

TO: Larry Schuetz  
 FROM: The Assessment Advisory Committee  
 RE: World of Ideas (WOI) Assessment Report  
 DATE: August 17, 2001

As a continuation of previous evaluations of term papers from WOI, a sample of 24 term papers was gathered in the fall, 2000. In addition, in the spring of 2001, 49 short (50-250 words), extemporaneous papers were gathered in one section of English 101 and one section of WOI. There were three primary objectives for gathering these papers: 1) To compare the quality of writing between the two groups, 2) To see if the rubric could be used successfully with short papers, and 3) To explore additional strategies for gathering samples of student writing. A summary of the findings, compiled by Steve Friedman for the Assessment Advisory Committee, follow. Mark Lencho then describes the process used for scoring the latest papers.

Three criteria—thinking, voice, and literacy—make-up the rubric. Each was scored on a six-point scale—6=outstanding, 5=strong, 4=adequate, 3=limited, 2=seriously flawed, and 1=fundamentally deficient. The three raters read each term paper and scored each criterion on the six-point scale. The results for last two years are presented below:

	<b>1999</b>	<b>2000</b>
<i>Averages for thinking</i>	Rater#1=3.87	Rater#1=3.25
	Rater#2=3.36	Rater#2=3.92
	Rater#3=3.13	Rater#3=3.54
	<b>Overall average=3.45</b>	<b>Overall average=3.57</b>
<i>Averages for voice</i>	Rater#1=3.97	Rater#1=3.46
	Rater#2=3.47	Rater#2=4.17
	Rater#3=3.18	Rater#3=3.29
	<b>Overall average=3.54</b>	<b>Overall average=3.64</b>
<i>Averages for literacy</i>	Rater#1=3.95	Rater#1=3.71
	Rater#2=3.16	Rater#2=3.67
	Rater#3=2.82	Rater#3=3.25
	<b>Overall average=3.31</b>	<b>Overall average=3.54</b>

Based on the definitions supplied with the rubric, the averages for 1999 fell between the “limited” and “adequate” descriptors, and there seems to be little difference across criteria.

The overall average scores across the three criteria were uniformly higher for the papers gathered in 2000. The largest gain occurred in literacy. The standard deviations across criteria and raters averaged about 1.00 and ranged between .90 and 1.19. There were some differences from year to year within and across raters. For example, like 1999, Rater#3 generally assigned the lowest scores across criteria. Rater#1’s scores tended to decrease from 1999 to 2000 while Rater#2’s increased over the same period. Like the findings in 1999, the raters are tending to agree with each other’s scoring of the papers. The correlation coefficients between the students’ composite scores (total of thinking, voice, and literacy) for pairs of raters range between .51-.68. Up to this point, averages for each criterion have been the focus of the analyses. However, when determining inter-rater agreement, total scores for each student across criteria were used. Like 1999, the results fell between “limited” and “adequate” with little variation across criteria.

The short, extemporaneous papers were gathered in a section of English 101 at the beginning of the spring term, 2001. The short papers for the WOI section were gathered well into the semester. When the short papers were given to the raters, they had been assigned a three-digit, random number, and the 101 and WOI papers were mixed. There were no

identifying marks on the papers except the numbers. The results for each group are presented below:

	<b>English 101</b>	<b>WOI</b>
<i>Averages for thinking</i>	Rater#1=2.90	Rater#1=3.50
	Rater#2=3.05	Rater#2=3.75
	Rater#3=2.57	Rater#3=3.21
	<b>Overall average=2.84</b>	<b>Overall average=3.49</b>
<i>Averages for voice</i>	Rater#1=3.05	Rater#1=3.61
	Rater#2=3.00	Rater#2=3.79
	Rater#3=2.29	Rater#3=3.07
	<b>Overall average=2.78</b>	<b>Overall average=3.49</b>
<i>Averages for literacy</i>	Rater#1=3.05	Rater#1=3.36
	Rater#2=2.24	Rater#2=3.00
	Rater#3=2.14	Rater#3=3.21
	<b>Overall average=2.48</b>	<b>Overall average=3.19</b>

The scores for the short papers for English 101, across criteria, were between the “seriously flawed” and “limited” descriptors. On the other hand, the scores for WOI papers fell between “limited” and “adequate”, and, with the exception of the overall average for literacy, were similar to the average scores for the term papers. For the English 101 short papers, the relationship between the students’ total scores (thinking, voice, and literacy combined and averaged across the three raters for each student) and final course grades (as measured by a correlation coefficient) was .45—a moderate, positive relationship. For the WOI short papers, the relationship between the same two variables was -.09. Here, the slightly negative correlation can be explained, in part, by the fact that the range of grades assigned in the course was quite small—i.e. “A”s and “B”s.

The correlation coefficients between the composite scores (total of thinking, voice, and literacy) for pairs of raters ranged between .60-.78. Overall, the correlation coefficients were higher when the rubric was used with the short papers compared to the term papers which can be explained, in part, by the fact that the short papers were written on the same topic while the topics for the term papers varied by section of WOI. Additionally, six of the WOI short papers were selected based on how well they represented the range of scores for the entire group. The term papers written by these same six students were then scored by the three raters. The relationship between the average total scores (thinking, voice, and literacy combined and averaged across the three raters for each student) for the term and short papers (as measured by a correlation coefficient) was -.58. A partial explanation is that the student who had the highest score on the short paper had the lowest score on the long term paper. Further, the sample size is only six.

New data gathered in 2000 added to what we had learned previously about our students’ writing skills. There was some consistency in the scores across traits from 1999 to 2000, and, overall, the performance was in the center of the scoring continuum. The raters generally agreed with each other’s scoring for both the term and short papers. The short papers appeared to measure writing skills that differ from the skills tapped by the term papers, and the WOI students exhibited stronger skills on the short papers compared to the 101 students. Ideally, the evaluative information in this report can focus our attention on several important issues. Hopefully, it informs our decisions as we strive to continually improve the quality of student writing—a worthy goal for our General Education program. Your feedback to the Assessment

Advisory Committee is encouraged. Steve Friedman can be reached at ext. 1970 or in 2048 Roseman.

### MEMO

TO: Steve Friedman  
Associate Dean, LEARN Center

FROM: Mark Lencho  
Associate Professor, Languages and Literatures  
Chair, Ad Hoc General Education Writing Assessment Team

RE: Spring Writing Assessments

DATE: July 27, 2001

cc: Emily Hipchen, Michael Longrie  
Members, Ad Hoc General Education Writing Assessment Team

#### ◆ **Background**

During the spring and summer 2001 terms, assessment of writing in the UW-Whitewater General Education program completed its third round. In the spring 1999 semester, 44 term papers were gathered from World of Ideas, the General Education capstone course, 15 of which papers were evaluated by the assessment team of Assistant Professor Emily Hipchen, Associate Professor Michael Longrie and myself, all from the Department of Languages and Literatures. As a result of this assessment, we developed a rubric for evaluating the “thinking,” “voice,” and “literacy” of student writing upon completion of the General Education program at UW-Whitewater. Then, using a refined version of the rubric, we assessed a second set of term papers collected in World of Ideas courses during the fall 1999 semester.

The assessments summarized in this report begin with a third set of papers solicited from World of Ideas instructors, reflecting student work submitted in the fall 2000 semester. The rubric for the current assessment remains the same as that used during the previous assessment; therefore, it is the first to be “field-tested.” Since the initiation of the General Education writing assessment two years ago, the team of readers has remained the same.

#### ◆ **Ongoing Collection Procedures**

Consistent with the 1999 assessment procedures, at the close of the fall 2000 semester, Larry Schuetz, Assistant Dean of the College of Letters and Sciences, organized the collection of essays from a cross-section of World of Ideas courses. All World of Ideas instructors for the fall 2000 semester were requested to submit a small set of student term papers, randomly chosen so that they might reflect the range of abilities in each class. As a result of this collection, 24 papers were submitted for assessment, attributions of writer, teacher, and nature of assignment having been excised.

#### ◆ **The Scoring Procedure**

As in the previous assessment, I copied and distributed sets of the papers to each member of the assessment team, along with rubrics consisting of instructions to score each paper with regard to the established three primary traits and six-point scale (see Appendix One for a copy of the rubric). Readers could also supplement scores with justifying commentary as necessary. After reading and scoring the set of papers independently, the three scorers of the assessment team met to discuss the papers and their scoring.

#### ◆ **Some Innovations in the Scoring Methods**

There is a school of thought that considers the selection of any method of analysis to be essentially an aesthetic decision. A subject such as writing can be carved up many ways, but one

wants to see the analytic knife wielded incisively and elegantly. There is a natural elegance to symmetrical, balanced categories, and in pursuit of this aesthetic point, I, in conjunction with the other two raters, have endeavored to make the three primary traits equally substantive and equally important. The implication for our scoring procedure is that the score for each primary trait should have equivalent weight. It follows that overall scores represent the simple average of the three scores given for the primary traits, a procedure rigorously followed for the first time in this current assessment.

Additionally, we decided to enhance the consistency of scoring by rigidly holding all raters to the six-point scale of rankings. For the first time scores are simple, whole number values, unburdened with further annotations in the form of pluses, minuses, or fractions. In short, for the first time attention was paid not to subvert our six-point scale with a scoring practice that layered on an indeterminate number of sub-classifications. However, since the overall score represents the simple average of the three primary trait scores, it will usually contain a fractional part.

Finally, for the first time selected papers were re-scored to determine to what extent the original scores would be duplicated factoring in the passage of time and the discussion of the writing samples with the other raters. Two months after the original scoring, two papers were re-circulated to be re-scored. The papers were selected because on at least one primary trait there were particularly marked discrepancies in the scores assigned by each of the three readers. The small adjustments which were made in the second attempt at scoring these papers brought the scores of all three raters more in line. The preliminary indication is that though raters remain somewhat consistent in their scoring, the opportunity to re-score samples does have a regularizing effect on the set of scores. Appendix Two provides a comprehensive chart of the scoring data.

#### ◆ **Agenda Established by the Scoring**

It seems best to defer comments relying on numerical manipulations until the completion of the forthcoming statistical analysis by Steve Friedman. Some important questions to be addressed at that time include the following: a) Are papers trending the same, better or worse from semester to semester, insofar as can be ascertained by inspecting assessment team ratings (as evidenced by a comparison of average overall scores spanning the three assessments)? b) Are there differences in the overall evaluation of “thinking,” “voice,” and “literacy” as we compare the current assessment with the preceding one (as evidenced by a comparison of average scores of fall 1999 and fall 2000 papers on each primary trait)? c) Are scorers consistent as a group (as evidenced by strong inter-rater agreement coefficients)? Or do scorers diverge from each other in quirky or unpredictable directions? d) Do the primary traits as they have been defined represent discrete and independent analytical components of writing (as evidenced by significant standard deviations for each scorer on each primary trait)? Or is it the case, for example, that a high score on “thinking” on a given paper will automatically entail a high score in “voice” and “literacy” on the same paper for some or all scorers? e) Does the six point scoring scale adequately tease out qualitative differences among the papers (as evidenced by respectable standard deviations within the set of scores for each primary trait and within the set of overall scores)? Or is there clumping around a certain segment of the scoring scale, so that a part of the discriminating function of the scale is not used and the papers are not optimally differentiated?

#### ◆ **Findings**

Holding in abeyance a discussion of numerical data until the completion of a formal statistical analysis, I would like to focus here on some substantive, non-quantitative issues that

arose during our discussions. As we have progressed through each round of assessment, the primary traits of “thinking” “voice” and “literacy” upon which we base our evaluations of student writing have become increasingly elaborated and refined, even as the written description of each trait has remained the same over the last two assessments. In the latest writing samples, there were several writers who were unable to reference source material in their research paper successfully in a way that allowed raters to distinguish their own contributions from their sources. All of us detected this liability, but we differed as to whether the problem should be categorized under “literacy” (little “grasp of writing conventions,” i.e., weak footnoting skills) or under “voice” (little ability “to show ownership,” i.e., to establish one’s own distinctive point of view). The language in the rubric admits to different styles of interpretation. We could have made a decision by fiat to go one way or the other, for the sake of our scoring, but we ultimately refrained from adjudicating this issue and others like it.

Under the circumstances of a large-scale, high-stakes assessment, like that conducted by Educational Testing Service to evaluate essays written by students seeking advanced college placement, training sessions are aimed to enforce assessment by consensus. But to obtain consensus, there must be an interpretive method that stands for the standard, toward which divergent graders can be made to adjust. Typically, in the case of an ETS AP reading, standard scores can be easily ascertained by the simple determination of a majority opinion in a room of 300 or so raters. Training can then focus on methods used to anticipate the majority opinion of the interpretive community of raters. Among an assessment team of three experienced scorers who disagree on the application of the traits in the rubric, there is not a similarly non-arbitrary authoritative source for establishing the standard by which all three should abide. We could *arbitrarily* decide, for example, that deficient attributions of source material in a research paper provide evidence for deficient “voice,” and should not impact our score for “literacy,” but one is naturally reluctant to permit too much arbitrariness in the development of the analytical instrument. For the sake of streamlining scores and increasing inter-rater agreement, we increase the artificiality of the assessment process, making the analytical labels less intuitively clear, therefore less comprehensible to the audience of the assessment report and more difficult for future scorers to master. If only a few arbitrary decisions could secure a transparent assessment process, then there would be a powerful incentive to intervene and reorient a rater even in such a small group as we have for our current assessment. However, it may turn out that in the search for inviolate demarcations for our (“domain-specific” but not “task-specific”) primary traits, a settled procedure could turn out to be indefinitely elusive. The primary traits as we have defined them do not provide a comprehensive analysis of writing, nor are they meant to. As we go through successive assessments, it becomes more and more clear that interpretive strategies we use to transform our reactions to the student writing into a primary trait rating are extremely nuance. Building up the descriptors of each primary trait in order to account for every interpretive point of departure may turn out to propel us into a never-ending process of refinement.

So far, we raters of World of Ideas papers have refrained from coercing each other’s scoring toward an arbitrary standard, though we occasionally make changes in our scoring as we compare our results with the others in the group. So at the cost of compromising somewhat inter-rater agreement coefficients, our rater responses give a truer picture of inherent variability of skilled, informed responses to something as subjective as the art and skill of writing.

Rather than looking at ways to remediate wayward raters, the discussion of the assessment team has paid more attention of the dynamics of rubric application. In principle, there could be

two kinds of problems in applying the rubric: overlap and omission. Overlapping primary traits appear to be the reason raters handled differently the problem of a writer's poor control over resource material, as discussed above. Overlap perhaps also accounts for divergence of rater opinion about how to deal with irrelevant writing. Language in the rubric under "voice" seems to be pertinent, as writers who do not make it "clear why they are writing and where they are coming from" may be charged with irrelevance. Likewise, language under "thinking" invokes the ability to "delimit knowledge," the violation of which might also lead to irrelevance. Rather than revising the description of either primary trait, it might be more appropriate to concede that irrelevance necessarily implicates both "thinking" and "voice." A third example of overlap we discussed involved our reaction to the sense of a writer's "honesty," conveyed by the writer's willingness to anticipate and address intractable cases given his or her perspective. We found a concessive element to some of this writing, which led us to reward the writer's "thinking," but we also valued the uncontrived "voice" of such a writer.

As for a characteristic of writing that seems not to be referenced in the rubric, we were all struck by the evident obedience shown by a group of writers, presumably to instructions in an assignment which scripted each segment of the paper. These papers, one just like the next, left a very poor impression on the raters, despite our efforts to try to find a way to recognize something constructive in such strenuous conformity in deference to an instructor's micro-management of the organization and content of a paper. It was almost as if the student writers had been forced into a choice of content and style of presentation which doomed the paper to a negative reception by the assessment group. This instance of omission points to an apparent fracture between behavior promoted by an instructor and skills adumbrated in the program goals for General Education, based upon which the rubric has been developed.

#### ◆ **New Assessments**

In an effort to introduce a longitudinal dimension to the assessment of writing in the General Education program at UW-Whitewater, for the first time writing samples were generated and collected from both World of Ideas students and students enrolled in the first semester of Freshman English. Students in World of Ideas represent those who are finishing their General Education sequence. Conversely, students entering Freshman English 101 are just commencing or have yet to begin the General Education block of courses. Assessing writing samples from these two groups lends insight into the writing skills of students as they exit General Education compared to the skills of their peers prior to their exposure to General Education courses.

Writing samples were generated by having students write an impromptu analysis of a vignette which has the flavor of a parable or Zen koan, a short story which seems simultaneously to resist and stimulate attempts at an interpretation (see Appendix Three). 24 writing samples were submitted from a section of Freshman English, and another 33 samples were gathered from a section of World of Ideas. From this set of papers, 49 samples were culled and combined into a single collection containing papers written by students from Freshman English and World of Ideas randomly mixed together. Names were excised from all the papers and numbers affixed, but an indexing of writer and paper number was held in reserve before sending the samples to the assessment team. By this means, a future study could be launched in order to gauge the extent to which rater assessments would correlate with the writer's end-of-course grade.

The 49 writing samples were then scored by the three-member assessment team using the same rubric and scoring scale that was used for World of Ideas term papers. As with the term paper scoring, the collection of short impromptu samples was scored independently by each rater over a time period that covered several weeks. After all the raters had completed scoring the

papers, we met to discuss our results. Following a discussion session on our scoring, three weeks later three samples were re-circulated for re-scoring on a primary trait where there was a significant discrepancy among the scores of the three raters on the first round of scoring. Subsequently, all the raters met for a final time to compare notes, discuss and defend the results of their re-scoring.

Finally, at the end of the spring 2001 semester, the assessment team having submitted the scoring of the impromptu writing, a set of six term papers written by World of Ideas students who had also participated in the impromptu writing exercise was targeted for scoring. These term papers were distributed and scored in a manner identical to the fall 1999 and 2000 assessments. Thus, the assessment team concluded their deliberations having scored both short, impromptu writing as well as long, premeditated term papers written by the same students.

#### ◆ **Writing context**

Students enrolled in one section of Freshman English were asked on the very first day of the semester, at the start of class, before any writing instruction, and even prior to the professor's own introduction, to write on the assessment prompt. They were asked to read the short story and analyze it in their best writing style, and then to indicate the amount of time they spent writing. They were told that they could take up to 10 minutes to complete the analysis. At mid-semester, an entire section of World of Ideas students were similarly enjoined, at the beginning of a class period, to carry out this surprise writing assignment, using the same instructions provided to the Freshman English students.

#### ◆ **Rationale for Assessment Design**

One of the main reasons for adding an assessment of short, impromptu writing to the on-going slate of assessment activities is to try to determine the extent to which the various skills and proficiencies targeted for development in the General Education courses could be assessed in a more streamlined format. Could short writing samples convey the same kind of insight into a student's thinking, voice, and literacy as a term paper? If so, then this analogue to the psychiatrist's Rorschach test might provide a similarly shorthanded diagnostic measure that might be able to supplant more onerous, labor-intensive assessment measures currently in use.

If a sampling of short, impromptu papers could replicate or predict the results of a sampling of longer, more thoroughly premeditated writing, then a host of reasons would argue for its implementation: a) It is capable of being administered with minimal disruption (10 minutes, no advance class preparation necessary). b) No external award system would be necessary to entice writers to participate, creating the possibility of evaluating dispositions to think and write. c) There are tight controls on variables (length of paper, subject matter, preparation time, prewriting activities, revisions, and opportunities for student-teacher conferencing) which are not employed in the present term paper assessment. d) It is easy to increase the sample size without creating an excessive burden of new work. e) Random sampling is more easily achieved when collections can be made from an entire class (When we entrust the World of Ideas instructors with selecting a representative sample from their class, there is a potential conflict of interest that pits the instructor's desire to participate honestly in the assessment with the competing desire put one's teaching in the best light possible). f) From the raters' perspective, the assessment is more focused and sharpened, as it is easier to compare samples that are similar in subject matter and length than it is to compare samples that may arise out of substantially diverse writing contexts. g) Short impromptus would be less likely to lead raters in irrelevant directions. There is more likelihood that everything would be pertinent to the evaluation. Finally, (h) the short impromptu writing is much easier to implement at two distinct educational levels than longer, graded term

paper writing, which would be much more likely to have class or grade level specific features not easily replicable at another level.

◆ **Agenda Setting Issues Related to the Implementation of Assessment of Short Impromptu Writing**

The implementation of an assessment of short impromptu writing in conjunction with the ongoing assessment of term papers raises several issues to be resolved: a) Does the rubric work for shorter writing samples? b) Does the assessment of short writing positively correlate with the educational level of the writers? c) Given that raters scored both short writing samples and longer, planned pieces from the same set of students, does the assessment of short impromptus predict our assessment of the writer's longer, more deliberative work? And (d) to what extent does the shift from short to long format influence the quality of the writing as evaluated by the rubric? Finally, because the identity of the writers of the papers could be retrieved, it becomes possible to inquire (e) if the assessments of short writing and long writing coincide with differences in the student writers' grades.

◆ **Findings Based on New Assessments**

As is the case with the ongoing term paper assignments, findings of a numerical nature await subsequent statistical analysis. However, there are several non-quantitative points that can be made at this time. Perhaps because of the nature of the writing assignment, or constraints on the writing context, or absence of a negative consequence for non-participation, basic writer engagement with the task turned out to be an issue. Some writers opted out of the assignment by expressing helplessness (see Appendix Four, benchmark paper for the assignment of a grade of "1"); others expressed resistance by writing off task. Interestingly, there was a significant element of non-analytic responses to the prompt when it was given as an assignment to a group of professors, harkening back to a finding of Sarah Freedman (1984), who explains that professional writers tend to be downgraded when their work is mixed in with an assessment of more inexperienced writers because of what she describes as "overstepping," which can manifest itself in behavior which either directly or indirectly rejects the assignment (Is this a reflection of the iconoclasm of the professional class? defensiveness? Or does it represent the relative absence of a power differential between test taker and test administrator?).

Writer resistance or disengagement could give rise to questions surrounding what is known as "instrument validity": if the examinee opts out of the task, then the task might be open to the charge of failing to deliver in the assessment of targeted skills. However, if the targeted skills include *disposition* to express thinking, voice, and literacy skills, then *only* an assessment that provoked a certain amount of non-compliance would have the chance of providing a comprehensive measurement of targeted skills. Reflection over the original program objectives for World of Ideas courses, devised with the input of the World of Ideas instructors, shows many objectives which use language like "deal with" and "appreciate" which highlight the dimension of disposition (see Appendix Five). The scholarship on "critical thinking" and "higher order thinking" regularly underscores the central importance of disposition to use the skill to exhibiting mastery of the skill. As Barry K Beyer explains in his article "Critical Thinking: What Is It?" invoking John McPeck, "critical thinking involves not only knowing 'when to question something and what sorts of questions to ask,' but an *inclination* to do so" [my italics, ML] (271).

Raters disagreed over the appropriateness of the prompt, one finding it excessively demanding in that the cultural context for understanding Zen koans is remote from the cultural

context of student writers typically found at UW-Whitewater. Nevertheless, if one of the objectives is to measure the students' familiarity with diverse cultural frames, then a prompt which highlights cultural differences might actually facilitate thinking and writing relevant to program objectives (see in particular objectives one-four, Appendix Five). In any event, difficulty with the prompt should be reflected in the scoring, the summary of which we await.

With the short writing samples, sometimes the rater discrepancies in scoring specific primary traits was very sharp. An especially dramatic case which illustrates the interesting debate over matching the writing to the rubric concerns paper 051 (see Appendix Six for full text). "Thinking" scores assigned to this paper ranged from "4" ("adequate") to "3" ("limited") to "1" ("fundamentally deficient"). After the raters discussed the scoring of this paper, it was included in the set to be re-scored. Although the scores were not identical to the first scoring, the rating discrepancy persisted, as now the paper was assigned "5" ("strong"), "3" ("limited"), and "2" ("seriously flawed"). For the rater assigning "1" and then "2" (and for the rater assigning the "3") the writing was entirely non-analytic. For the rater assigning this paper a "4" and then a "5", the writer had inventively developed a reading "tied to 'normal' logical connections" (which, however, do not reflect the paradoxical connections in the Nasrudin's story). The scoring discrepancy appears to relate to whether writing should be evaluated in relation to the task at hand, or if the student writing can be evaluated to a certain extent without reference to the prompt, and where any analysis is at most implicit in the writing, or, more remote still, implicit in the act of rejection of analysis (!?).

Michael Longrie notes that the scoring fissure for paper 051 and others like it highlights differing attitudes among the raters about the nature of analytical thinking ability, and whether it should include solving logical puzzles like that presented by the koan. He indicates that when there is something clearly premeditated in the text to which we have our students respond, whether or not students "get it" and convey the "right answer" has no bearing on their insight (maybe because it only indicates something about their success in being able to read our minds). As he states, "right answer formulations . . . DON'T measure insight." However, in defense of the koan selection, I believe it could be argued that its very paradoxical nature makes it an especially effective catalyst of more free ranging thinking that transcends any facile attempt at closure. At the same time and *in addition*, there are difficult surface facts that may not be contravened. Therefore, the puzzle solving dimension injects a small "fact of the matter" to the problem at hand without pre-empting, in my opinion, an open ended range of analyses reflecting the criteria surveyed in the rubric under "thinking." Is problem solving an aspect of thinking? I think it is, but this turns out not to be the universal position of our committee.

Ultimately, I would hope that primary trait analysis would permit us to distinguish between papers like 051 and 565 (see Appendix Seven) that, though ranked close in their overall score, nevertheless leave a very different impression, and, in my own mind, differ markedly in exactly ways the rubric is designed to detect.

#### ◆ Scoring Prospects

Having gone through three rounds of scoring papers, we have finally accumulated enough of a sample to propose a set of benchmarks illustrating the scoring scale as it is applied to both the World of Ideas term papers and the short impromptus (see Appendices four and eight for benchmarks and Appendices nine and ten for the full scoring results on the new assessment). The extreme ends of the scale have proven to be the hardest to benchmark: we have no benchmark for "1" ("fundamentally deficient") among the term papers for the simple reason that we have received no "1" term papers. And the benchmark for the "5+" ("strong/outstanding") impromptu

is not without controversy. Although this paper was the only one that garnered a group response that averaged higher than “5,” there is some feeling that the weaknesses of the writing compromise it too much for it to stand as an anchor, or orientation point, for our highest category. Apart from this obscuring at the edges of our scale, these benchmarks should enhance inter-rater agreement on overall scores in the future. Developing benchmarks for each primary trait is considerably more challenging. Ideally, one contemplates a set of primary trait benchmark papers which show the range of scores the scoring scale makes available for the target primary trait, while at the same time controlling for the other non-target traits, thereby minimizing irrelevant variability. Of course, this ideal training scenario makes the assumption that the primary traits represent truly discrete analytic components, an assumption that remains to be validated.

#### ◆ Aspirations

What the process of assessment shows is the need for some conversation between World of Ideas instructors, personnel administrating assessment at UW-Whitewater, and the raters carrying out this assessment. Though the assessment rubric was constructed around the program objectives established when the World of Ideas course was being developed, there has yet to be a discussion which brings together all stakeholders on the success of the fit between General Education objectives and its assessment criteria.

In addition to discussing the fit between program goals and assessment mechanisms, the assessment team is keen to discuss how different classroom strategies and assignments impact on the assessment of student writing. With this assessment, it has become apparent that the type of assignment generating the student writing is a significant indicator of student success in our assessment. We wonder how we can make the assessment fairer to the students and more valid for all of us interested in gauging the success of students in meeting program objectives. At the very least, all of us involved in this writing assessment feel that we need more program coordination so that all sections of World of Ideas DO have students write papers responding to similar prompts. Ultimately, we are interested in how best to advance our student writers by improving our program and by making sure that each and every significant academic success is validated.

#### References

- Beyer, B. K. (1985). Critical Thinking: What is it? *Social Education*, 47(4), 270-276.
- Freedman, S. W. (1984). The registers of student and professional expository writing: Influences on teachers' responses. In R. Beach and L.S. Bridwell (Eds.), *New directions in composition research* (pp. 334-347). New York: Guilford Press.