (13) • Must it be so trite? (20) • Thesis appears on pg. 2. That conflict in entire region is dominated by “economic approach” – vague and not operational (21).
<i>Hypothesis Component</i>	<ul style="list-style-type: none"> • No clear hypothesis (2) • The student's hypothesis struck me as a “virtual” hypothesis (4) • Never stated – mostly about Japan's effect (5) • (descriptive?) (11) • Paper comes down to: does case fit paradigm (13) • More the application of a theory than its explanation power (14) • There is no hypothesis. No systematic use of rational choice theory either (21) Meets test of theoretically appropriate...explanation of a significant political problem (23)
<i>Evidence Component</i>	<ul style="list-style-type: none"> • Evidence present but not widely sourced (4) • Somewhat difficult to assess this. Student has an unclear thesis but has evidence generally appropriate to demonstrate a potential thesis (12) • For what = in the end the contention is that symbolic politics explains the “ethnic war” (13) • Most of actual data is a chronicle account not consciously linked to the theory (14) • Needed to tie evidence to theory more (19)
<i>Conclusions Component</i>	<ul style="list-style-type: none"> • There is no evidence or conclusion relative to this contention (13) • Are asserted but far from demonstrated. Still a credible effort (14) • Odd conclusions section (17) • Over reach and not tied to rational choice theory contentions (21)

Research Criteria	
<i>Sources Component</i>	<ul style="list-style-type: none"> • Many cites but no political science sources (2) • I count 4-5 academic sources actually used (13) • Few scholarly sources (21)
<i>Citations Component</i>	<ul style="list-style-type: none"> • No year in embedded citation (4) • Does not include year in embedded citation form. Sometimes in sentence, sometimes not (7) I have some reservations about student's citations. For instance, cites Rabuskla and Shepsle's notion from a separate source (12)
<i>Bibliography Component</i>	<ul style="list-style-type: none"> • But largely unused (13) • Organization – limited scholarly sources (21)
Articulate and Communicate Criteria	
<i>Organization Component</i>	<ul style="list-style-type: none"> • Highly redundant – argument goes nowhere (2) • Some problems with all these, but better than many CSU students (4) • Long. Long historical section does not advance argument much (14). • Not clear how the sections contribute to the economic argument (21).
<i>Paragraphs Component</i>	<ul style="list-style-type: none"> • Some problems with all these, but better than many CSU students (4) • Endless repetition (20)
<i>Sentence Structure Component</i>	<ul style="list-style-type: none"> • Student's writing is making errors consistent with non-native speaker (4) • Some problems with all these, but better than many CSU students (4)
<i>Diction Component</i>	<ul style="list-style-type: none"> • Student's writing is making errors consistent with non-native speaker (4) • Some problems with all these, but better than many CSU students (4) • Generally OK with some minor problems (21)
<i>Grammar Component</i>	<ul style="list-style-type: none"> • Student's writing is making errors consistent with non-native speaker (4) • Some problems with all these, but better than many CSU students (4) • Some mistakes (21)

One evaluator submitted the following:

"Overall comment:

With the exception of paper #7, they weren't particularly good. They were mostly descriptive, not using theoretical material very effectively. On the other hand, there was some description of theory, they had obviously done a fair amount of work, and they presented their stories in a fairly coherent manner – certainly better than many CSU students.

So, even though I was unsatisfied on some level, I gave them mostly "2"s."