ED 475 954 IR 021 952

AUTHOR Vendlinski, Terry; Underdahl, Jennifer; Simpson, Elise; Stevens, Ron

TITLE Authentic Assessment of Student Understanding in Near-Real Time!

PUB DATE 2002-06-00

NOTE 16p.; In: NECC 2002: National Educational Computing Conference Proceedings (23rd, San Antonio, Texas, June 17-19, 2002); see IR 021 916.

AVAILABLE FROM For full text: http://confreg.uoregon.edu/necc2002/ .

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC01 Plus Postage.

DESCRIPTORS Chemistry; *Computer Assisted Instruction; *Computer Assisted Testing; Educational Researchers; High Schools; Instructional Effectiveness; Instructional Materials; Multimedia Materials; Pedagogical Content Knowledge; Science Instruction; Science Teachers; Secondary Education; Student Evaluation

ABSTRACT

While most would probably agree that the ultimate success of an educational system should be graduates who are able to evaluate, synthesize, analyze and apply what they have learned, routinely assessing such understanding in students has proven difficult, for a variety of reasons, using most traditional methods. Recent developments in cognitive science, information technology, and analytical tools, however, have provided educators and researchers a means to overcome many of these barriers. This paper explains how high school teachers and educational researchers worked together to apply these developments to successfully implement and assess case-based problem-solving in a typical high school chemistry curriculum. A quasi-experimental study was conducted with 134 first-year high school chemistry students using the Interactive Multi-Media Exercises (IMMEX[TM]) computer based learning and assessment tool to solve qualitative chemistry problems. The students' teacher constructed the problem for her students using the same tool. Results of this study show a strong and significant correlation between computer-aided measures of student understanding and the teacher's manual evaluation of student understanding using her own rubric. While both measures demonstrated that content knowledge was important, neither proved just a surrogate measure of content knowledge. Neither measure demonstrated gender, ethnic or socioeconomic bias. The results suggest that the data collected by this tool not only allows for valid inferences of a student's content understanding, but allows such assessment to happen in a classroom setting in near real time. (Contains 46 references and 5 tables.) (Author)

Authentic Assessment of Student Understanding in Near-Real Time!

Terry Vendlinski[1], Jennifer Underdahl, Elise Simpson, and Ron Stevens
UCLA / IMMEX Lab
5601W. Slauson Ave. Suite 255
Culver City, CA 90403
[1] Email: vendlins@ucla.edu

A paper presented at the
National Educational Computing Conference
San Antonio, TX
June 17 – 19, 2002

Abstract

While most would probably agree that the ultimate success of an educational system should be graduates who are able to evaluate, synthesize, analyze and apply what they have learned, routinely assessing such understanding in students has proven difficult, for a variety of reasons, using most traditional methods. Recent developments in cognitive science, information technology, and analytical tools, however, have provided educators and researchers a means to overcome many of these barriers. This paper explains how high school teachers and educational researchers worked together to apply these developments to successfully implement and assess case-based problem-solving in a typical high school chemistry curriculum. We conducted a quasi-experimental study with 134 first-year high school chemistry students using the Interactive Multi-Media Exercises (IMMEX™) computer based learning and assessment tool to solve qualitative chemistry problems. The students' teacher constructed the problem for her students using the same tool. The results of this study show a strong and significant correlation between computer-aided measures of student understanding and the teacher's manual evaluation of student understanding using her own rubric. Moreover, while both measures demonstrated that content knowledge was important, neither proved just a surrogate measure of content knowledge. In addition, neither measure demonstrated gender, ethnic or socioeconomic bias. The results suggest that the data collected by this tool not only allows for valid inferences of a student's content *understanding*, but allows such assessment to happen in a classroom setting in near real time.

Background

The assessment of student learning continues to occupy a prominent place in the debate about the efficacy of American education. As of January 2002, each of the 50 states and the District of Columbia had instituted some program of comprehensive student assessment (Olson 2002), and the most recent reauthorization of the federal Elementary and Secondary Education Act (ESEA), dubbed "The No Child Left Behind (NCLB) Act of 2001," significantly increases the stakes of these state tests by predicating federal funding to the states on improved student test results. Although many of these tests are not currently aligned with the specific standards guiding actual classroom instruction, in every case the results of such testing are intended to measure student learning and to ensure accountability (Goertz and Duffy 2001). While these two ends seem to justify the need for some form of external assessment, the consequences of current tests have nonetheless sparked controversy. In fact, some feel strongly that these consequences alone can adversely affect the validity of the interpretations made from the data produced by such testing (Messick 1989). For many teachers, these consequences include the amount of curricular time devoted to student testing (academic, administrative and reporting), the content that will appear on such tests, and the fact that these tests are added to, not integrated with, the curriculum. Since the late 1950's, there has also been a growing consensus that students must be able to do more than just recall concepts. According to Codding and Rothman (1999), "decades of research on student learning suggest that instruction should begin with clear expectations of what students should know and be able to do, and should provide students with opportunities to demonstrate their understanding in increasingly complex ways until they meet those expectations" (pg. 15). If, as researchers from Bloom (1956) to Busching (1998) suggest, the goal of education is to teach concepts which students can then evaluate, synthesize, analyze and apply to solve new problems in often unanticipated contexts, our assessments of student learning must not only overcome the problems described above, but must also be able to assess how well a student *understands* the concepts taught in America's classrooms. Nevertheless such higher order thinking skills have been difficult to measure, especially if classroom teachers are not encouraged to propose novel problems for which students have not already memorized the answers (Stiggins, Rubel et al. 1988). As a result, most educators and educational systems have been constrained to evaluate student academic ability almost exclusively by measuring the quantity of knowledge a student can recall (Bloom 1956; Wiggins 1993). As Resnick and Resnick (1992) put it, "What is easy to measure gets measured."

Unfortunately, as Alberts (2002) so aptly notes, "Memorization is not understanding." Limiting the determination of student ability and the measure of school effectiveness to how well a student recalls specific facts can have serious negative ramifications for both student learning (Resnick 1987) and instructional quality (Corbett and Wilson 1991; Viadero 2000), can leave serious foundational misconceptions intact among students (Mestre 1991), and encourages students to verify "expected" outcomes instead of learning (Rudd, Greenbowe et al. 2002). Moreover, because of working memory limitations, when students merely memorize concepts rather than form rich relationships between them, students are less likely to solve problems using these concepts and are unlikely to show the ability to transfer these concepts to other problems (Gabel and Bunce 1994). Arguably, these consequences are amplified when test results are used to make high-stakes decisions (Corbett and Wilson 1991; Shepard 1995).

In an effort to overcome these limitations, many have advocated the use of assessments that require students to not only learn important concepts, but to actually apply these concepts to solve realistic problems indicative of what they might encounter outside of the classroom (see for example USDoEd 1983; Rutherford and Ahlgren 1990; USDoEd 1993; NRC 1996; NRC 1999). Such "authentic," case-based scenarios have been used for some time in medical and business schools (Elstein 1993), and they are now gaining increasing popularity among undergraduate, secondary, and even primary educators (e.g. Libarkin and Mencke 2002). While such assessments can provide evidence of the higher order skills explicated in Bloom's taxonomy (1956) and may hold the promise of improving education, many of the same barriers that make more traditional assessments difficult to implement can also be barriers to such "authentic" assessments as well. Authentic test items can entail logistic (Quellmalz, Schank et al. 1999) and pedagogical hardships (Lowyck and Poysa 2001), have been difficult to validate (Barton 1999), and can complicate both individual student and system-wide evaluation. In fact, many have cited good standards and rubric development as two of the most difficult obstacles to credible performance-based assessment (Raizen 1990; Arter and McTighe 2001). Recent developments in cognitive science, information technology, and data analysis tools, however, suggest that educators, researchers, and policy-makers now have the tools to conquer these barriers.

As educational assessment has become increasingly informed by developments in the field of cognitive science, the focus of such assessments has shifted away from the specific items and tasks a student must perform to the constructs necessary to understand what a competent performance in a domain looks like (Pellegrino, Chudowsky et al. 2001), p. 344). Fischer (1997) suggests that in order to take advantage of this progress and improve our assessment of student understanding, we must capture the skills and context of the diversity of human problem solving and find order in that diversity. A static student model is insufficient to account for the variability inherent in such performances and, as Wiggins (1993) has suggested, patterns of student performance will emerge when data is collected using a variety of means. In fact, federal statues expect the use of such multiple measures as well (Goertz and Duffy 2001). One benefit of moving from the more static Piagetian states to dynamic models of development has been our ability to find order in student diversity.

While the large amounts of detailed data required for such a process were difficult to collect and analyze before the age of information technology, recent developments in information and computer science now offer the tools necessary to create more dynamic and valid representations of student problem solving inexpensively and in near-real time. By using such tools, teachers and educators can determine not only what a student answered, but also how the student went about arriving at such an answer. The information provided by these types of assessments allows both for new insights into how a student conceptualizes a problem space and, as will be shown, for more valid inferences about student *understanding*. In addition, "stimulating and engaging assessments improve student learning, not just measure it, by providing students with opportunities to perform tasks that challenge them to use their knowledge and by providing teachers with examples of the kinds of performances students can produce in their classrooms day after day." (Tucker 1999, p. 38). "Whatever the project or problem," says Roberta Furger (2002), "well-crafted performance assessments share a common purpose: to give students the chance to show what they know and can do and to provide teachers with the tools to assess these abilities."

Concurrent with developments in the field of human cognition, advances in computer technology have also made possible more advanced data analysis tools. Specifically, new pattern recognition algorithms (both statistical and adaptive "artificially intelligent") make it quick and inexpensive to discern recurring problem-solving strategies that students use in a particular context. In addition, such tools overcome inherent human limitations on the amount of information we are able to process and remember, and thereby increase the number and types of classifications student performances can be separated into (Miller 1956). Furthermore, these advances have now afforded us the ability to fit cognitive models to the student a posteri and not constrain a student to some pre-conceived model a priori. The real significance of these developments, however, is not just that we can observe students as they change the types of strategies they employ to solve problems in various contexts, but that we might very well be able to identify when specific interventions would be most appropriate to improve the understanding of individual students, and even what those interventions might be. Hartley and Bendixen

(2001), for example, argue that it is this use of testing that will ultimately improve teaching and learning in our schools.

The purpose of this paper is to explain how high school teachers and educational researchers worked together to apply these developments in order to successfully assess the understanding of individual students based on the performances of these students solving case-based problems in a typical high school chemistry curriculum.

Methodology

We conducted a quasi-experimental, time interrupted sequence study on a group of 134 first year high school chemistry students to assess how well they understood a curricular unit on qualitative chemistry. The students lived in a largely middle to upper-middle class suburban community in Southern California. Student grade point averages, first semester grades, and student demographic data suggest this population is typical of student populations at suburban American high schools (NCES 1996) with the exception that African American students were under represented and Asian American students were overrepresented in this group. We have, however, found trends similar to those reported here in student groups where African American students were overrepresented and Asian American students were underrepresented in the study cohort. Nevertheless, the small number of students in specific non-white ethnic groups in these studies made it impossible to investigate the statistical relationships between particular ethnic groups and other variables in this research. Consequently, this study considers ethnicity to be a dichotomous variable.

As anticipated from the work of Gardner and others (Crouse and Trusheim 1988; Gardner 1993), pre-treatment data suggested that the best predictor of a student's future grade or overall teacher ranking of student ability was previous grade. Consequently, we first investigated correlation between various grade measures to determine if these metrics were consistent in their measurement of student ability.

As part of the post-treatment phase of this study, the students also took the Stanford Achievement Test, 9th edition (SAT-9). The SAT-9 taken by these students was a norm-referenced, content-based, multiple choice achievement test (Harcourt Educational Measurement 2001). The science and math portions of the test were entirely multiple-choice and were, at the time of this research, the primary vehicle used by the state of California to measure the math and science proficiency of its high school students. The results of this statewide assessment of math, science and reading allowed us both to compare evaluations of student ability as determined by grade with this assessment, and to compare the metrics of understanding with these same results. Because written, forced-choice items were used almost exclusively to evaluate these students, we suspected that these tests might not be fully measuring what a student *understood* and could apply in solving problems within a specific domain.

Consequently, we have used a new tool, the Interactive Multi-Media Exercises (IMMEX™) to develop strategy-based assessments of student understanding. The

IMMEX software consists of three modules: a problem authoring module, a problem presentation module, and an assessment module that unobtrusively records a student's progress through the problem space for later analysis.

Teachers use the IMMEX authoring tools to develop problem sets that address their pedagogical and curricular needs, and the contextual learning goals in their individual classrooms or courses. The authoring module can also be used to adapt or expand existing problem sets to meet the needs of different students, teaching styles or instruction contexts. At a minimum, each IMMEX problem set consists of a scenario and enough information to solve that scenario in multiple ways. Together, these define a problem space. By making small changes to the problem space, authors can easily create numerous similar but distinct versions (cases), each with different information and a different solution, in a matter of minutes. Educators design problem sets based on real world tasks that require their students apply curricular content to solve the problem. As such, scenarios are flexible enough to allow students the opportunity to approach problem solving in a way that makes sense to each individual student (Hurst, Casillas et al. 1997). The qualitative chemistry problem used in this study (called *Hazmat*) has 23 such cases and all cases share an identically structured problem space.

The IMMEX presentation module delivers *Hazmat* cases as a series of web pages via the worldwide web. Initially, IMMEX presents the student a statement of the problem they are required to solve. The teacher of the students involved in this study designed the scenario in each of the 23 *Hazmat* cases around a hypothetical earthquake which caused a number of chemicals, some of which were hazardous, to fall off the stockroom shelf. Since the labels are no longer with the spilled chemicals and time is of the essence, the school hired some of its chemistry students to identify the spilled chemicals. In each *Hazmat* case, the student then has access to the results of conducting up to 11 different physical or chemical tests on the unknown substance and to nine general reference items, which s)he can use to identify the unknown substance. As students form various hypotheses about a solution to the posed problem, they can access any of these items of information, in any order, to validate their hypotheses. The software presents the results from each test to the student graphically – as still or video images – audibly, or as text.

After the student has either successfully solved a case or has exhausted his or her available attempts at a solution (two for these students), IMMEX provides the student and teacher multiple performance measures. While some of these metrics measure problem solving efficiency or proficiency (for example, cases performed or % correctly solved), a more detailed description of student performance is available to assess understanding. The IMMEX assessment module provides a graphical representation of a student's performance in which each information item in the problem space is represented by a unique rectangle. As a student solves an IMMEX case, by moving from one information item to the next, the assessment module records each of these steps and it builds a unique graphical representation of the student's path to a solution. The resulting artifact is termed a Search Path Map (SPM). A student can access their SPM any time after finishing a case, and the student's teacher can view this map at anytime during or after the student performance. Initially, the SPM consists only of a "Start" box, but the assessment module adds a new rectangle to the map each time a student views a new piece of information. In addition, IMMEX also inserts a line between each rectangle indicating the order in which a student selected each piece of information and adds a graphical timeline to the SPM representing the relative amount of time the student spent viewing each item.

When one examines student search path maps, two characteristics quickly become evident. First, the student either solved or did not solve the problem. Second, a student either gathered enough information to solve the problem or they had viewed insufficient information with which to do so. In this research, we term a student that both had the necessary information and actually solved the problem to have demonstrated an *understanding* of the concepts required to solve the case. We used this dichotomous metric as one measure of student understanding. While the process will eventually be automated, for this study researchers used this metric to manually score the SPM generated by each of the students during their end-of-chapter performance assessment. All 134 search path maps were scored in less than one hour.

For this study, each student also completed written worksheets as they solved each *Hazmat* case. The worksheets were not a roadmap through the problem space, but were designed to allow students to record the results of each test they conducted, their interpretation of the information the test provided, and their logic at arriving at an answer. Prior to the study, the students' teacher had also created a rubric to score the degree of student *understanding* demonstrated on each worksheet. Ultimately, the teacher assigned each worksheet a grade of from one to five points according to the following scale:

- Five points: A performance explicitly shows the logic behind the student's correct answer. The student clearly had eliminated all other possible answers. Full understanding.
- Four points: A performance shows the logic behind the student's answer, but the logic required the student to guess between at least two possible answers.
- Three points: A performance gave the information necessary to solve the problem, but the student appeared unable to interpret that information and arrive at a conclusion.
- Two points: A performance in which the student began a series of tests (i.e. identified part of the unknown compound) or made only one or two initial observations.
- One point: A performance that was apparently random or showed little logical direction.

This rubric served as a second measure of student understanding since it required students to provide a written explanation of their activity. Consequently, we hypothesized a strong and significant correlation would exist between the researchers' evaluation of student understanding and the teacher's evaluation of each student's understanding using her rubric.

A possible limitation of this study was that most students only performed one case under examination conditions so we were concerned that the performances might not have been representative of the more global dimension of student understanding. Consequently, we used the demonstrated pattern recognition capabilities of artificial neural networks (see, for example, Principe, Euliano et al. (2000)), to cluster the performances of these students into groups that represented similar problem solving strategies. These neural networks cluster performances into similar groups based on the items a student selected. Accordingly, each group represents a particular item selection *strategy* which describes how the students within that group solved *Hazmat* cases. We have published details of this procedure and of its validity elsewhere (Stevens and Najafi 1993; Casillas, Clyman et al. 2000; Vendlinski and Stevens 2000; Vendlinski 2001). The resulting strategies often differ in at least two very important ways. First, different strategies often rely on different amounts of information. For example, one common guessing strategy is to immediately guess at the solution to an IMMEX case. Obviously, this strategy needs no menu items to implement. Conversely, another common strategy used by students is to view each and every available menu item before solving. Accordingly, these strategies fall into two distinct groups. Another characteristic that distinguishes various strategies is their effectiveness. Some strategies consistently produce high solve rates, while others seldom result in the correct identification of an unknown. A distinct pattern emerges when the amount of information used by each strategy is plotted against the strategy's solve rate. Students who access few or no information items seldom solve the case they are working on. Similarly, students with unfocused searches often view large amounts of information but, like their more penurious peers, seldom correctly solve the problem. On the other hand, students using more focused strategies view only enough information to reach a single, logical and, more often than not, correct conclusion. Individual strategies, therefore, may be further classified into more general strategy types termed "Limited," "Prolific," and "Efficient," respectively.

Results

As suspected, the overall GPA of the students in this study correlated strongly and significantly with both first semester ($r = .673$) and second semester ($r = .757$) chemistry grade. Furthermore, the correlation between the two semester chemistry grades were just as strong and significant ($r = .747$). All three correlation coefficients were significant at the $p < .001$ level. All other math and science grades were similarly correlated suggesting to us that grades consistently measure the same attributes of these students. Based on the literature and for the reasons we now address, however, we do not believe that these attributes fully assess how well students *understand* the concepts they are taught.

On average, the students in this study scored at the 60th percentile in science, at the 67th percentile in math, and near the 56th percentile on the reading portion of the SAT-9. All three tests showed a strong and significant correlation with GPA and first semester chemistry grade, but the SAT-9 science test was the only test that did not significantly correlate with second semester chemistry grade. While a significant correlation between student grade measures and each of these tests was expected, that the largest correlation would be between grades and SAT-9 reading was not. In fact, the students' SAT-9 reading score was correlated more strongly than either of the other SAT-9 tests with almost every math and science grade the students had received in high school. Moreover, SAT-9 reading scores showed a strong and significant correlation with ethnicity (_^{2} = 20.83; df = 3; p < .001). White students overwhelmingly scored above the mean on this test, while non-white students overwhelmingly scored below the mean. These results suggest to us that the largest variability in a student's grade might result from their English language literacy rather than the content knowledge of the course a student is studying.

Because both of the understanding metrics described previously were arguably measuring the construct of understanding, we expected them to produce highly correlated evaluations of this trait in each student, even though the methods used to arrive at such a conclusion were very different. Moreover, we expected that neither would share the same degree of correlation with the traits measured by grades and standardized testing.

As Table 1 illustrates, the two metrics seem to be measuring the same construct.

		Teacher Classification				
		Rubric Score "1"	Rubric Score "2"	Rubric Score "3"	Rubric Score "4"	Rubric Score "5"
IMMEX Classification	Understanding	0	1	2	9	37
	Did not fully understand	1	23	16	18	8

Table 1. This table is a cross-tabulation of the score a student received on the notes s)he made to support his or her answer on the *Hazmat* assessment problem versus how the same student's performance was classified using their IMMEX performance data alone. The chi-square statistic (_^{2} = 52.38; df = 4; p < .001) suggests that the distribution did not occur by random chance.

The results in Table 1 suggest there exists a significant statistical relationship between students classified as understanding based on their IMMEX performance and those similarly classified by their teacher's rubric. The two metrics agreed on almost 83% of the student performances (95 out of 115). Both measures agreed that thirty-seven students had demonstrated understanding, and that the performances of fifty-eight students demonstrated less than full understanding (i.e. a rubric score of between 1 and 4). In fact, the actual data suggests the correlation may be even stronger than is obvious from a quick review of Table 1. In half of the eight performances that received a "5" on the teacher's rubric but were classified "Did not fully understand" based on their IMMEX performance, the student never actually attempted the problem recorded on her or his worksheet. Rather, the student merely made up a non-existent performance on paper so s)he would not have to actually work the case that was presented by the computer. In every instance, this "invented" performance replicated a successful practice performance by that student from earlier in the semester. While the large number of student performances made these invented performances almost impossible for the teacher to detect without physically matching the paper and computer records, these counterfeits were easily uncovered using the computerized tool. Furthermore, the 12 performances classified as demonstrating understanding based on the performance recorded by IMMEX, but given scores of less than "5" by the teacher usually indicated the student had actually reviewed more information in IMMEX than s)he had actually recorded on his or her worksheet. Were reclassifications of the data made based on these facts, the two classifications of student *understanding* are in agreement almost 97% of the time. Moreover, the computer made detecting false performances simple and ensured that the teacher could see all the steps a student used to arrive at an answer, whether or not the student considered such a step important enough to report. Finally, the computer technology allowed such an evaluation to occur in a small fraction of the time the teacher would have required to physically set up such an experiment or to evaluate each performance using a rubric.

In addition to confirming that these two metrics are highly correlated and so are likely to be measuring the same construct, it is important to ensure that the measures do not correlate to unrelated traits (Cronbach 1989). Table 2 suggests that the two measure of understanding have a much greater correlation with one another than either metric has with other typical classroom evaluations such as a forced choice chapter test, or the student's self reported frequency of guessing at a solution.

	IMMEX Understanding	Teacher Rubric	Unit 12 Test	Student Self-evaluation
Teacher Rubric	.638			
Unit 12 Test	.295	.397		
Student Self-evaluation	.303	.375	.273	
Student self-reported guessing	-.302	Not Significant	-.256	-.374

Table 2. These correlations suggest that, in fact, the IMMEX and teacher measures of understanding demonstrate the highest degree of correlation, and that neither measure correlates as well with metrics measuring content or guessing. All coefficients are significant at the p < .01 level.

While the metrics called "IMMEX Understanding" and "Teacher Rubric" seem to be measuring the same construct, neither seems to be merely another measure of a construct like content knowledge or guessing. The correlation coefficients reported in Table 3 suggest that the same is true for these two understanding metrics and high-stakes content tests such as the SAT-9. Here again, correlation coefficients less than .638 suggest that neither metric of understanding is actually measuring only the content knowledge reputedly tested by grades and high-stakes tests.

	IMMEX Understanding	Teacher Rubric	Unit 12 Test	Student Self-evaluation	Student self-reported guessing
GPA	.283	.368	.545	.248	-.350
1st Semester Chemistry	.205*	.358	.637	.281	-.249
2nd Semester Chemistry	.246	.437	.580	.216*	-.245
SAT – 9 Reading	Not Significant	Not Significant	.456	Not Significant	Not Significant
SAT – 9 Math	Not Significant	.236	.352	.325	Not Significant
SAT – 9 Science	Not Significant	Not Significant	.437	.293	Not Significant

Table 3. This table shows the correlation coefficients between the two measures of understanding and student self-evaluation, student self-reported guessing, and other pencil and paper evaluations. Pencil and paper evaluations demonstrate the greatest correlations with measures of content "knowledge." The measures of understanding, however, show a much lower correlation with content measures. All coefficients are significant at $p < .01$, unless noted. Coefficients with an asterisk are significant at $.01 < p <.05$.

Furthermore, the lack of correlation between these measures of understanding and SAT-9 reading suggests that, unlike grade measures (see Table 4), neither measure is biased by a student's English language literacy. In fact these results suggest that these other measures are probably measuring something other than what we have termed understanding. Just as important, neither construct of understanding proposed here is significantly correlated with a student's gender, ethnicity or SES.

	GPA	1st Semester Chemistry	2nd Semester Chemistry	SAT – 9 Reading	SAT – 9 Math
1st Semester Chemistry	.673				
2nd Semester Chemistry	.757	.747			
SAT – 9 Reading	.465	.414	.311		
SAT – 9 Math	.288	.356	.237	.336	
SAT – 9 Science	.291	.363	Not Significant	.642	.540

Table 4. This table shows the large inter-correlation coefficients between grades and between part of the Stanford Achievement Test, 9th Edition (SAT – 9). These values suggest they each are measuring similar constructs. Furthermore, the moderately large correlations between grades and SAT – 9 Reading suggests that English language literacy may play an important role in grade evaluations. All coefficients are significant at $p < .01$.

To determine if the student's examination performance was representative of their overall performance, we conducted a correlation analysis comparing the student's typical practice strategy type with the strategy type a student used during their exam.

		Strategy type student used on assessment		
		Limited	Efficient	Prolific
Strategy type student used most often	Limited	12 : 5	7 : 11	13 : 17
	Efficient	4 : 4	21 : 10	4 : 15
	Prolific	1 : 8	9 : 17	41 : 26

Table 5. A crosstabulation of the strategy type a student used most often when solving *Hazmat* cases and the type of strategy a student used to solve the *Hazmat* case given on the assessment. The numbers in each cell represent the observed value : the value expected if the distribution was entirely random. The chi-square statistic (_^{2} = 51.8; df = 4; ; < .001) suggests that the distribution did not occur by random chance and the larger than expected values on the left to right downward diagonal suggest that students are very likely to use the same type of strategy to do assessment problems as they ordinarily use to solve other cases.

As shown in Table 5, this analysis suggests that the examination performance was, indeed, representative of the students overall performance. In fact, without pedagogical intervention, the students in this study repeatedly used the same type of strategy to solve *Hazmat* cases, even when that type of strategy was consistently ineffective. These results suggest that practice cases were, in themselves, indicators of the degree of student understanding during the course of instruction, that we can identify when a student is having conceptual difficulty in a course, and that we may be able to posit why they are having difficulty. Unfortunately, because manually rating each class of student note pages required such a large commitment of the teacher's time, it was impossible to rate each student's practice assessment and then correlate teacher and IMMEX assessments of student understanding for these performances.

Conclusion

The significant correlation between a standard measure of student *understanding* (such as a teacher using a rubric to assess written, open-ended student answers) and the automated measure (in this case the information a student used to arrive at and the correctness of an answer) suggest that this technology may allow educators to accurately gauge student understanding in near-real time. Furthermore, the low correlation between the IMMEX measure and other content and guessing metrics suggest the former and latter metrics are not simply measuring the same traits.

The results reported here suggest we can discern how well a student's actions explain their thinking, how they use facts to support an explanation, and how they apply that explanation to situations that are similar, but not identical, to situations they have practiced before. In short these findings suggest we can use the data produced by such technologies to infer if a student *understands* the content we are teaching and assessing. Moreover, these findings suggest that we now have the tools to finally begin addressing

Bloom's (1956) lament that the educational system overvalues and assesses knowledge while ignoring the more important aspects of education. Most importantly, this technology allows schools to embed the assessments within the normal curriculum, adapt their content to the curriculum students are actually exposed to, and to reduce the amount of time necessary to conduct and report assessment results.

References

Alberts, B. (2002). Appropriate Assessments for Reinvigorating Science Education. http://glef.org, George Lucas Educational Foundation.

Arter, J. and J. McTighe (2001). Scoring rubrics in the classroom. Thousand Oaks CA, Corwin Press.

Barton, P. E. (1999). Too much testing of the wrong kind; too little of the right kind in K-12 education. Princeton, NJ, Educational Testing Service.

Bloom, B. S., Ed. (1956). Taxonomy of educational objectives. New York, David McKay Co.

Busching, B. (1998). Grading inquiry projects. Changing the way we grade student performance: classroom assessment and the new learning paradigm. R. S. Anderson and B. W. Speck. San Francisco, Jossey-Bass. 74: 89 - 96.

Casillas, A. M., S. Clyman, et al. (2000). "Exploring alternative models of complex patient management with artifical neural networks." Advances in health sciences education 5(1): 23 - 41.

Codding, J. B. and R. Rothman (1999). Just passing through: the life of an American high school. The new American high school. D. D. Marsh and J. B. Codding. Thousand Oaks CA, Corwin Press: 3 - 17.

Corbett, H. D. and B. L. Wilson (1991). Testing, reform, and rebellion. Norwood NJ, Ablex.

Cronbach, L. J. (1989). Test validation. Educational measurement. R. L. Linn. Washington D.C., American Council on Education: 443-507.

Crouse, J. and D. Trusheim (1988). The Case Against the SAT. Chicago, University Of Chicago Press.

Elstein, A. S. (1993). "Beyond multiple-choice questions and essays: the need for a new way to asses clincal competence." Academic medicine 68: 244 - 249.

Fischer, K. W. and T. R. Bidell (1997). Dynamic development of psychological structures in action and thought. The Handbook of child psychology: theoretical models of human development. Lerner. New York, Wiley. 1: 467 - 561.

Furger, R. (2002). Assessments for understanding. George Lucas Educational Foundation: www.glef.org.

Gabel, D. L. and D. M. Bunce (1994). Research on problem solving: chemistry. Handbook of research on science teaching and learning. D. L. Gabel. New York NY, Macmillan Publishing: 301 - 326.

Gardner, H. (1993). Multiple intelligences: the theory in practice. New York, Basic Books.

Goertz, M. and M. Duffy (2001). Assessment and acountability across the 50 states. http://www.cpre.org/Publications/Publications_Policy_Briefs.htm, Consortium for policy research in education: 8.

Harcourt Educational Measurement (2001). You can count on Stanford 9!, Harcourt, Inc. www.hemweb.com/trophy/achvtest/sat9view.htm.

Harley, K. and L. D. Bendixen (2001). "Educational research in the Internet Age: examining the role of individual characteristics." Educational Researcher 30(9): 22 - 26.

Hurst, K., A. Casillas, et al. (1997). Exploring the dynamics of complex problem-solving with artificial neural network-based assessment systems. Los Angeles, UCLA.

Libarkin, J. C. and R. Mencke (2002). "Students teaching students." Journal of college sceince teaching 31(4): 235 - 239.

Lowyck, J. and J. Poysa (2001). "Design of collaborative learning environments." Computers in human behavior 17: 507-516.

Messick, S. (1989). Validity. Educational Measurement. R. L. Linn. New York, Macmillan Publishing Co.: 13 - 103.

Mestre, J. P. (1991). "Learning and instruction in pre-college physical science." Physics today: 56 - 62.

Miller, G. A. (1956). "The magical number seven, plus or minus two: some limits on our capacity for processing information." The psychological review 63: 81 - 97.

NCES (1996). Data from the National Assessment of Educational Progress (NEAP). Washington, D.C., National Center for Educational Statistics.

NRC (1996). National science education standards. Washington D.C., National Academy Press.

NRC (1999). Transforming undergraduate education in science, mathematics, engineering, and technology. Washington DC, National Academy Press.

Olson, L. (2002). Grade by grade testing policies. Education week on the web. 21: 26.

Pellegrino, J., N. Chudowsky, et al., Eds. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington DC, National Academy Press.

Principe, J. C., N. R. Euliano, et al. (2000). Neural and Adaptive Systems. New York, John Wiley & Sons.

Quellmalz, E., P. Schank, et al. (1999). "Performance Assessment Links in Science." Practical Assessment, Research & Evaluation 6(10).

Raizen, S. A. (1990). Assessment in science education. The prices of secrecy. J. L. Schwartz and K. A. Viator. Boston MA, Harvard Graduate School of Education: 57 - 68.

Resnick, L. B. (1987). Education and learning to think. Washington, D.C., National Academy Press.

Resnick, L. B. and D. P. Resnick (1992). Assessing the thinking curiculum: new tools for educational reform. Changing assessments: alternative views of aptitude, achievement and instruction. B. G. Gifford and M. C. O'Conner. Boston, Kluwer Academic Publishers: 37 - 75.

Rudd, J. A., T. J. Greenbowe, et al. (2002). "Recreafting the general chemistry laboratory report." Journal of college sceince teaching 31(4): 230 - 234.

Rutherford, F. J. and A. Ahlgren (1990). Science for all Americans. Oxford, Oxford University Press.

Shepard, L. (1995). Effects of introducing classroom performance assessments on student learning. Los Angeles, UCLA/CRESST.

Stevens, R. H. and K. Najafi (1993). "Artificial neural networks as adjuncts for assessing medical students' problem solving performances on computer-based simulations." Computers and biomedical research 26(2): 172 - 187.

Stiggins, R. J., E. Rubel, et al. (1988). Measuring thinking skills in the classroom. Washington DC, National Education Association.

Tucker, M. S. (1999). How did we get her, and where should we be going. The new American high school. D. D. Marsh and J. B. Codding. Thousand Oaks CA, Corwin Press: 18 - 34.

USDoEd (1983). A nation at risk: the imperative for educational reform. Washington, D.C., U.S. Department of Education.

USDoEd (1993). Goals 2000: Educate America. Washington, D.C., U.S. Department of Education.

Vendlinski, T. and R. Stevens (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. International Conference of the Learning Sciences, University of Michigan, Lawrence Erlbaum Associates.

Vendlinski, T. P. (2001). Affecting U.S. education through assessment: new tools to discover student understanding. Technology Management and Policy. Cambridge MA, Massachusetts Institute of Technology: 183.

Viadero, D. (2000). Students in dire need of good teachers often get the least qualified or less experienced. Education week on the web: www.edweek.org/ew/ewstory.cfm.

Wiggins, G. P. (1993). Assessing Student Performance: exploring the purpose and limits of testing. San Francisco, Jossey-Bass.