Senior Thesis Assessment Project
May 2006

Organizers
Steven Weisler, Dean of Academic Development and Professor of Linguistics, Hampshire College
Carol Trosset, Director of Institutional Research, Hampshire College

Participants
Amy Clore, Assistant Professor of Biology, New College
Mimi Czarnik, English, Alverno College
Charlene D'Avanzo, Dean of Natural Science and Professor of Ecology, Hampshire College
Mike Ford, Dean of the College and Associate Professor of Social Science, Hampshire College
Heidi Harley, Associate Professor of Psychology, New College
Carolyn Haynes, Professor of Interdisciplinary Studies, Miami University
Susan Marks, Assistant Professor of Judaic Studies, New College
Karen Spear, Director of CIEL
Chris Wolfe, Professor of Interdisciplinary Studies, Miami University

Rubrics

The original version of the rubric was developed before the workshop began, by Steve Weisler and Carol Trosset. This version was used for the norming exercise the first evening.

Original Version	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Rationale	no clear rationale for the project	weak rationale	provides an appropriate rationale	persuasive and creative rationale
Dealing with Complexity in Framing Topic	frames complex questions as simple ones	acknowledges some complexity but defines the topic in a way that simplifies it	frames the question by addressing some - but not all - of the relevant dimensions of the topic's complexity	frames the topic with a full appreciation of its complexity
Approach	not clear what was done	clear what was done but not explained/justified	clearly described and explained, but may be missing some approach that should be used, or imperfectly executed	clearly described and justified, well-chosen and appropriate, and well-executed
Scholarly Context	author does not demonstrate awareness of the relevant scholarly literature	author demonstrates some awareness of the literature	author demonstrates broad awareness and situates own work within the literature	author does these things and makes a contribution to the field, or identifies a new direction for investigation

Position	does not take a clear position or draw a clear conclusion	describes a position that is already in the literature	extends or critiques a position that is already in the literature	develops a clear position of his/her own, draws a significant conclusion
Argument	no argument, perhaps a simple assertion	a weak or invalid argument	a valid argument, well supported with evidence	an argument that is both well supported and genuinely tested against conflicting explanations
Use of Data/Evidence	draws on little or no data or evidence	skims the surface, leaves much available data/evidence unused, or used selectively to support author's position	attempts to deal with full range of evidence but does not offer a fully satisfactory explanation or does not consider counter-evidence	fully exploits the richness of the data/evidence/ideas
Seeing Patterns and Connections	treats related issues, ideas, or data as if they were unrelated	draws weak or simplistic connections between related data or ideas	brings together related data or ideas in appropriate ways	develops insightful connections and patterns that require intellectual creativity
Writing				
Grammar and Spelling	many errors	some significant errors	a few minor errors	no errors
Organization	seriously flawed	would benefit from re-organization	good, easy to follow	outstanding, including strong introduction and conclusion and coherent transitions
Clarity, Style, Readability	poor	gets in the way of reading for content	good, easy to follow and read for content	exceptional, including elegant style, transparent argument structure
Size of Project (treat as continuum)	equivalent to work for one course			full-time work for two semesters

At several points during the workshop, the group discussed the rubric and made revisions. Revisions were motivated by several goals: to find language that seemed appropriate across disciplines, to provide sufficient detail to assist readers in placing marginal works, and to raise the standard represented by Level 2 so that more weak student work would fall into Level 1, and that Level 3 would represent quite high quality work.

Revised Version	Level 1	Level 2	Level 3	Level 4
Rationale	no clear rationale or a weak rationale for the project	some rationale presented, begins to motivate the work	provides and discusses a suitable rationale	persuasive and creative rationale
Dealing with Complexity in Framing Topic	frames complex questions as simple ones	invests question with some complexity, may over-simplify or over-extend	reasonable balance between focus and complexity	frames the topic with a full appreciation of its complexity while retaining appropriate focus
Approach	not clear what was done or why, or an inappropriate method	approach is generally appropriate and properly executed	clearly described and justified, well-chosen and appropriate, and well-executed	creative and sophisticated methods
Scholarly Context	author does not demonstrate awareness of the scholarly literature, may over-rely on too few sources	author demonstrates a reasonable awareness of the literature	author demonstrates broad awareness and situates own work within the literature	author does these things and makes a contribution to the field, or identifies a new direction for investigation
Position	does not take a clear or defensible position or draw a clear conclusion	clearly describes, or begins to support/test/extend/ critique a position that is already in the literature	thoroughly and effectively supports, tests, extends, or critiques a position that is already in the literature	develops a clear and defensible position of his/her own, draws a significant conclusion
Argument	weak, invalid, or no argument, perhaps a simple assertion	some arguments valid and well supported, some not	main arguments valid, systematic, and well supported	arguments both well supported and genuinely compared to conflicting explanations
Use of Data/Evidence	draws on little or no evidence, mostly relies on assertions or opinions, or evidence not clearly presented	some appropriate use of evidence but uneven	feasible evidence appropriately selected and not over-interpreted	fully exploits the richness of the data/evidence/ideas, and is sufficiently persuasive
Insight, Seeing Patterns and Connections	treats related ideas or data as unrelated, or draws weak or simplistic connections	begins to establish connections and perceive implications of the material	brings together related data or ideas in productive ways, thoroughly discusses implications of material	develops insightful connections and patterns that require intellectual creativity

Writing Mechanics				
Usage, Grammar and Spelling	significantly impairs readability	frequent or serious errors	some minor errors	virtually no errors
Organization	needs significant reorganization	structure is of inconsistent quality, may have choppy transitions and/or redundancies or disconnections	structure supports the argument, clearly ordered sections fit together well	structure enhances the argument, strong sections and seamless flow
Clarity, Style, Readability (as appropriate to disciplines)	gets in the way of reading for content	beginning to be comfortable with appropriate conventions, style is inconsistent or uneven	effective prose style, follows relevant scholarly conventions, emergence of voice	mastery of the genre, including elegant style, established voice
Size of Project (treat as continuum)	equivalent to work for one course			full-time work for two semesters (equivalent to eight courses)

The Process

Each institution brought theses in the academic areas covered by the attending faculty members. Piles of theses appropriate to each participant were assembled. No one read any theses from their own institution (except Trosset, who had never before read a Hampshire thesis). It was not possible to disguise the institutional identity of the theses while they were being read.

The first evening, we all read the same thesis from Hampshire, applied the original rubric to it, and discussed our ratings on each parameter. There was a very high level of agreement among all readers. Discussing the areas where we disagreed helped us to begin the process of editing and improving the rubric.

Over the next day and a half, the rubric was applied 90 times to 81 different theses. (9 theses were read twice, to provide a beginning indicator of inter-rater reliability.) Table 1 shows the number of theses falling into the following categories by subject and institution. The first number in the cell gives the number of total readings in that category, while the number in parentheses gives the number of different theses read in that category.

Table 1. Theses by Subject Area and Institution.

	Cognitive Sciences	Humanities	Natural Sciences	Social Sciences
Hampshire	11 (10)	17 (15)	11 (9)	1
New	8 (7)	12 (11)	7	1
Miami	2	4	2 (1)	2
Alverno	6	1	0	6 (4)

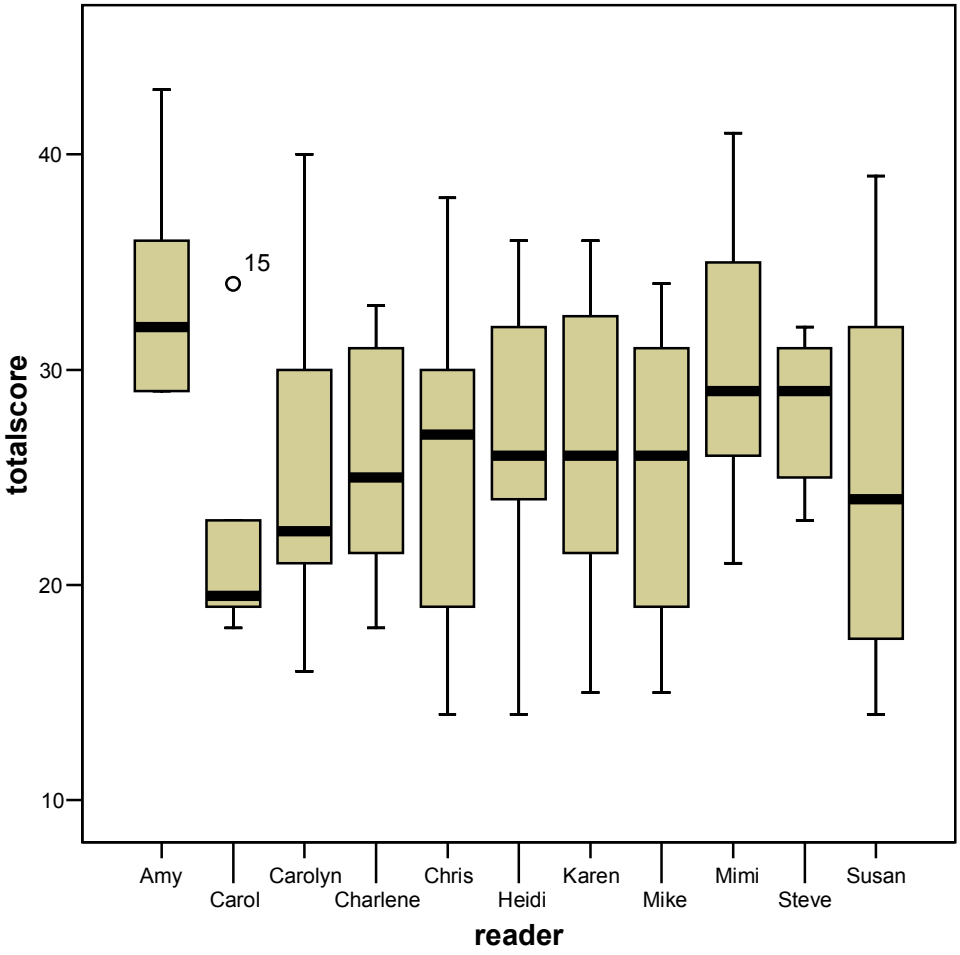
Note that the theses read were not distributed evenly across institutions or across subject areas. Most theses came from Hampshire and New Colleges, while most social science theses came from Alverno College, etc. These imbalances will necessarily influence the way the results are analyzed below.

Inter-rater Reliability

The main goal of this workshop was to test the rubric, to see how effectively it could be used across disciplines and by different people. We appear to have been quite successful in these respects.

Treating second readings of a single thesis as separate data points, Figure 1 below shows the ranges of scores assigned to the various theses by different readers. Note that although Amy gave the highest average scores and Carol the lowest, everyone's ranges overlapped to a considerable degree (even though they were not reading the same subject areas and may not have been reading theses of comparable quality). Differences between the readers were not statistically significant.

Figure 1 – Total Scores by Different Readers.



Nine theses were read by two different people. Three of these pairs were read using different versions of the rubric; the others used only the revised version. In the pairs using different versions, the newer version always yielded the lower score. However, inter-rater reliability did not improve when only the revised version was used.

Table 2 shows the range of variation for the nine these that were read twice each.

	College of thesis	Subject Area	# of 11 rubric dimensions with different scores	Difference in total score (points)	Difference in # course equivalents
1	New	Humanities	9	9	2 – 5
2	Hampshire	Natural science	4	2	1 – 8
3	Alverno	Social science	7	5	0.5 – 1
4	Miami	Natural science	6	3	2.5 – 3
5	Hampshire	Cognitive science	6	4	2 = 2
6	Hampshire	Humanities	4	0	0.5 – 5
7	New	Cognitive science	2	2	4 – 6
8	Hampshire	Humanities	7	6	1.5 – 6
9	Hampshire	Natural science	3	3	5 – 8

With such a tiny sample it doesn't mean very much, but in general, our inter-rater reliability with respect to quality seems fairly good.

- In four cases more than half of the dimensions of the rubric received the same score from both readers. In five cases more than half the dimensions were rated differently.
- There was no difference between the dimensions in how consistent readers were in using them. Each of the eleven dimensions was rated differently 3-5 times out of the nine theses read twice.
- Of the 48 times that one thesis received two different scores on one dimension of the rubric, the two scores differed by more than one point only 4 times (8% of the time).
- Total score was calculated by assigning points for each component equivalent to the level selected, and summing them. The maximum possible total score was 44 and the minimum possible was 11. Only three of these nine theses received total scores differing by more than four points.
- The level of inter-rater reliability does not appear to vary with subject area.

On the other hand, our estimates of the scope of effort required by the various projects was not reliable in most cases. Only three of the nine were estimated at the same level of time and effort by the two readers. Five theses received estimates that differed by at least three courses' worth of work.

In the rest of this report, the two readings of these nine theses will be treated as if they referred to different theses.

Overall Results

Of the 90 theses read, 51 (57%) were evaluated using the first version of the rubric and 39 (43%) were evaluated using the second version.

As mentioned above, when both versions were used by different readers to evaluate the same thesis, the second version always resulted in the lower score. However, when considering all theses read, the average score assigned using the second version was slightly higher.

Table 3. Scores by Rubric Version.

Version	N	Minimum	Maximum	Mean	Standard Deviation
First	51	14	40	25.8	7.2
Second	39	14	43	28.1	7.1

Given that we have no independent measure of the quality of the theses read using the two versions, we should not infer anything important from this and for the rest of this report the analysis will ignore the different between versions.

Total scores assigned ranged from 14 to 43. The top 15% received scores between 35 and 43, while the bottom 15% received scores between 14 and 19.

Table 4 shows the range of scores assigned on each dimension of the rubric.

	Level 1	Level 2	Level 3	Level 4
Rationale	14	34	35	7
Complexity	14	32	33	11
Approach	18	39	28	5
Context	15	43	25	7
Position	13	42	28	7
Argument	19	35	27	9
Evidence	20	31	30	9
Insight	11	36	30	13
Usage	2	16	57	15
Organization	13	31	40	6
Style	5	26	54	6

Table 5 repeats the previous one but gives percents instead of counts.

	Level 1	Level 2	Level 3	Level 4
Rationale	15%	38%	39%	8%
Complexity	15%	37%	36%	12%
Approach	20%	43%	31%	6%
Context	16%	48%	28%	8%
Position	14%	47%	31%	8%
Argument	21%	39%	30%	10%
Evidence	22%	35%	33%	10%
Insight	12%	40%	33%	15%
Usage	2%	18%	63%	17%
Organization	14%	34%	45%	7%
Style	5.5%	29%	60%	5.5%
Average	14%	37%	39%	10%

The remaining statistics in this section of the report are an attempt to investigate to what extent the eleven dimensions of the rubric measure different things as opposed to the same thing, and to discover which dimensions best predict the overall quality of the thesis.

A factor analysis of the 11 dimensions yielded two factors. The first was made up of all the dimensions except usage/grammar/spelling, which dominated the second factor.

Table 6 shows the degree to which each rubric dimension correlated with the total scores. All these correlations were significant at the 0.01 level. However, we can see that approach, argument, insight, and complexity correlate most closely with the total score.

	Correlation with Total Score
Rationale	0.71
Complexity	0.85
Approach	0.87
Context	0.83
Position	0.81
Argument	0.87
Evidence	0.80
Insight	0.86
Usage	0.46
Organization	0.77
Style	0.74

Table 7 shows which pairs of components correlate most closely with each other.

Complexity	Approach	0.78
Argument	Approach	0.77
Position	Argument	0.74
Scholarly context	Insight	0.73
Complexity	Insight	0.73
Evidence	Argument	0.72

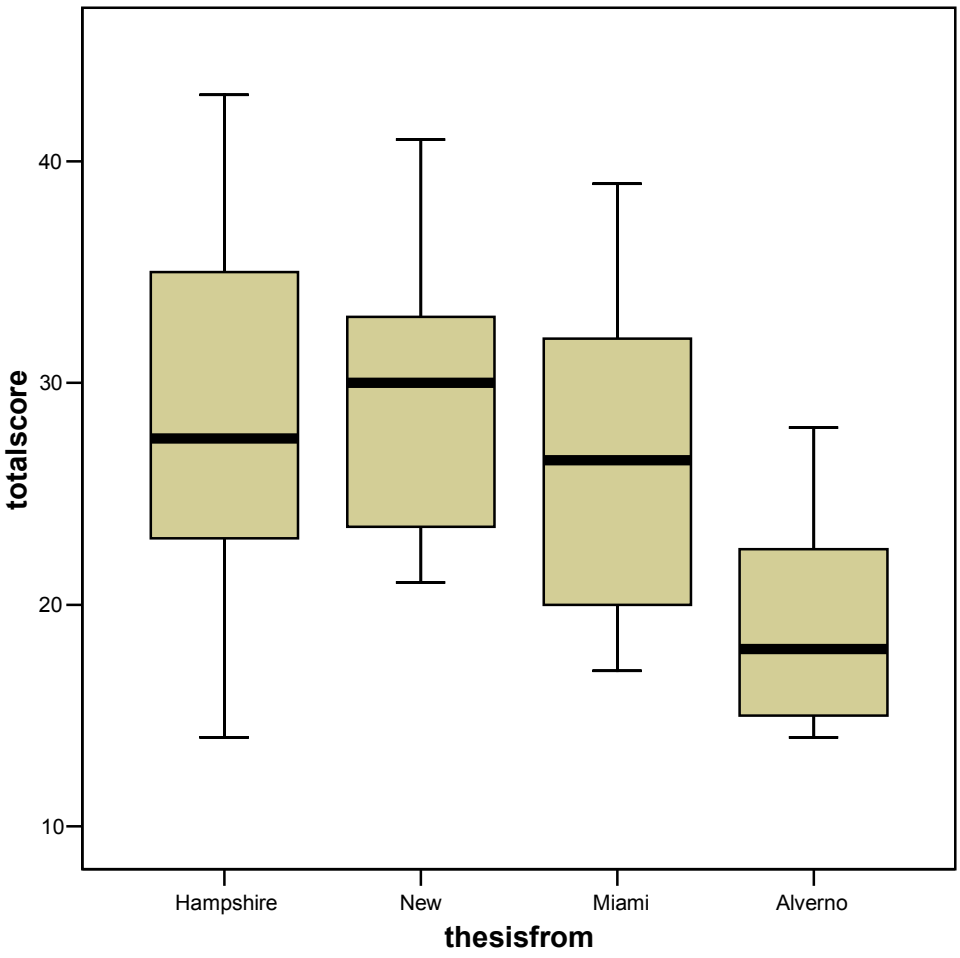
If there were more distinct factors or a near-perfect correlation between certain components, that would suggest that we could simplify the rubric while obtaining approximately the same results. For example, both “approach” and “argument” correlate strongly with each other and with the total score. However, given that the rubric also has the teaching function of reminding students, advisors, and readers of the important dimensions of a research paper, these correlations are not so strong as to justify merging these two dimensions into one.

Results by Institution

Table 8 shows how many theses were read from each institution and gives the total score averages. The difference between the Alverno average and the others is statistically significant.

	N	Average Total Score
Hampshire	40	27.8
New	28	28.9
Miami	10	26.7
Alverno	12	18.7

Figure 2 shows the distribution of scores assigned to theses from each institution.



In order to investigate any differences between the theses from Hampshire, New College, and Miami, it was necessary to remove the Alverno theses from the remaining analysis. Doing so revealed that there was no statistically significant difference between the total scores at those three institutions. There was only one statistically significant difference between these three institutions with respect to the eleven individual dimensions of the rubric: the Miami theses received a lower average score with respect to the use of evidence.

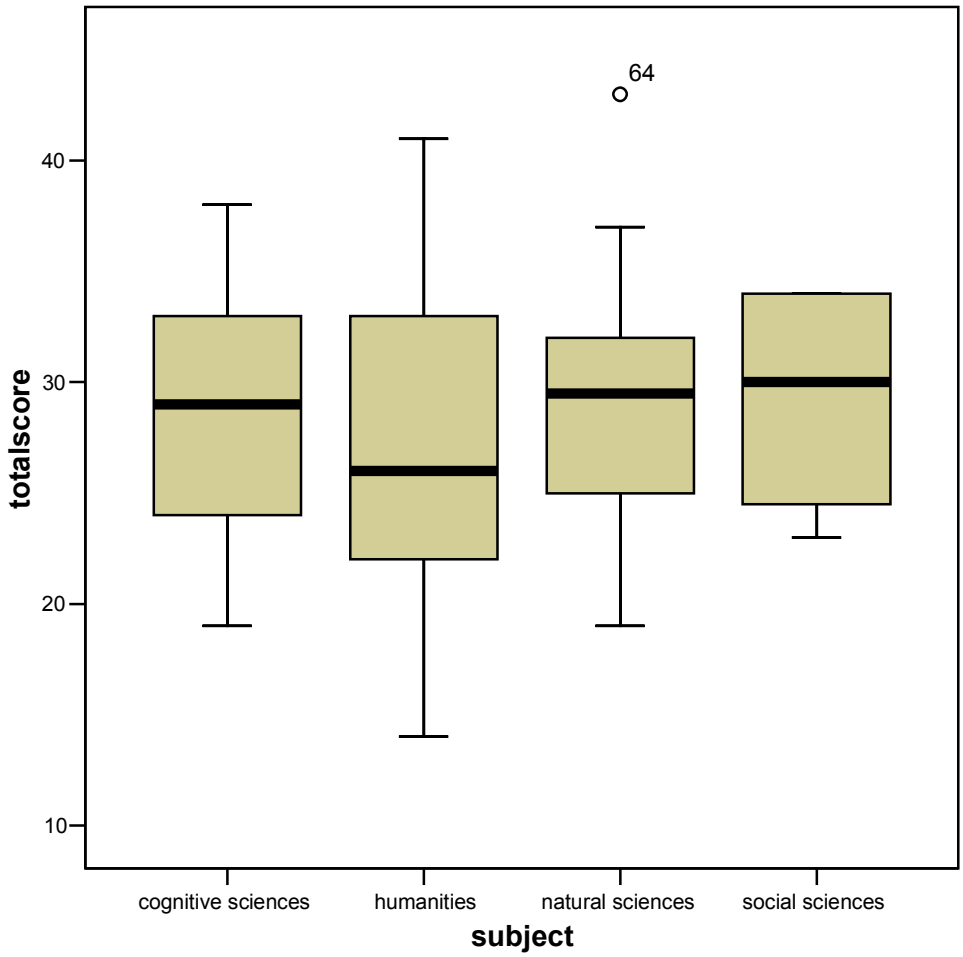
Results by Subject Area

Because the Alverno scores were significantly lower overall and because Alverno brought the majority of social science theses and no natural science theses, it was necessary to remove Alverno from the analysis of results by subject area. Having done so, there were no statistically significant differences between subjects.

Table 9. Scores by Subject Area (no Alverno)

Subject	N	Minimum	Maximum	Mean	Standard Deviation
Cognitive Science	21	19	38	28.2	6.0
Humanities	33	14	41	27.0	7.7
Natural Science	19	19	43	29.3	6.1
Social Science	4	23	34	29.3	5.6

Figure 3 shows the very high degree of similarity in the range of scores given in each subject area.



Conclusion

Although at the onset of this project we were not at all sure that it was possible to develop a scoring rubric that would work independent of topic and across the curriculum, we think we did it! Our revised rubric appears to work well for all research-based disciplines (although something different will be needed to evaluate creative and autobiographical writing, along with other creative and performing arts).

Based on this experiment, we believe that something very close to this rubric could be an effective tool for both students and advisors. Some of the language may still need adjusting, in particular for the components referring to writing style, clarity, and "voice."

What are the next steps? Some of us intend to try using this rubric at our own institutions with a wider range of these and of faculty readers. This could have at least two positive effects: further editing of the rubric to make it more widely useful across disciplinary approaches, and encouraging discussions across faculty members and disciplines about what qualities a high-quality thesis should have. In this final connection, our uneven estimates of the course "credit" equivalents of the various projects deserve further examination.