A Novel Strategy for Assessing the Effects of Curriculum Reform on Student Competence

John C. Wright
Department of Chemistry, University of Wisconsin, Madison, WI 53706

Susan B. Millar, Steve A. Kosciuk, and Debra L. Penberthy
LEAD Center, University of Wisconsin, Madison, WI 53706

Paul H. Williams
Department of Plant Pathology, University of Wisconsin, Madison, WI 53706

Bruce E. Wampold
Department of Counseling Psychology, University of Wisconsin, Madison, WI 53706

Curriculum reform efforts require assessment methods that credibly determine improvements in student learning and skill development (*1–9*). In particular, the current national science standards require instructional methods that develop the habits of the mind that characterize successful problem solving (*10*). Although many studies are directed at assessing the effects of active learning (*11*), few assess the habits of the mind in ways that university chemistry faculty find convincing. Dissemination to skeptical faculty requires development of assessment methods that can assess these skills in order to guide curriculum reform efforts and evaluate their success.

University faculty who introduce active learning methods often perceive striking changes in student competence, but no studies have provided unbiased assessments that document their observations. In this paper, we describe a new assessment strategy that was designed to determine whether such changes in student skills are observable by independent and unbiased observers. The methods were developed by representatives of the University of Wisconsin Chemistry faculty to assess reform success in ways that they would believe. The strategy is applicable to course comparisons that are often found in curriculum reform projects where the project design does not allow the controlled course settings that are sought for educational research. Two sections of a large analytical chemistry course for first-year undergraduates were assessed. One section was taught using methods that focused on lectures that carefully led the student to mastery of the course material using methods that encouraged student questions and participation. The other section was taught using cooperative learning methods that emphasized group work and self-discovery (*12*). These sections are labeled responsive lecturing (RL) and structured active learning (SAL). The SAL approach had been developed in 1992 and refined during the subsequent spring semesters. Unbiased external evaluation judged that both sections represented best practice for each method.

The assessment strategy used an integrated combination of qualitative sociological research methods and oral examinations conducted by 25 university faculty.[1] Both parts of the assessment were done independently of the course faculty. The qualitative methods were chosen to address the problem that the assessment design did not use controls to isolate the cooperative learning parameter. The faculty implemented their own approaches without change in order to teach the courses at the highest level. This approach introduces uncertainties because there are multiple differences between courses. On the other hand, it has the advantage that it can be applied to most courses without appreciably perturbing the course. The success of the approach requires that there are substantial enough differences in the way that students learn the course material that standard qualitative sociological research methods (*13*, *14*) can reliably identify the differences. The qualitative research showed conclusively that student learning in the RL section was primarily individual and was centered on the professor and graduate teaching assistants (TA's), whereas learning for the SAL section was primarily cooperative and was centered on interactions among students.

The oral examinations showed conclusively that independent, unbiased faculty perceived that students in the SAL section had quantifiably better reasoning and communication skills. In addition, student questionnaires showed the differences were apparent to the students. Study of the oral examinations with qualitative research methods showed that the differences in perceived competence were most marked when the oral examinations assessed how students thought about problems. These results should not be interpreted as passing a definitive judgment on the relative merits of the two educational approaches that were compared. Rather, they should be viewed as an indication that it is possible to measure in an unbiased and quantitative way the extent to which the goal of increasing student competence can be achieved. Additional research is required to understand the factors important in defining the reasons for the faculty perceptions of increased competence.

Description of Experiment

Course Structure

The two sections studied were part of the Chemistry 110 course at the University of Wisconsin–Madison, a second-semester accelerated course in analytical chemistry for well-prepared first-year students. The students are primarily science and engineering majors with advanced preparation that places them in the upper 10–15% of entering chemistry students. The student enrollment came from two sections of a first-semester course in general chemistry (Chemistry 109). Approximately 2/3 of the students continue to the Chemistry 110 course. There was little correlation between the sections chosen by the students for the first and second semesters.

The SAL section for the 1995 assessment had 108 students and was characterized by interactive classroom settings, cooperative student assignments and examinations, and open-ended group projects and laboratory experiments (*11*). The students were engaged in the cooperative activities throughout the semester, including the open-ended laboratory projects. The RL section had 95 students and was taught using responsive lectures, spreadsheets, and difficult homework problems. Independent group projects were used during the last 3 weeks of the semester. The faculty in both sections have strong reputations for teaching excellence and equivalent teaching evaluations, and they teach the Chemistry 110 course with similar material and depth.

There were other differences between the sections. In addition to the traditional analytical topics (analytical techniques, absorption spectroscopy, and acid/base, complexation, and oxidation–reduction equilibria), the SAL course covered thermodynamics and chromatography but did not cover precipitation equilibria. The RL course covered the traditional analytical topics including precipitation equilibria. These differences in course content were taken into account in the faculty assessor study by providing the faculty assessors with copies of the course syllabi and the textbook chapters from the sections common to both courses. They were encouraged to ask questions about applications in their disciplines that were unfamiliar to the students in either section.

Faculty Assessor Study

The SAL course instructor perceived a marked increase in the student competence associated with the active learning techniques when SAL was first introduced. To obtain an independent and credible measure of the effects on student competence, faculty, staff, and students at Wisconsin cooperated in an experiment involving a multidimensional assessment strategy. The strategy emerged from an ad hoc committee of skeptical chemistry faculty who met prior to the 1995 course. They concluded that the only type of assessment data they would find credible would be faculty-conducted oral examinations of all students. It was important that the assessment be done orally in order to probe student understanding and problem-solving ability. It was also important that the assessment involve external faculty who are independent of the course faculty.

To implement the ideas, we enlisted 25 University of Wisconsin–Madison faculty assessors from science, mathematics, and engineering departments other than chemistry.[1] All but the four mathematicians used the ideas in Chemistry 110 in their own research. Each assessor was asked to design a 30-minute oral examination with which they would rank the competence of the students they examined. The faculty were allowed to make their own definition of student competence and design their own oral examinations. This approach provides representation for the diversity of faculty definitions of student competence. The faculty were asked to provide written descriptions of their criteria for judging competence and were interviewed by UW-Madison LEAD Center (Learning through Evaluation, Assessment, and Dissemination) researchers to better understand the nature of their oral examinations.

The students were divided into octiles of roughly 24 students based on their rank in class in Chemistry 109. An average of three faculty were assigned to examine each octile and each assessor typically saw eight students for 30 minutes each. The students were equally divided between the SAL and RL sections. The faculty were not apprised of this student assignment strategy, the lecture section of any student, or the teaching methods used by either section. The students were told that a small portion of their grade would come from the oral assessment.

Qualitative Research Study

It was important to have independent evaluation data that would identify the reasons for differences between the two sections. In particular, it was important to determine whether the teaching methods were representative of cooperative learning methods and individual, lecture-focused learning methods and whether the student learning patterns reflected the teaching approach. It was also important to understand and document the nature of the faculty's oral assessments. These goals were reached through a study conducted by an independent third-party evaluation team from the UW-Madison LEAD Center.

The LEAD Center used interviews, observations, and surveys to study the students, teaching assistants, and faculty in both sections throughout the semester. It also collected and analyzed the interview, ranking, and questionnaire data for the faculty assessor study. Interviews were transcribed and analyzed by established inductive analysis methods in social science (*13, 14*). The analysis produced two case studies, as the patterns of social interaction and student and instructor experiences that emerged from the interviews, observations, and open-ended survey responses were too different to sustain a comparative analysis. Comparative statistical methods were used to analyze the ranking and numeric questionnaire data.

Results of LEAD Qualitative Assessment

The RL lecture section was characterized by a lecturing style that was effective in eliciting student responses and involvement in the lecture material. This approach also included very challenging homework problems, quantitative examination questions, well-defined laboratory experiments, an open-ended laboratory project, and an unannounced absolute grading scale. Cooperation among students was neither encouraged nor discouraged. The SAL lecture section used open-ended question–answer sessions during lectures and added cooperative learning methods that included an absolute grading scale, cooperative computer projects, open-ended laboratory projects, oral examinations on the project results, research paper analysis, a "student board of directors", and cooperative examinations. Both sections required the same textbook (*15*) but each used it primarily as a reference.

The LEAD Center classroom observation and interview data established that the teaching strategy implementation for each lecture section was at the high end of the performance scale. Student attendance, attentiveness, and participation were very high in comparison with other large lecture courses and students gave their lecturer high marks for the skill and care with which the lectures and course components were implemented. The interview data also established that although enrollment from Chemistry 109 showed some difference in self-selection between sections, the differences did not correlate with the faculty assessor rankings. The average rank in class from Chemistry 109 was the 60.3 percentile for SAL and the 60.7 percentile for RL, so students entering both lecture sections had performed equivalently.

The interview data showed sharp differences in the nature of the learning interactions. The SAL learning was characterized by student–student interactions. Students stressed the importance of the research-oriented, structured group activity and indicated that group interactions helped connect the lecture, laboratory, and other course components. Generally, students felt the class atmosphere fostered support and cooperation, although 20% of the students either expressed distress about dysfunctional student groups or expressed a strong preference for working independently. The people who flourished in the cooperative environment enjoyed the challenge of solving open-ended problems and acquired greater self-reliance. All students, even those with dysfunctional groups, also felt that the course structure fostered an awareness of the complexity and frustrations that accompany genuine research. Most SAL interviewees commented that the workload was very heavy, although many also added that it was worth it. These observations show that the SAL section is an appropriate example of courses stressing active learning.

The RL learning was characterized by a focus on the lecturer as the authority for knowledge. There was an appreciation and respect for the professor, with some viewing him as a role model. There was also a strong sense of accomplishment in mastering the material using the lecturer's step-by-step problem-solving approach and mathematical modeling method. The professor was viewed as practically the sole source of information and understanding by some, while others mentioned the TAs and other students as valuable resources. Many students spontaneously formed groups that greatly assisted their learning, but other students preferred to work alone, or were unsuccessful in becoming a member of a suitable group. Some students reported frustration about their inability to connect lecture concepts and laboratory experiences. Many RL students commented that the workload was heavy, while others considered it about right. These observations suggest that the RL section is an appropriate example of courses stressing the lecturer as the focus for learning.

Results of External Faculty Assessment and Questionnaires

The faculty assessors were coordinated by an objective, external faculty member and the LEAD Center. Assessors constructed their own structures for the student orals, formed their own criteria for assessing competence, and filled out a pre-examination survey about their approach to the oral examination. They completed a second survey after the oral examinations that reported changes in their ideas, methods, and criteria for competence. Both faculty and students filled out questionnaires for each oral. LEAD Center personnel conducted follow-up interviews with all faculty assessors to document the nature of the orals.

The questionnaire data are summarized in Table 1. The probability values or *p* values represent Mann–Whitney tests of significance (*16*) and indicate the probability that the

Table 1. Faculty and Student Questionnaire Results

Question	Mean SAL	Mean RL	*p* Value
	1 disagree–6 agree		
Faculty Questionnaire			
Student felt at ease	4.74	4.70	.89
Student is well-prepared for other science courses	4.84	4.48	.045
Confidence that student performance on oral reflects true competence	4.68	4.45	.22
This student demonstrated overall competence	4.79	4.17	.0013
Relative Rank (1 is most competent)	3.68	4.80	.0066[a] (.023)[b]
Adjusted Absolute Rank[c] (1 is most competent)	1.77	2.22	.0002
Student Questionnaire			
I felt at ease	4.90	4.95	.64
I demonstrated what I learned	4.63	4.18	.026
I demonstrated ability to relate knowledge in new contexts	4.68	4.36	.12
I was fluent in responding to questions	4.39	3.91	.014
I demonstrated I am knowledgeable in chemistry	4.74	4.27	.0066
I appeared nervous	3.06	2.93	.54
I feel well prepared for other science courses	4.99	4.42	.0005
Compared to other college courses, Chemistry 110 is at the top	4.77	4.05	.0027
What grade would you assign yourself based on the competence you demonstrated?[d]	88.5	78.6	<.001
How many hours did you spend out-of-class in Chem. 110?[e]	3.96	3.47	.024
What portion of time did you spend working with other students on out-of-class work[f]	3.43	2.32	<.001

[a] *p* value based on Wilcoxon matched-pair sign rank test. [b] *p* value based on signed test. [c] Faculty placed students on an absolute scale, similar student performances were grouped together, and the groups were ordered in rank from first to last. [d] Grade based on 0–100 scale. [e] Numbers refer to the seven choices on the questionnaire. The choices were (i) 0–3; (ii) 4–7; (iii) 8–11; (iv) 12–14; (v) 15–17; (vi) 18–20; or (vii) more than 20 hours/week. [f] Numbers refer to the five choices on the questionnaire. The choices were, in terms of percent, (i) 0–20; (ii) 20–40; (iii) 40–60; (iv) 60–80; or (v) 80–100%.

observed differences are statistically identical. The Mann–Whitney test was used because it does not depend on the nature of the data distribution.

First, it is important to realize that the questionnaire data show that the students in both sections had very favorable answers to all the questions about satisfaction and accomplishment in the courses. This comfort with the course was reinforced by the LEAD interview data showing that students' feeling of accomplishment and satisfaction with the course was higher for both sections than for other comparable courses. The faculty assessor data also show that the assessors were impressed with the knowledge and ability levels of the students in both sections. This observation reinforces the LEAD Center data showing that both sections were taught at the high end of the faculty performance scale. Data from both the faculty and student questionnaires show no differences between lecture sections in student nervousness, but there are marked differences in all the questions on preparation and performance. Both student and faculty questionnaires showed very significant differences in the perception of the student preparation for future science courses. The student answers also showed substantial differences in the students' perception of how well they demonstrated their learning, how fluent they were in answering questions, and how knowledgeable they appeared. The SAL students appeared to spend 15% more time in out-of-class work and 56% more of that time working with other students.

These differences in perception are also reflected in the faculty assessments of performance. Assessors defined both a relative rank where their students' competence was ranked from first to last (we label this approach "forced relative assessor ranking") and an absolute score where the individual students' performance was placed on a continuum from low to high competency (labeled "absolute assessor ranking"). The relative ranking strategy forces decisions that accentuate differences between students and make it easier to discern differences associated with teaching method. At the same time, it obscures the magnitude of the differences. It also introduces correlation between the two sections, because if one section does better, the other must do worse. This correlation prevents the use of statistical techniques based on the normal distribution. The absolute score gives complementary information about the magnitude of differences. An "adjusted absolute rank" was defined by grouping students with similar scores to distinguish them from students who differed more markedly. For example, if an assessor clustered 4 students near the top of the continuum, 1 student above average, 2 students near the middle, and 1 student near the bottom, we would assign an adjusted absolute rank from 1 (assigned to the 4 best students) to 4 (assigned to the student with the lowest score). Although the number of values and the assigned scores differed between faculty members, this ranking strategy provides an absolute measure of student competence for any particular faculty assessor. It also eliminates the effects of outlying values and the correlation problem, so normal statistical methods are justified.

The differences in relative ranking between sections are the most significant indicators. The typical assessor had 4 students from each section, whom they ranked first though eighth. If all 4 students in one section were ranked ahead of the other 4, the average rankings would be 2.5 and 6.5 and their difference would be 4 for that assessor. When this approach was averaged over all 25 assessors, the maximum possible difference was actually 3.46 because some faculty gave some students equal ranks and some faculty saw different numbers of students. To see whether there were statistically significant differences between sections, the Wilcoxon matched-pair signed-rank test was employed using the relative ranks of each assessor and the sign test was employed using the sign of the rank difference between sections (*16*). Both methods are robust, nonparametric statistical tests that are independent of the type of statistical distribution. They are the most common tests of significance when the data are not normally distributed. Table 1 shows that the relative rank is 4.80 for the RL section and 3.68 for the SAL section. The 1.12 difference between sections is statistically significant with a p value of .0066 for the Wilcoxon matched-pair signed-rank test and .023 for the more rigorous sign test. It is noteworthy that this difference is 1/3 of the maximum possible difference that could have occurred.

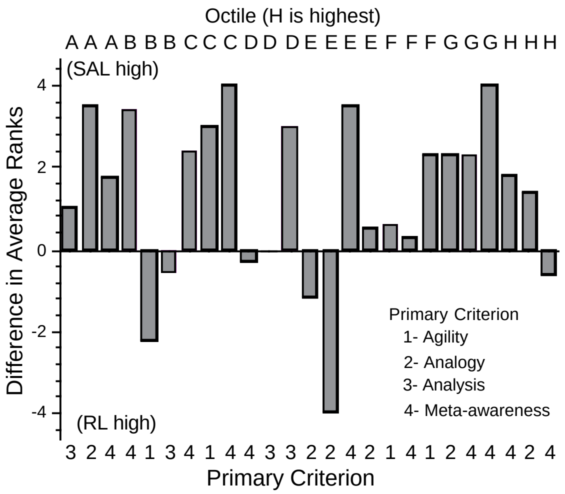

Figure 1. The bars show the difference in student ranks, for each assessor, between SAL and RL students. The letters above the bars indicate the octile of students that the assessor interviewed, A being the lowest and H the highest octile. The numbers below the bars indicate the classification of the criterion used by the assessor according to the code indicated in the figure.

The section differences from each assessor are shown in Figure 1. The student octile seen for each assessor is indicated at the top of the figure. The largest differences are seen at the bottom and the middle octiles but the differences are significant for all octiles.

Similar results were found for the adjusted absolute rank of overall competence (see Table 1). The differences in sections for the adjusted absolute rank and the overall competence question were very significant at a much lower p value because they represented independent observations. The relative rankings, adjusted absolute rankings, student competence question, and grades the students assigned to themselves all reflect substantial differences in the competence demonstrated by the students.

Analysis of Correlations

To determine the nature of the individual assessor's exams and identify further reasons behind the differences, the LEAD Center analyzed the faculty assessors' criteria for the relative

Table 2. Student Performance for Different Assessor Approaches to the Oral Discussions

Category Subcategory	Assessor Faculty (No.)	Mean Relative Rank		*p* Value[a]
		SAL	RL	
Outcomes	10	4.2	4.5	> .37
Analogy	6	4.4	4.7	> .60
Analysis	4	3.5	4.2	> .28
Process	15	3.3	5.0	< .01
Meta-awareness	11	3.3	5.3	< .01
Agility	4	3.2	4.2	> .27

[a] Matched-pair sign rank test.

Table 3. Results of Analysis of Variance and Correlation between Experimental Variables

Factor 1	Factor 2	Factor 3	*p* Value
	Self-Selection Effects		
Self-selection	109 lecture		> .31
Self-selection	octile		> .23
Self-selection	gender		> .38
Self-selection	110 lecture	assessor rank	> .81 (RL) > .17 (SAL)
	Gender Effect		
Gender	assessor rank		> .85
	Grade Effects		
109 rank in class	110 grade		< .05
109 rank in class	assessor rank		< .05
110 grade	assessor rank		< .05
109 rank in class	110 grade	assessor rank	< .05
	Chemistry 109 Effects		
109 lecture	assessor rank		< .05
110 lecture	assessor rank		< .05
	Student Effects		
Time spent out of class	assessor rank		< .10
Time spent out of class	110 lecture		< .10
Time spent out of class	110 lecture	assessor rank	< .10
Time spent in groups	110 lecture	assessor rank	> .64

rankings using the data from the personal interviews, pre- and post-assessment surveys, faculty reports on each student, and student surveys about the oral exam. All the assessors asked students to demonstrate a basic knowledge of chemistry and an ability to use the knowledge in a way that required an integration of abstract principles. The differences in assessor approach fell into two broad categories labeled “outcomes” and “process.” Assessors in the outcomes group tended to use the examination to measure the command of the material, whereas assessors in the process group tended to use the examination to observe how the students approached new problems. The outcomes category was further subdivided into an “analogy” subgroup (6 of 25 assessors) that used problems requiring students to relate a new problem to their course material and an “analysis” subgroup (4 of 25) requiring students to solve a problem that was unrelated to the course material. The process category was subdivided into an “agility” subgroup (4 of 25) that measured how rapidly students could analyze and solve a problem or how effectively they could react to new information and a “meta-awareness” subgroup (11 of 25) that focused on the thinking patterns (did the students self-correct, have a variety of perspectives, understand the larger context surrounding a particular problem, or relate theory and practice?). The primary criterion used by each assessor is indicated on the bottom of Figure 1.

The relative rankings showed a strong correlation with the assessor subcategory. The results are summarized in Table 2. Although SAL students outperformed RL students in all subcategories, the assessors in the meta-awareness subgroup found the largest differences between sections—almost 1/2 the maximum possible differences. This finding indicates that the major reason for the large difference in student competence was the thinking process that students displayed during the oral examination. In the analysis and agility subgroups, the differences became smaller but were still 25% of the maximum possible difference. The differences are not significant, however, because of the small number of assessors in each category. In the analogy subgroup, the differences are still smaller and are not significant.

These results have three important implications. First, they suggest that the differences in perceived student competence are correlated with the ability of students to demonstrate higher-level thinking skills. Second, if developing higher-level thinking skills is a central goal, assessments must involve problems and provide opportunities where the complexity of the student’s thinking process is exhibited. Third, it is interesting that although a large proportion of the faculty used meta-awareness as their primary criterion for competence, interviews with the faculty showed that these same faculty did not use this criterion for their own courses.

The qualitative research on the faculty oral assessments also addressed the question of whether the faculty felt their interviews were a credible assessment of student competence or simply reflected improved oral communication. All the faculty felt that the oral examinations did provide a true assessment of student competence and were not unduly biased by student communication skills. Many provided specific examples of students with poor communication skills who nevertheless demonstrated a superior ability to handle unfamiliar problems. Many also provided examples of how students’ discussion reflected their logical thinking as they sought a path towards a problem’s solution.

To discover the correlation between the variables in the experiment, an analysis of variance (ANOVA) was performed (*16*). This method identifies the factors that are statistically correlated. The variables studied were whether the students self-selected into a particular section; the section of Chemistry 109; the section of Chemistry 110; the student octile in Chemistry 109; the student gender; the relative rank given by the assessor; the grade in Chemistry 110; the time spent out of class; and the time spent in groups. Table 3 summarizes the results of log-linear ANOVA analyses of the course data for two-way and three-way correlations. A two-way correlation between two variables means that the values of one variable helped determine the value of the other; a three-way correlation means that two variables together were correlated with the value of the third.

There were no significant two-way correlations between students who self-selected into specific lecture sections and the previous Chemistry 109 lecture section; Chemistry 109 octile; gender; or assessor relative rankings. There were also no significant effects of gender on the assessor relative rankings. There were significant two- and three-way correlations between the Chemistry 109 rank in class, Chemistry 110 grades, and assessor relative rankings. The two-way interactions indicated that for each lecture, students receiving AB (intermediate grade between A and B) or better were proportionately overrepresented among students in the top assessor ranking. The three-way interaction indicates that students who received an AB or better in Chemistry 110 and were in the upper half of their Chemistry 109 class were overrepresented among students in the top assessor ranking. Both two-way correlations between the assessor ranking and the Chemistry 109 lecture section or the Chemistry 110 lecture section were significant. It is interesting that the Chemistry 109 lecture section that produced students with the better assessor rankings was taught by a university teaching award winner. The three-way correlation, however, was not significant. Some two- and three-way interactions between time spent, lecture section, and assessor ranking were fairly significant. A two-way interaction indicated that proportionately more students in the SAL section spent more than 15 h/week out of class than in the RL section. Interestingly, there was also an over-representation of the 0–7 h/week students in the top rank. The three-way interaction indicated that among the >15 h/week students, a student in the RL section was more likely to be ranked second than first compared to an SAL student. Finally, the time spent working with others did not correlate with the assessor rankings for either lecture section. This finding is interesting and should be investigated further. It suggests the presence of threshold effects where some degree of student–student interaction causes the differences in student competence. A better understanding of this effect may allow optimization of the time required to implement cooperative learning methods.

Effects on Faculty Evaluations

The changes to an active learning format initiated in 1992 caused significant changes in the student course evaluation forms of the SAL instructor between the 12 times it was taught with traditional lectures and the 4 times it was taught with SAL. Answers to the questions “Was the course interesting?” and “Was the instructor effective?” improved significantly from 3.87 to 4.36 and 4.22 to 4.55, respectively, on a 5-point scale. There were no significant changes in the student evaluations of the instructor’s preparation level, the background assumed for the course, or the pace of the course. The amount of problem work assigned did change significantly from 2.91 to 3.24 on a 5-point scale, where 3 is labeled “about right.”

Conclusions

This study was designed to implement a new assessment strategy and use it to determine whether unbiased external faculty assessors could identify differences in student competence resulting from use of an SAL strategy. The definition and measurement of student competence in the two sections was controlled by individual faculty assessors in client departments, so the assessment represents the range of competence definitions and questioning strategies that one would expect in a major research university. It relies on individual faculty members determining student competence in a manner that they find individually credible. Although there may be individual differences in the success of the faculty methods, the overall performance represents the best measure that faculty can form by individualized oral assessment. The forced relative assessor rankings hid the fact that the assessors felt almost all the students were quite competent and accentuated the differences that resulted from the SAL experiences, especially when the faculty used competence criteria that monitored the sophistication of student higher-order critical thinking skills. The student survey data indicate that students felt the performance differences were related to their ability to demonstrate what they had learned and their preparation levels. The faculty assessor data indicate that differences were related to the thinking skills the students exhibited.

It is important to recognize that the goals for the two lecture sections were different and that this study was designed to test the attainment of the SAL section’s goals, that is, improved student competence in thinking and solving new problems. The study was not designed to test mastery of course content. The oral exam format did not stress the specific material actually covered in Chemistry 110, but emphasized the application of the material to new contexts. There was no common written examination to test relative performance on the specific course material. There is no reason to believe that the SAL students would do better or worse on traditional examinations. In fact, when the SAL instructor gave traditional timed examinations with challenging quantitative problems from examination questions used prior to changing to cooperative methods, the student performance was not significantly different. This latter result is important because it suggests that the oral examinations measure different aspects of student skills than written examinations. We believe that the oral examinations reflect the scientific maturity of a student while traditional examinations reflect command of the subject matter. This hypothesis indicates that curriculum reform is most effective in changing student habits of the mind. It is important to test this hypothesis further and identify the specific factors responsible for the differences before one can use oral assessment as a reliable method for evaluating changes in the habits of the mind.

There are many additional questions that this study did not address and that future work should consider, including: Were the effects instructor dependent? Can others repeat the results? Are the same differences observable with students who are not accelerated? Are the differences observable in courses that are carefully controlled to isolate the differences in teaching method? Do the differences persist over time? and Is the content knowledge improved as well?

This study may have implications for efforts to change learning because it supports the idea that oral discussions are more effective in measuring changes in thinking skills than standardized examinations. It also represents an approach that can be used in settings where controlled educational research studies are not feasible. For example, the approach can assess differences in student preparation at the beginning of a course. A separate consequence of the assessment was that 25 faculty from across the university became more interested in active learning methods as a result of interacting with enthusiastic students. This new assessment strategy may therefore be part of a dissemination strategy as well. It is important for other faculty to try similar experiments, using their insights and ideas to discover the methods that will optimize science education.

Acknowledgments

This work was sponsored by the National Science Foundation under grant DUE-9455928, the Advanced Research Projects Agency under grant EEC-8721545, and the College of Letters and Science at the University of Wisconsin. We would like to express our gratitude to the cooperating faculty members in the comparison lecture section, the students in the two courses, and the 25 faculty volunteers whose generous donation of time made this study possible.

Note

1. The faculty assessors and departments were Biochemistry: A. Attie, T. Martin, D. Nelson; Biomolecular Chemistry: P. Bertics; Chemical Engineering: K. Bray, D. Cameron, J. DePablo, C. Hill, S. Kim, T. Root; Clinical Chemistry: M. Evenson; Geology: L. Baumgartner, J. Valley; Materials Science: S. Babock, E. Hellstrom, R. Matyi; Mathematics: S. Baumann, M. Certain, T. Millar, A. Nagel; Pharmacy: R. Burnette, K. Conners, C. Royer, G. Zografi; Soil Science: P. Helmke.

Literature Cited

1. Rubinstein, E. *Science* 1994, *266*, 843–875.
2. Barr, R. B.; Tagg, J. *Change* 1995, *1995*(Nov/Dec), 13–25 .
3. *They're Not Dumb, They're Different*, 1st ed.; Tobias, S., Ed.; Research Corporation: Tucson, 1990.
4. *Revitalizing Undergraduate Science*, 1st ed.; Tobias, S., Ed.; Research Corporation: Tucson, 1992.
5. Newmann, F. M. *Phi Delta Kappan* 1991, *72*, 458–463 .
6. MacGregor, J. *New Directions for Teaching and Learning*. No. 42; Jossey-Bass: San Francisco, 1990; pp 19–30.
7. Schoenfeld, A. H. In *Cognitive Science and Mathematics Education*, 1st ed.; Schoenfeld, A. H., Ed.; Lawrence Erlbaum: Hillsdale, NJ, 1987; pp 189–215.
8. Heller, P.; Keith, R.; Anderson, S. *Am. J. Phys.* 1992, *60*, 627–636.
9. Heller, P.; Hollabaugh, M. *Am. J. Phys.* 1992, *60*, 637–644.
10. American Association for the Advancement of Science. *Benchmarks for Science Literacy*; Oxford University Press: New York, 1993.
11. Johnson, D. W.; Johnson, R. T. *Cooperation and Competition—Theory and Research*, 2nd ed.; Interaction Book Company: Edina, MN, 1989.
12. Wright, J. C. *J. Chem. Educ.* 1996, *73,* 827–832.
13. Patton, M. Q. *Qualitative Evaluation and Research Methods*; 2nd ed.; Sage: Newbury Park, CA, 1990.
14. Coffey, A.; Atkinson, P. *Making Sense of Qualitative Data: Complementary Research Strategies*; Sage: Thousand Oaks, CA, 1996.
15. Harris, D. C. *Quantitative Chemical Analysis*, 4th ed.; Freeman: New York, 1995.
16. Johnson, R. *Elementary Statistics*; PWS-Kent: Boston, 1988.