INTERACTIVE SCENARIO-BASED ASSESSMENT SYSTEM (ISBAS): PILOT STUDY

Seth Mayotte
University of Minnesota
Department of Educational Psychology
Minneapolis, MN / U.S.A.
mayotte@stat.umn.edu

Abstract

We have moved beyond a time when simple multiple choice exams are sophisticated enough to assess the complex knowledge we expect of students in today's world. At the same time, we have moved beyond the scale at which individualized assessments, like oral, essay, or performance exams, are efficient enough to be utilized. We must develop an efficient, objective assessment method which is complex and robust enough to assess knowledge and skills necessary to succeed in our modern world.

This design proposal for an interactive scenario-based assessment system (ISBAS) aims to address these concerns. The system is designed to allow students to troubleshoot complex scenarios, ask diagnostic questions, and make diagnoses all through an interactive web-based interface. Students receive feedback each step of the way, aiding them in navigating through a decision tree to solve the problem at hand. The choices that the student makes are automatically scored based on predetermined, objective values used to assign a grade or proficiency level. This proposed system is efficient, flexible, objective, and able to encompass the complexities of complex knowledge and skills.

A functioning demonstration of the proposed system will be presented along with usability and preliminary pilot testing results.

STATEMENT OF PROBLEM

George Madaus, of the National Board of Educational Testing and Public Policy describes three basic ways a school can measure academic achievement:

- select an answer from several options (multiple-choice, matching, etc.)
- produce an answer in essay form
- ask students to do something (fix something, perform, present)

(2000, as cited by Nathan & Johnson, 2000, p.25-26)

Unfortunately, each of these methods has a series of inherent problems. Select an answer test formats are used due to the efficiency of automated scoring. They may be presented in paper formats, including handwritten and Scranton, or in computerized formats, both in person and online. All of these forms and formats are subject to the same limitations, namely test-wise students, short-term memory retention (e.g., cramming for the test), and superficial knowledge assessment (i.e., lack of complexity).

Essay exams allow students do demonstrate a greater level of complexity then select an answer tests, but they also suffer two major limitations. Grading essays is extremely time-consuming and however qualified the grader is, a great deal of subjectivity in grading is unavoidable.

Finally, traditional Authentic Assessment (asking a student to do something) again allows a student to demonstrate competence in a highly tangible manner, yet it suffers similar limitations to essay exams. The setup, administration, and scoring of the assessments is extremely cost prohibitive in many cases. And however qualified the observers/assessors are, a limited number of human observers introduces a great deal of subjectivity into the scoring.

This pilot study is designed to evaluate the effectiveness of an alternative interactive scenario based assessment system (ISBAS). ISBAS addresses the limitations of all three of the previously discussed assessment methods. It is designed to allow students to troubleshoot complex scenarios, ask diagnostic questions, and make diagnoses. All the while, the choices that the student makes are tracked and scored based on predetermined, objective values. A final analysis of the student's choices and decisions are used to assign a grade or proficiency level and provide feedback to the student and instructor. ISBAS is efficient, flexible,

objective, and able to encompass the complexities of procedural knowledge and skills.

The overarching research question is: Is ISBAS *more* effective in assessing complex knowledge and skills in students than any of the three traditional methods of school assessment? This is a complex question that will take a great deal of research to address. This pilot study is designed to be the first step in that research. Its purpose is to address the following needs:

1. Demonstrate the implementation of ISBAS
2. Establish concurrent validity with traditional assessment methods
3. Collect student attitudes about using ISBAS
4. Identify future research priorities

REVIEW OF LITERATURE

What is Assessment?

An assessment instrument in an educational context serves as one piece of evidence to support the evaluation question: how do we know when we have gotten our learners to where we want them to be (Smith & Ragan, 1999, p. 92)? The answer to the question "where we want them to be" follows directly from the learning objectives laid out in the curriculum design process. It is predicated on the assumption that learners have effectively encoded information in the brain and have a retrieval method that allows them to access the information when appropriate.

In designing effective curriculum, we must consider in what ways the information is encoded and in what ways the learner will be asked to retrieve the information, both in terms of assessment and for later use in life, work, etc. Encoding Specificity Theory (Tulving & Thompson, 1973) indicates that the environment in which information is encoded should be similar to the environment in which it is expected to be retrieved to better facilitate retrieval. Asking a learner to take a pencil and paper test (of any sort) about how to adjust a carburetor on an automobile would not lead to effective retrieval.

Furthermore, the Spread of Activation network memory model (Warren, 1977) would suggest that priming the learner prior to or during the assessment with cues that activate similar memory nodes would lead to better retrieval. Contrary to the belief that more information "gives away the answer," this priming better sets a more realistic context for retrieval. For example, asking a medical student to answer a multiple choice question about which symptoms are indicative of disease X is not realistic to the retrieval with which the student will be likely to encounter in an employment setting. Asking the student, on the other hand, to respond to a patient that comes in with a description of presiding symptoms and durations with a potential diagnosis would activate those memory nodes appropriate to the retrieval.

Having a series of assessments that build on each other or themselves also provides cognitive scaffolding essential for challenging the limits of a learner's ability (Vygotsky). For example, an assessment which iteratively goes into greater and greater detail or specificity in response to the learners' decisions would be ideal. It would probe the limits or scope of the skill set or knowledgebase.

The Heuristic Systematic Model (HSM) proposed by Chaiken (1993) illustrates two different approaches to cognition. The heuristic approach is automatic, requires little cognitive effort and focuses on salient cues. Systematic processing is more carefully thoughtout. It requires attention, reasoning, and a great deal of mental effort, along with capacity and motivation. Both methods are used in problem solving at different times. Take for example, an auto mechanic's certification assessment. One simple way to set up an assessment would be to bring in a car with a problem and give them mechanic-trainee a chance to repair it. But what if the mechanic simply gets luck and guesses this correct problem on his first attempt. Does that illustrate competence? What if, on the other hand, the mechanic takes a very systematic approach, is very thorough about working through the problem step-by-step, but forgets one minor detail. Does this illustrate incompetence? An assessment must be able to capture both aspects.

What is Assessment Used For?

The education world generally breaks down the purposes of assessment down into two categories based on the intended use. Norm-referenced assessments like the Graduate Record Exam (GRE) are designed to compare or rank learners' abilities. Norm referenced assessments do not provide much in the form of useful information about an individual learners' competence in a particular skill or knowledge.

Criterion-referenced assessments on the other hand are designed to determine levels of competence. They can also be used to identify "where individuals' weaknesses are" to target future instruction and learning activities (Smith & Ragan, 1999, p.93). These assessments are ideal for formative assessment.

Unfortunately, some educational systems use the wrong types of assessments in the wrong settings. Recently the emphasis on school accountability in the No Child Left Behind laws and on student accountability in graduation standards exams both rely heavily on summative, norm-referenced assessments to rank schools and students. These high-stakes goals would be better suited by early administration of criterion-referenced formative assessments designed to aid in altering curriculum to meet the needs of the learners. Curriculum is hardly standardized between schools, districts, and states, so it is meaningless to rank order with standardized summative

assessments. The goal here should not be to rank schools and punish or reward based on rank, it should be to identify areas which need improvement and successful strategies to meet those needs.

Traditional Educational Implementations of Assessment

The Socratic Method could be considered one of the first widespread assessment methods used in Western society. This method is characterized by "the use of questions…to develop a latent idea in the mind of a student or elicit an admission from an opponent." (Random House Webster's College Dictionary, 1992, p.1271) This method evolved into what we now categorize as oral exams. Oral exams used to be the mainstay in educational assessment, but have fallen out of use in many contexts in the last century. The main problems associated with oral exams are their subjective nature and the time required to implement them. Oral exams are still fairly common in non-Western educational systems, but are rarely found in the west, with the exception of the dissertation defense. Even in the case of dissertation defenses, multiple raters are essential to reducing subjectivity or rater bias.

About a century ago, standardized, Multiple-Choice testing gained large scale popularity due to its efficiency and objectivity. "Critics of standardized tests are quick to argue that such instruments place too much emphasis on factual knowledge and on the application of procedures to solve well-structured, decontextualized problems. Pleas for higher order thinking skills are plentiful." (Linn, Baker & Dunbar, 1991, p.19) Bloom (1956) laid out his famous taxonomy of objectives in the cognitive domain, consisting of six types: recall, comprehension, application, analysis, synthesis, and evaluation. And while critics of Bloom argue that there may be as few as two dimensions, all seem to agree that there are distinctions in the complexity of knowledge and skills that can be assessed in a learner (see Merrill, 1983, Gagne, 1985).

The critiques of standardized test models essentially fall into two categories: the inherent limits of the assessment model and the lack of quality control in individual item or test design. Hence an overview of recommended assessment design features is warranted. In "A Practical Guide to Assessment and Accountability in Schools," Nathan and Johnson (2000, p.13) describe the importance of laying out measurable goals: "Not every important goal can be measured easily. But unless a

goal can be stated explicitly, it is very difficult to know whether a school and its students are making progress – and if so, how much." Too often, assessments are designed without first explicitly laying out goals, outcomes, and objectives (sadly, the same is true of much instruction).

In addition, we must consider how often test item writers fall into patterns that test-savvy test takers can pick up on (e.g., longest answer, all/none of the above are more often correct and can be detected easily). Even if we set these arguments aside and assume that our standardized test was created based on the best standards, we must consider the level of complexity or depth of knowledge that the test is capable of measuring. Would you want a medical school to rely solely on standardized tests to assess surgical skills? Of course not. This is the reason medical residencies exist. This is also the primary reason for the recent resurgence in "Authentic Assessments." One must keep in mind that most forms of traditional authentic assessments predate standardized testing (oral exams, oral presentations, performances, essays, etc.).

Peterson and Neill go on to describe the main criticisms of our traditional authentic assessment category, which they dub "Alternative Assessments." Critics argue that these alternatives are not trusted as accurate representations of student knowledge or skills because they cannot be "statistically and 'objectively' determined and analyzed." They attribute this lack of trust to a societies "predominant approach to thinking and learning" and point out that historically, the consequences of this approach can be dangerous (1999, p.2). The implication here is that "objectivity" is a loaded term. We think of it as being the gold standard we should all adhere to, but the flip side of the coin is that objective measures are only as objective as the creators of the measures and that once we layout the framework of "objectivity" we stifle any alternate interpretations of the problem.

Another common criticism of alternative assessments is that such assessments generally cost more to administer and score in either time or money. Finally, a common criticism of traditional authentic assessments is that the scoring is subjective and does not take into account the needs of minority or ESL students. "Clearly, this is a serious issue. At the same time, it is a problem that pervades all forms of assessment. Who, for example, chooses the questions on standardized tests?" (Peterson and Neill, 1999, p.4)

The educational system has thus entered this vicious cycle of assessment. Problems of subjectivity and expense lead to more reliance on standardized tests. Problems with depth of knowledge and decontextualization lead to more reliance on authentic assessments. This puts us right back where we started.

The Vicious Cycle of Assessment

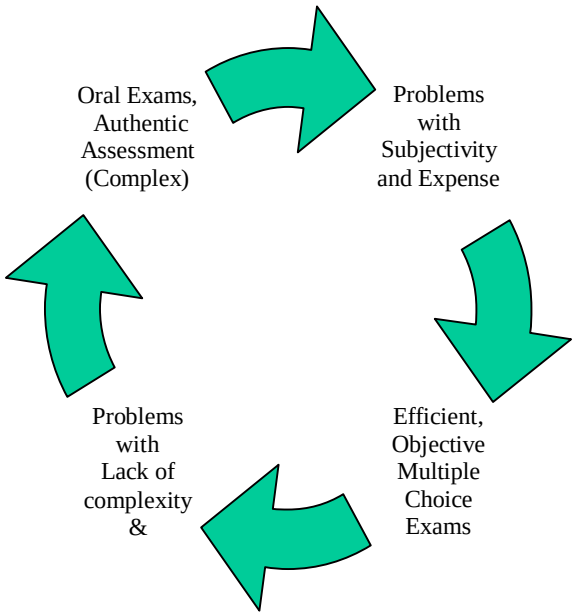


Figure 1

Improving Assessment Development

A great deal of effort has gone into improving the quality of standardized tests and the efficiency of traditional authentic assessments. Educational Testing Service (ETS) serves as an example of both, with their continuing refinement of the Multiple Choice (MC) components of the SAT and GRE exams and their exploration of automated, computerized scoring of essay components.

Can standardized assessments be written to address higher levels of complexity or higher-order thinking? Yes. See Appendix A for examples of this. Can authentic assessment be scored in an objective manner? Yes. The use of rubrics, multiple scorers, inter-rater reliability and training can all be used to increase the objectivity. Both of these problems can be solved by carefully designed assessment instruments. The problem is that creating such carefully designed assessments is difficult and time consuming.

Teachers often lack the training to design quality assessment instruments. Even if they are well trained, they often lack the time. This does not however, stop

teachers from designing assessments. "Stigging and Bridgeford found that the use of teacher-made objective tests increased between 2nd and 11th grade." (1985, as cited by Rodreguez, 2004, p.3) Other teacher rely on textbook provided assessments, written by textbook authors or their reviewers. In short, most assessments are designed in a hurry by people who may not be very well qualified. Good criterion-referenced assessments are often described as having "The Five C's: congruence, completeness, consistency, confidence, and cost." (Smith & Ragan, 1999, p. 95)

Designing good quality test items is a difficult and time consuming task that is often beyond the scope of training of those who create tests (i.e., teachers, assessment coordinators, etc.). How often do teachers simply rely on pre-made standardized tests that come from questionable sources (i.e., textbook publishers)?

We must also consider the development or design costs of the assessment. If we consider any large-scale standardized assessment measure, such as the ETS's SAT, it is clear that vast amounts of time and money went into test writing, expert review panels, piloting, testing validity and reliability, and maintaining test integrity (e.g., finding, removing, and replacing compromised test items). Peterson and Neill summarize it well by stating "decent assessment can't be done cheaply, any more than can decent education." (1999, p.3)

New Models of Assessment

As our world becomes more and more complex, the nature of the skills and knowledge we attempt to assess also becomes increasingly so. Simulations may be better able to capture these complexities. Simulations have been successfully used as instructional tools in many different settings. For example, instructional simulation activities have been shown to improve students' statistical reasoning skills (delMas, Garfield & Chance, 1999, p. 1). Computer simulations are often used in the teaching of new technology. Despite these promising uses, few examples of simulations being used specifically for assessment exist.

"The challenges in the development of innovative testing methods lie primarily in the scoring arena. Complex test stimuli result in complex responses, which require complex models to capture and appropriately combine information from the test to create a valid score." (Melnick, 1996, as cited by Mislevy, 2002, cite) Mislevy lays out a conceptual assessment framework consisting of a complex interaction between

student, evidence, and task models in "Making Sense of Data from Complex Assessments." Messick (1994, p.16) summarizes the key features of the model:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics.

Much recent research has gone into designing new forms of assessment that attempt to address the shortcomings of the different assessments presented thus far. One good example of this is Harvey, a cardiology patient simulator developed at the University of Miami, School of Medicine. It is essentially a highly complex and interactive talking Annie manikin (as used in CPR training). It is capable of physically simulating 27 cardiac conditions (Issenberg et al, 1999). It has been successfully integrated into the curriculum and used for assessment purposes in many medical schools in the United States and in the United Kingdom (Issenberg et al, 2003).

The Department of Defense (DoD) and the US Air Force (USAF) have also been investigating the use of complex computer-based assessments. The DoD has investigated and contracted for the development of machine scoring of essays. Streeter et al (2003) describe the advantages of this type of system over human graders. In short they are: shorter time, more detailed, completely consistent, entirely objective, more complex analysis, free of reasoning and judgment errors. The USAF has been working on an even more ambitious project, developing a computer-based team performance assessment technology designed to assess the performance of military teams in a computer simulated design (Swezey, et al, 2000).

Problems to be Overcome with New Models of Assessment

One of the biggest initial problems with these new forms of assessment is the high cost involved in developing them. For many, it is difficult to calculate the total costs. Harvey, which has been brought to the market and adopted by many medical

schools, is being sold for $75,000 each. Developing and testing a new assessment system would cost even more.

Even more important though, are the obstacles to implementation. The DoD and USAF have a great deal of money to throw into assessing for military purposes. Schools, on the other hand, do not. Imagine the costs involved in training educators in customizing, administering, and scoring these new computer-dependent assessments. Furthermore, there is a tendency for teachers as a group to refuse to implement new technologies into their classrooms. This "culture of refusal," as described by Steven Hodas (1993), involves teachers viewing technology as a threat to either their way of teaching (or assessing) or their jobs altogether.

Finally, the way in which these new assessment systems are evaluated is by comparing the results with those of standardized tests and/or traditional authentic assessments. If these new models truly aim to more accurately measure these complex skills, then they should not correlate perfectly with these older models of assessment. Some theoretical explanation of the ways in which these assessments should differ is needed.

METHODOLOGY

Interactive scenario based assessment system (ISBAS) Overview

A computer simulation can be designed in such a way that it can solve all of these issues. This computer program would present a scenario to the learner. The learner would then have the option to ask a series of diagnostic questions or tests based on a keyword search. After receiving a response to each of the questions or tests, the learner must make a diagnosis and choose a course of action.

The learner will then be scored according to the appropriateness of the questions or tests asked for and the quality of the diagnosis. If the learner asks appropriate questions and comes to a succinct diagnosis, he/she will score high. If the questions or diagnosis are inappropriate, he/she will receive a lower score.

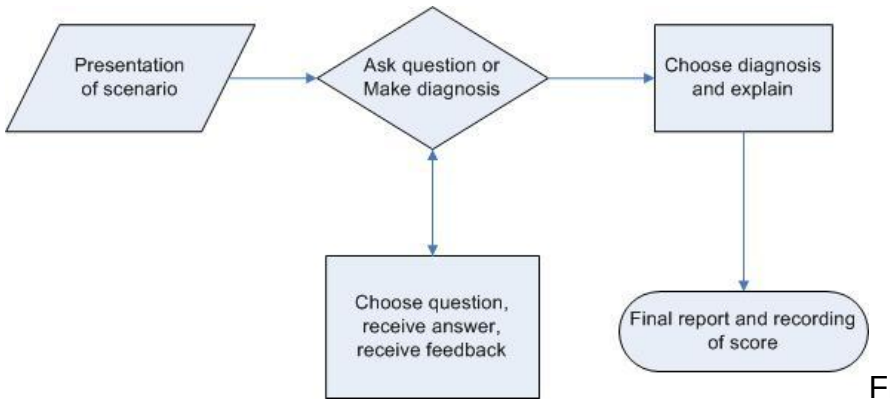


Figure 2

Take for instance, a driving simulator example. If used in a state driving proficiency exam, it could go beyond taking a nice drive around the block and parallel parking the car. A simulation could present learners with challenging driving conditions and evaluate how they respond. If they sped up to counter a tailgater, they would decrease their score; if they pulled over to allow for an emergency vehicle to pass, they would increase their score.

This proposed Interactive scenario based assessment system (ISBAS) has a number of advantages over the current alternatives. It is based on a complex scenario and it is not predictable, so it does not suffer from the deficiencies of the objective written assessments. Furthermore, it is scored by objective criteria in a computer environment, thereby making it efficient to administer and objective to score.

One of the additional goals of this proposal is to create a program that is easy for the layperson to use. The initial target audience for this program would be professional certification programs for the IT, medical, and other procedural skills based industries. However, the long term target audience for this program would be classroom teachers. Imagine if this system were designed simply enough so that a high school chemistry instructor or a middle school world cultures instructor could design procedural assessments for both formative and summative assessment practices in the classroom. Due to the efficiency of the system, more time could be devoted to classroom preparation and less time to assessment scoring. This system would be appropriate for any subject matter which already uses problem solving, critical incidents, or troubleshooting diagnostic procedures.

The initial system will be largely text-based in nature. Future versions may include graphics, photos, audio, video, and other multimedia components. Initially, the only input devices supported will be keyboard and mouse. Future versions may include more complex input devices. Content will be delivered and the assessment will be administered through a secure browser (SHTML) system.

The proposed system will use a simple spreadsheet based input file. Furthermore, an easy-to-use template will be available to simply fill in the values for Scenarios, Questions (appropriate or not), Answers, and Diagnoses with appropriate scoring weightings. The question that remains is who will decide these scoring criteria? Clearly, the experts in the field or the instructors assigned to the course would be the most appropriate assessors. Given that, the design will be simple enough for anyone who knows how to use a spreadsheet to design assessment scenarios.

Screen shots of ISBAS are included in Appendix B.

Piloting Plan

ISBAS will be piloted during the spring semester 2006 with students enrolled ITEC 1310 (Microcomputer System Maintenance) at Minneapolis Community and Technical College. This site and class was chosen because the researcher is the assigned faculty for this class.

Toward the end of the semester, students will be introduced to the ISBAS program by walking through a demonstration scenario with the entire class on a LCD projector. Students will then be asked to go through 4 scenarios related to their coursework as a review activity for the class. Students who choose not to participate will be given other printed materials from which to review course content. After completing 4 scenarios, students were asked to complete an online survey related to ISBAS (see Appendix C).

After the administration of this pilot test, the researcher will perform data analysis of the results in comparison to other forms of assessment generally administered to the course (multiple choice test items, essay questions, and performance-based assessments). All student names will be removed from the dataset; data will be organized by student ID number. The purpose of this analysis will be to establish concurrent validity with other forms of assessment.

Survey results will be analyzed to provide qualitative evidence about participant attitudes to usability and accuracy questions.

RESEARCH FINDINGS

10 students from ITEC 1310 at MCTC participated in the ISBAS pilot study on 4/19/2006. All of the students completed at least 4 scenarios. 8 of the 10 student took the survey afterwards.

Analysis of the resulting dataset was very challenging. At the heart of this challenge is the purpose of ISBAS, namely, to create a better assessment tool. If the ISBAS results were to correlate perfectly with traditional assessment methods, then ISBAS could only be thought of as being as good as, not better. What ISBAS is designed to measure is a slightly different construct then traditional assessments. Specifically, it measures a learner's ability to approach, troubleshoot, and resolve a complex scenario. This integrates problem solving skills and content knowledge.

The best measure of this construct, albeit a poor measure, is the cumulative total of all other assessments of the students' abilities throughout the rest of the course (labeled: assess_only). This measure explicitly removes many of the other factors that students are often graded on from it, including attendance, effort/participation, and growth/improvement. While these factors may be appropriate as evidence to support an evaluative grade, they are not the same as the ability construct. The measure used from ISBAS is the mean points received for the diagnoses chosen (labeled: diag_mean). The Pearson correlation between assess_only and diag_mean is r = .489 An ANOVA analysis shows a significant effect (F = 9.199, sig. = .048).

A much more robust linear relationship can be demonstrated if we control for scenario success. 3 of the possible 7 scenarios had significantly higher success rates than the others. Upon investigation, these 3 scenarios were also the most practiced and assessed of the options for this class. If we eliminate the other 4 scenarios from the data analysis, we get an ANOVA showing a relationship between assess_only and the trimmed diag_mean (F = 19.11, sig. = .008). r^2 = .867.

Finally, we also show a similar linear relationship between the diag-mean and total-mean and the practical assessment made in the class (labeled: pract). R^2 = .824; ANOVA (F = 11.726, sig. = .013).

Ideally the mean total points measure would have better captured the intricacies of the troubleshooting steps students took. This variable could not be used because of the limited nature of the pilot study. Students did not have an opportunity to practice with a system which rewarded or penalized them for the interactive steps they took and thus were prone to ask too many questions, thus skewing this variable. Furthermore, the program set the minimum total points to 0 and the maximum to a number arbitrarily chosen by the scenario writer. While these minimum and maximum limits can be justified for grading purposes, the raw numbers are necessary for this level of data analysis as they show the true range of all of a student's decisions throughout the simulation.

Qualitative analysis of the survey data suggest that the more difficult a student found learning and using ISBAS, the less accurate the student thought the assessment was. Conversely, the less difficult a student found learning and using ISBAS, the more accurate the student thought the assessment was. This illustrates a need to spend more time familiarizing students with the instrument before data collection or assessment occur.

The open-ended comments on the survey yield some interesting commonalities. Students commented on ISBAS being a better tool for [formative assessment] or self-assessment, rather than [summative assessment]. They also expressed concerns about how the keyword search was structured and many would have preferred a different method by which to search for questions. These comments will help with further refinement and implementation of the assessment instrument.

DISCUSSION

This pilot sample data is problematic for a number of reasons. As was clear from the survey data, students did not have enough exposure to the instrument to feel comfortable using it. They also indicated some technical problems and/or features that are needed to increase usability. The scope limitation making the mean total score variable unusable were also unfortunate.

There was some minor evidence to support concurrent validity, but in reality, concurrent validity will not answer the main question. Linn lays out a model for

evaluating the validity of alternative assessments as needing to include the following 8 evidences:

- regarding the intended and unintended consequences
- the degree to which performance or specific assessment tasks transfer
- fairness of the assessments
- cognitive complexity of the processes students employ in solving assessment problems
- meaningfulness of the problems for students and teachers
- content quality
- comprehensiveness of content coverage
- cost of the assessment

(Linn, Baker & Dunbar, 1991, p. 20)

Future research needs to address these (or other) more complex ways of evaluating validity. Refinement of the instrument and greater exposure to students prior to data collection are necessary. Eventually, it will be necessary to broaden the subject matter assessed by the instrument to demonstrate generalizability. The long term plan to accomplish this goal is to package the software in a portable format and distribute it to assessors in a number of settings. Free use of the software will be tied to providing data and survey information back to the researcher.

REFERENCES

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistics Education*, 1999, v.7, n.3.

Eagly, A. H. & Chaiken, S (1993). *The Psychology of Attitudes*. Learning, Belmont, CA: Wadsworth Group/Thompson.

Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher*, Dec. 1989, 27-32.

Hodas, S. (1993). Technology Refusal and the Organizational Culture of Schools. *Educational Policy Analysis Archives,* v1 n10 Sept. 14, 1993.

Issenberg, S. B., Pringle, S., Harden, R. M., Khogah, S. & Gorden, M. S. (2003). Adoption and integration of simulation-based learning technologies into the curriculum of a UK Undergraduate Education Programme. *Medical Education,* 2003; 37 (Suppl. 1):42-49.

Issenberg, S. B., McGaghie, W. C., Hart, I. R., Mayer, J. W., Felner, J. M., Petrusa, E. R., Waugh, R. A., Brown, D. D., Safford, R. R., Gessner, I. H., Gordon, D. L. & Ewy, G. A. (1999). Simulation Technology for Health Care Professional Sills Training and Assessment. *Journal of the American Medical Association,* Sept. 1, 1999, v282, n9, 861-866.

Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher*, Nov., 1991, 15-21.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessment. *Educational Researcher*, March 1994, 13-23.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G. & Johnson, L. (2002). Making Sense of Data From Complex Assessments. *Applied Measurement in Education,* 15(4), 363-389.

Nathan, J. & Johnson, N. (2000). What Should We Do? A Practical Guide to Assessment and Accountability in Schools. *Center for School Change, Hubert H. Humphrey Institute of Public Affairs, University of Minnesota.*

Peterson, B. & Neill, M. (1999). Alternative to Standardized Tests. Retrieved Nov. 28, 2005, from http://www.pde.state.pa.us/alt_disruptive/lib/alt_disruptive/ALTTEST.pdf

Random House (1992). Random House Webster's College Dictionary. New York, NY: Random House.

Rodriguez, M. C. (2004). The Role of Classroom Assessment in Student Performance on TIMSS. *Applied Measurement in Education*, 2004, 17(1), 1-24.

Smith, P. L., Ragan, T. J. (1999). Instructional Design, 2nd Edition. Upper Saddle River, New Jersey: Prentice-Hall, Inc..

Streeter, L., Psotka, J., Laham, D. & MacCuish, D (2003). The Credible Grading Machine: Automated Essay Scoring in the DoD. Retrieved Nov. 28, 2005, from http://www.k-a-t.com/papers/essayscoring.pdf

Swezey, R. W., Hutcheson, T. D., & Swezey, L. L. (2000). Development of a Second-Generation Computer-Based Team Performance Assessment Technology. *International Journal of Cognitive Ergonomics,* 2000, 4(2), 163-170.

Tulving, E. & Thompson, D. (1973). Encoding Specificity and retrieval process in episodic process. *Journal of Experimental Psychology*, 1973.

Warren, R. E. (1977). Time and the Spread of Activation in Memory. *Journal of Experimental Psychology: Human Learning and Memory*, 3, 4, 458-466, July 1977.

APPENDIX A – MULTIPLE CHOICE QUESTIONS

Multiple Choice questions can be written to address higher levels of complex thinking (or higher order thinking). Below are three examples and where they would be classified in Bloom's taxonomy.

MC example: Knowledge

Which one of the following persons is the author of "Das Kapital"?

a. Mannheim

b. Marx

c. Weber

d. Engels

e. Michels

MC example: Application

Which one of the following values approximates best to the volume of a sphere with radius 5m?

a. 2000m³

b 1000m³

c. 500m³

d 250m³

e. 125m³

MC example: Evaluation

Judge the sentence in italics according to the criteria given below: *"The United States took part in the Gulf War against Iraq BECAUSE of the lack of civil liberties imposed on the Kurds by Saddam Hussein's regime."*

a. The assertion and the reason are both correct, and the reason is valid.

b. The assertion and the reason are both correct, but the reason is invalid.

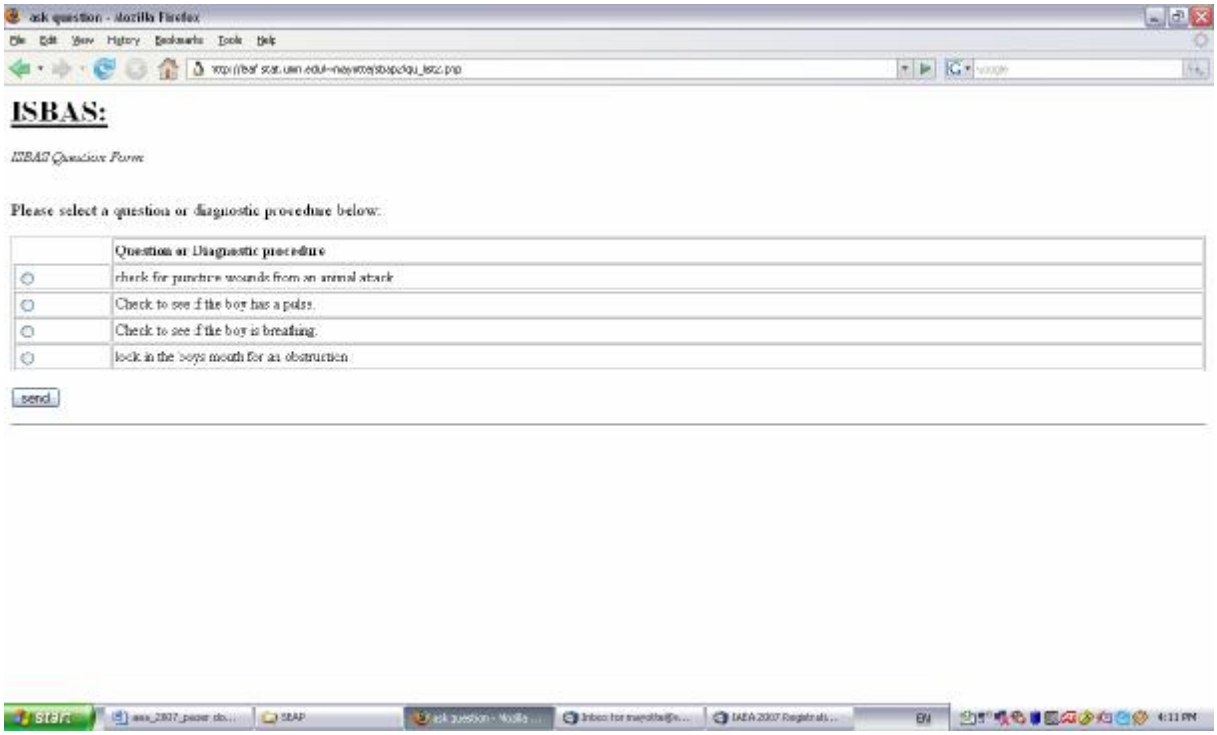
c The assertion is correct but the reason is incorrect.

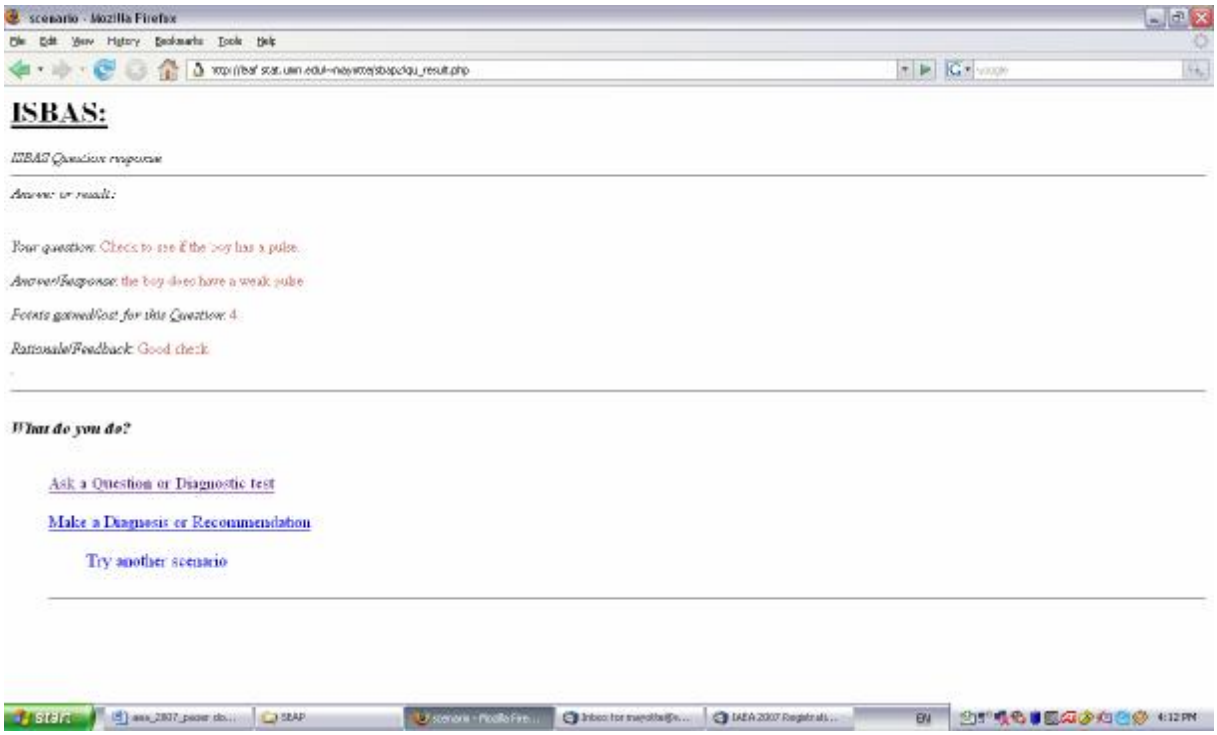
d. The assertion is incorrect but the reason is correct.

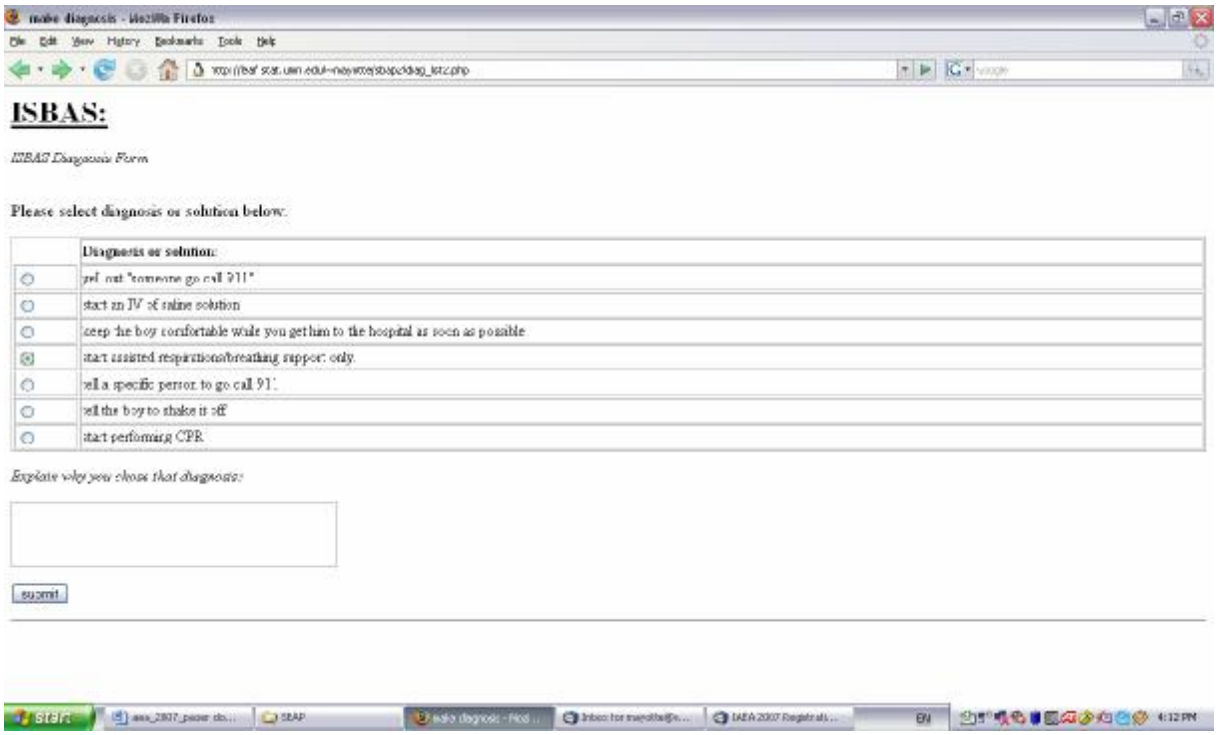
e. Both the assertion and the reason are incorrect.

Source: University of Cape Town, South Africa, Handbook on Designing and Managing Multiple Choice Questions,
http://web.uct.ac.za/projects/cbe/mcqman/mcqappc.html

APPENDIX B - SCREEN SHOTS OF ISBAS






ISBAS:
ISBAS Diagnosis Form
Please select diagnosis or solution below:
Diagnosis or solution:
yell out "someone go call 911"
start an IV of saline solution
keep the boy comfortable while you get him to the hospital as soon as possible
start assisted respirations/breathing support only
tell a specific person to go call 911
tell the boy to shake it off
start performing CPR
Explain why you chose that diagnosis:
submit

APPENDIX C - ONLINE SURVEY RELATED TO ISBAS

ISBAS:

ISBAS Survey Form

How accurate was ISBAS in assessing your abilities? (1=least accurate, 4=most accurate)

☐ 1 ☐ 2 ☐ 3 ☐ 4

How difficult was it to learn ISBAS? (1=least difficult, 4=most difficult)

☐ 1 ☐ 2 ☐ 3 ☐ 4

How difficult was it to use ISBAS? (1=least difficult, 4=most difficult)

☐ 1 ☐ 2 ☐ 3 ☐ 4

How useful was the feedback received at the end of each scenario? (1=least useful, 4=most useful)

☐ 1 ☐ 2 ☐ 3 ☐ 4

How comfortable would you be with ISBAS results being used as part of your course grade (don't worry; they will not be this semester)? (1=least comfortable, 4=most comfortable)

☐ 1 ☐ 2 ☐ 3 ☐ 4

How does ISBAS compare to Multiple Choice tests in accurately measuring your abilities?

- ☐ ISBAS is much worse
- ☐ ISBAS is worse
- ☐ ISBAS is the same
- ☐ ISBAS is better
- ☐ ISBAS is much better

Additional comments on the previous question:

How does ISBAS compare to Essay exams in accurately measuring your abilities?

- ☐ ISBAS is much worse
- ☐ ISBAS is worse
- ☐ ISBAS is the same
- ☐ ISBAS is better
- ☐ ISBAS is much better

Additional comments on the previous question:

How does ISBAS compare to <u>Hands-on performance assessments (teacher observing)</u> in accurately measuring your abilities?

- ISBAS is much worse
- ISBAS is worse
- ISBAS is the same
- ISBAS is better
- ISBAS is much better

Additional comments on the previous question:

What features would you suggest be added or changed?

Additional Comments:

Which best describes you:

MCTC student UofMN graduate student Other

submit survey