

Lynch School of Education

inTASC Publications

Boston College

Year 2002

How Computer-based Technology Can
Disrupt the Technology of Testing and
Assessment

Michael Russell
Boston College,



TECHNOLOGY AND ASSESSMENT STUDY COLLABORATIVE

How Computer-Based Technology Can Disrupt the Technology of Testing and Assessment

Michael Russell
Technology and Assessment Study Collaborative
Boston College
332 Champion Hall
Chestnut Hill, MA 02467

www.intasc.org



How Computer-Based Technology Can Disrupt the Technology of Testing and Assessment

Michael Russell
Technology and Assessment Study Collaborative
CSTEED, Boston College
Released April 2002
(Commissioned by the National Research Council, November 2001)

Over the past decade, the presence of and access to computer-based technology in K–12 schools has increased rapidly. In turn, computer-based technologies are changing the tools with which teachers teach and students learn. As computer-based tools continue to evolve and become more prevalent in K–12 classrooms, their use provides challenges to and opportunities for assessment. In some cases, the challenges result from pressure applied on testing programs as a result of classroom uses of technology. In other cases, the technology itself can increase the efficiency of testing. And in still other cases, computer-based technology provides opportunities to radically transform testing and assessment. In this paper, I briefly discuss how classroom uses of technology and the efficiency afforded by technology impact testing. The bulk of this paper, however, focuses on disruptive applications of computer-based technology to educational assessment.

Pressure from the Classroom Up

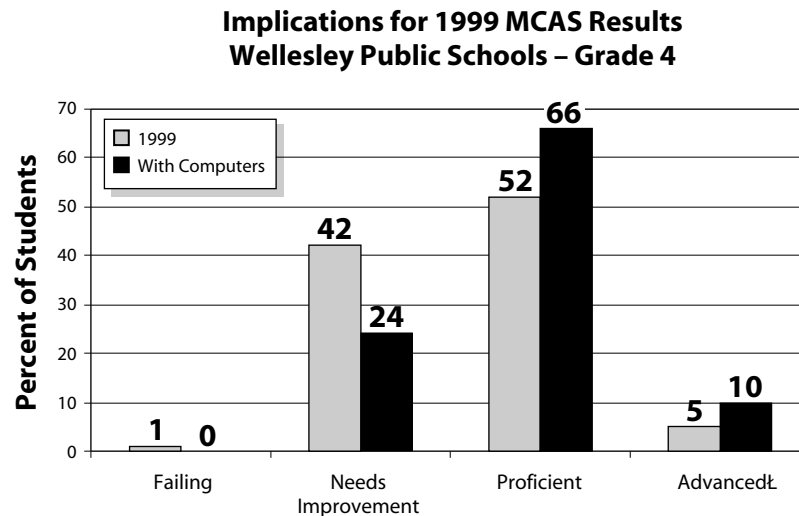
As the use of computer-based technologies gradually becomes a regular component of classroom teaching and learning, the tools with which students solve problems and produce work are evolving from pencil-and-paper-based to computer-based. As students become increasingly accustomed to learning and working with these computer-based tools, a misalignment develops between the tools students regularly use to learn and the tools students are allowed to use while their achievement is tested. In turn, students, teachers and educational systems begin to pressure testing programs to allow the use of these instructional tools during testing. As two examples, students are increasingly using calculators during mathematics instruction and word-processors for writing.

During the mid 90's there was much debate over whether students should be provided access to calculators during testing (Dion, et al, 2000; Dunham & Dick, 1994; Kenelly, 1990). When the debate over calculators first arose several concerns were raised. These concerns related to:

- equity issues: do all students have access to calculators both in the classroom and during testing?
- standardization: should all students use the same type of calculator during testing?
- construct validity: is the construct measured the same when students do and do not use calculators?

While test developers and those who use test results are still concerned about these issues, the widespread and regular use of calculators in the classroom has led many testing programs to allow students to use calculators on items that do not specifically measure students' arithmetic skills.

Similarly, the wide-spread use of word processors for developing and refining written work in K–12 schools poses similar challenges to current paper-and-pencil testing practices. As a series of studies has shown, the writing ability of students accustomed to writing with computers is seriously underestimated by paper-and-pencil tests of writing (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001). In a series of randomized experiments, this mode of administration effect has ranged from an effect size of about .4 to just over 1.0. In practical terms, the mode of administration found in the first study indicated that when students accustomed to writing on computer were forced to use paper-and-pencil, only 30 percent performed at a “passing” level; when they wrote on computer, 67 percent “passed.” In a second study, the difference in performance on paper versus on computer for students who could keyboard approximately 20 words a minute was larger than the amount students' scores typically change between grade 7 and grade 8 on standardized tests. However, for students who were not accustomed to writing on computer and could only keyboard at relatively low levels, taking the tests on computer diminished performance. Finally, a third study that focused on the Massachusetts Comprehensive Assessment Systems Language Arts Tests demonstrated that removing the mode of administration effect for writing items would have a dramatic impact on the study district's results. As figure 1 indicates, based on 1999 MCAS results, 19% of the fourth graders classified as “Needs Improvement” would move up to the “Proficient” performance level. An additional 5% of students who were classified as “Proficient” would be deemed “Advanced”.

Figure 1: Mode of Administration Effect on Grade 4 1999 MCAS Results

As new technologies develop and become a regular component of classroom instruction, it is quite likely that they too will place similar pressures on testing. While it is difficult to see into the future, we are already starting to see calls for the use of graphic calculators during math tests (Forster & Mueller, 2001). As voice recognition rapidly improves, it is likely that it will become the preferred method of composing text for many students. Already some schools are using voice recognition software with students with learning disabilities. As students become comfortable and dependent on these and other emerging learning and production tools, it is likely that tests that prohibit the use of these tools will again underestimate the performance of students who have become accustomed to working with these tools. In turn, testing programs will be challenged to adapt their policies and procedures to accommodate these emerging technologies in order to maintain construct validity.

Efficiency

Beyond their use at the classroom level, computer-based technologies can greatly increase the efficiency of testing. To a large extent, testing programs have already capitalized on the efficiencies afforded by technology. As one example, computer-adaptive testing combines a computer-based delivery system with algorithms that select items targeted at the test-takers estimated ability. Based on the test-takers success or failures on each targeted item or sets of items, the algorithm refines its ability estimate and then presents additional items targeted at the refined ability estimate. This iterative process occurs until the test-takers ability estimate stabilizes. Although computer-adaptive testing complicates the item selection and delivery process, it is usually able to obtain an ability estimate based on a smaller set of items, and thus is more efficient than traditional paper-based tests.

Computer-based technologies are also impacting the efficiency with which open-ended items are scored. Work on computer-based scoring dates back to the work of Ellis Page during the late 1960's. Since Page's (1966; 1968) pioneering efforts, four

approaches to computer-based scoring have evolved and have begun to be used to score student work both in the classroom and on large-scale testing programs (see Rudner, 2001 for an overview of these four methods). For all four approaches, studies have demonstrated that the scores produced by these computer algorithms are as reliable as scores produced by two independent human readers. Clearly, once in a digital format, the use of computer scoring systems can dramatically increase the speed with which open-ended responses are scored and reduce the costs required to compensate human scorers. The use of computer-based scoring systems also could allow examinees to obtain more immediate, or even preliminary, feedback, thus increasing the utility of open-ended tests to inform instruction in a timely manner.

Similarly, moves to administer tests via the internet have the potential to greatly increase the efficiency and utility of testing (Bennett, 2001). Rather than distributing, collecting, and then scanning paper-based tests, the internet can streamline distribution, administration and scoring into a seamless and nearly instant process. In turn, the rapid return of test results could provide valuable information to students and teachers in a timely manner.

As test developers continue to grow familiar with new and developing computer-based technologies, it is likely that they will discover other ways to improve the efficiency of testing. Already, some testing programs are experimenting with ways to generate large banks of test items via computer algorithms with the hopes of saving the time and money currently required to produce test items manually (Bennett, 1999).

Disruptive Applications of Computer-based Technologies

As Madaus has long emphasized, testing is its own technology with its own “body of special knowledge, skills, and procedures” (2001, p. 1). While the application of computer-based technologies described above may increase the validity of inferences based on tests and/or may increase the efficiency of testing, they do not fundamentally impact the technology of testing itself. Even in the cases of adaptive testing and item generation, the psychometric principles and “rules” for test construction developed over the past fifty years are applied without significant alteration to determine which items are to be used in a given situation. In this way, applications to improve the validity and/or efficiency of testing are layered on top of the existing and long established technology of testing.

Computer-based technologies, however, hold tremendous opportunities to dramatically alter the technology of testing. The ability of computers to present complex, multi-step problems that may incorporate several types of media, can have several different paths to reach a solution, and/or can have multiple solutions, coupled with the computer’s ability to record the examinee’s every action, creates opportunities to learn about students’ knowledge, conceptual understanding, and cognitive development in ways that today’s technology of testing cannot.

Although the principles and procedures of the current technology of testing are sound, several shortcomings arise. Despite efforts to incorporate open-ended items into some tests, most test items result in binary information about a student, namely did s/he answer correctly or incorrectly. While scoring guides for some open-ended items focus on the procedures and cognitive process students use to solve problems,

these items are dependent upon students' descriptions of their processes, which are often incomplete and/or inaccurate reflections of their actual processes. As a result, these items provide very indirect and crude insight into examinees' cognitive processes.

Similarly, while the educational community uses tests for a variety of purposes including diagnosing students' strengths and weaknesses, measuring achievement, aptitude and ability, assessing the impact of instruction on student learning, and examining the quality of education students receive within a school, district, state or even country, test experts have long argued that the current technology of testing should be applied to meet a single purpose at a time. As Haney, Madaus and Lyons argue, the fundamental problem with using a single test or assessment for multiple purposes "is that such tests require...fundamentally different characteristics." (1993, p. 264). Nonetheless, many current testing programs attempt to use a single test or set of closely related tests to fulfill multiple purposes. As an example, the Massachusetts Comprehensive Assessment System uses results from tenth grade language arts and mathematics tests to: 1) make decisions about student competency and eligibility for graduation; 2) make decisions about the quality of education within individual schools; 3) to identify exemplary educational programs; 4) to assess the effectiveness of state and local interventions (such as tutoring); and 5) to help teachers and schools diagnose student weaknesses. Despite including multiple item formats and requiring several hours to complete, the tests contain roughly fifty items that are performed by all students across the state. While performance on the same set of items helps reassure the public that decisions about student competency and graduation eligibility are based on the same information, this limited set of items attempts to assess such a broad domain that only a handful of items are used to measure the sub-domains. As a result, there is very little information available to usefully diagnose students' strengths and weaknesses. Moreover, the tests do not attempt to probe why students may have performed poorly within a given sub-domain. Similarly, by administering the same set of items to all students rather than spiraling item sets across sets of students, schools and districts are provided with very limited information about the strengths and weaknesses of their educational programs. In short, while tests like the MCAS ambitiously attempt to satisfy several purposes, they fail to adequately meet these needs.

Beyond the MCAS, several state developed and commercial tests attempt to help teachers diagnose student weaknesses. These tests, however, focus on specific content within a given domain and often times use multiple-choice formats to measure student performance within the several sub-domains. As a result, the diagnostic information provided to educators is typically limited to whether or not students tend to succeed or fail on items within a given sub-domain. While this information helps educators identify those sub-domains that may be in need of further instruction, these diagnostic tests tend to provide little or no information about why students may be struggling within a given sub-domain. Rather than diagnosing the misconceptions and/or specific skills sets that interfere with students mastery of the sub-domain, most current diagnostic tests provide little more information than an achievement or mastery test.

Among other shortcomings of current testing practices is that most testing currently occurs outside of instruction. As a result, the amount of instructional time is

decreased. Ironically, this problem is exacerbated in settings that administer diagnostic tests on a regular and frequent basis to help focus instruction and/or use a series of achievement tests to better measure the impact of instruction on student learning over time. With each test administration, regardless of whether the use is internal or external to the classroom and whether it is teacher-developed or developed external to the classroom, instructional time is decreased. While some educators argue that embedded assessment (see Wilson & Sloane (2000) for an example of an embedded assessment system) will better streamline the traditional instructional and assessment cycle, externally developed or mandated tests still diminish instructional time.

It is these shortcomings that disruptive applications of computer-based technology to the technology of testing could well address. By building on learning systems currently in use and/or under-development, there is tremendous potential to capture information about students and their learning during the actual learning process. By presenting complex problems as part of the instructional process and examining the strategies students use to solve these problems and then comparing these strategies to those of novices, learned, and experts in the field, the notion of mastery could be expanded from the ability to consistently answer problems correctly to the ability to incorporate knowledge and skills in a way that resembles expertise. Information collected as students interact with the learning system could also be used to diagnose student learning styles, common errors or tendencies, and misconceptions. Once identified, the systems could help teachers intervene immediately and could help structure future instruction in a way that is compatible with the students' learning style. In addition, as students master sub-domains, the systems could track student achievement. Since achievement is tracked throughout the year, attempts to assess the educational quality or effectiveness of schools, district and/or states could be based on the full range of content and skills addressed during the entire year. And because the information would be broader and deeper than that provided by current achievement tests, the need for external exams might be eliminated.

Below, I describe two collaborative efforts that have recently been initiated to apply computer-based technologies in a manner that substantially depart from current approaches to testing and assessment.

Surgical Simulation

The US Army is often credited with sparking the growth of large-scale standardized testing. With the onset of World War I and the need to quickly assess and assign recruits to various positions believed to require different levels of intelligence, the Army administered the Army Alpha and Beta intelligence tests to over 1.75 million recruits (Gould, 1996). Soon thereafter, school systems began using standardized achievement tests to evaluate program effectiveness (Madaus, Scriven and Stufflebeam, 1983). Since then, testing has grown into a billion dollar industry (Clarke, Madaus, Horn and Ramos, 2001).

Given the initial merit and stimulus the US Military gave to the standardized testing industry, it is fitting that the US Military is now playing a major role in reshaping future assessment methodologies. In response to a 1998 report issued by the General Accounting Office that underscored the need to provide military medical personnel

with trauma care training that reflected the injuries encountered during wartime, the US Army Medical Research and Materiel Command (USAMRMC) Telemedicine and Advanced Technology Research Center (TATRC) has launched several initiatives involving medical simulations. While the main purpose of these initiatives is to develop medical and surgical simulators to efficiently and more effectively train Army Medics, these simulators are providing unique opportunities to assess kinesthetic, content knowledge and medical decision making skills.

Working collaboratively, the Center for the Study of Testing, Evaluation and Educational Policy (CSTEPP) at Boston College and the Center for the Integration of Medicine and Innovative Therapy (CIMIT) are applying computer-based technologies to assess several aspects of medic training and proficiency. As one example, CIMIT has developed a chest tube and surgical airway simulator. The simulator is intended to train medics how to alleviate three conditions commonly caused by chest trauma, namely tension pneumothorax (collapsing lung with trapped air under pressure), hemothorax (collapsed lung with blood in the chest cavity), and hemopneumothorax (blood and air in the chest cavity). All three conditions are life threatening if not alleviated in a relatively short period of time. As part of the learning system, medic recruits first interact with a web-based tutorial that provides information on basic first aid, detailed descriptions of these three conditions, protocols and video demonstrations of the procedures required to alleviate these conditions, and detailed descriptions of common complications and appropriate counter-actions. As part of this component of the learning system, opportunities for recruits to demonstrate the acquisition of the basic knowledge through traditional multiple-choice items are being incorporated. When mastery of this information is not demonstrated, recruits are presented with additional information to help them master the content. While this component of the learning system does not expand upon the current technology of testing, the actual simulator does.

Upon demonstrating mastery of the content knowledge, recruits are then introduced to the simulator. The simulator combines a sophisticated mannequin (figure 1 below) that contains flesh-like tissue, bone-like ribs, and pockets of blood-like liquid with a computer that contains an exact model of the mannequin with the addition of internal organs. In addition, all surgical tools employed for these procedures are connected to tracking devices that record all movements made with the device inside and outside of the chest cavity. By combining the instrument tracking with the simulated model of the mannequin's internal organs, the simulator is able to record the amount of time it takes to perform each task required for a given procedure, while also monitoring all movements of the instruments in 3 dimensional space. Based on these recorded movements, it is then possible to calculate factors that can impact the success of the procedure such as the speed with which instruments enter the cavity, their angle of entry and their depth of entry. In addition, it is possible to examine factors such as acceleration and deceleration and changes in the direction and angle of movement. In addition, using the recorded movements the learning system is able to reproduce the procedure on screen and show the relationship between surgical tools, ribs and key organs that could be harmed.

Figure 2: CIMIT Chest Tube Trauma Mannequin (also known as VIRGIL)



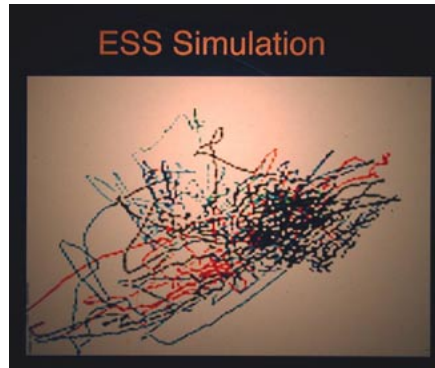
As a training tool, the simulator provides several benefits. Whereas medics typically practice these procedures on animals and do so only a couple of times, procedures can be performed repeatedly on the simulator (and without harm to animals). In addition, since the mannequin is a reproduction (both externally and internally) of a real person and has tissue and bone properties very close to real flesh and bones, the training more accurately reflects procedures that will likely be performed in the field. Finally, the portability of the mannequin allows training to occur just about anywhere (even on a helicopter en route to a battlefield).

From an assessment perspective, the simulator enables unique approaches to diagnostics and mastery testing. As the simulator provides opportunities for the recruit to practice new procedures or introduces new complications, the system can identify tendencies such as inserting an instrument at a dangerously steep or flat angle or inserting instruments too deep or shallow. This information can then be shared with the recruit and the instructor.

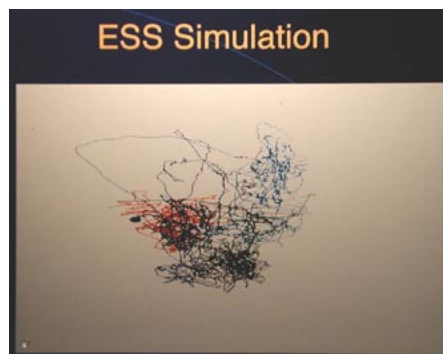
The simulator also has potential to compare the skill of the recruits with those of masters. These comparisons can be made at the macro or the micro level. At the macro level, the enhanced-reality reproductions of the recruits can be layered on top of the reproduction of an expert allowing the recruit to visually compare the “track” of their performance with that of the expert. Through this macro comparison, important differences in technique may become apparent. On subsequent attempts, the recruit can then adjust his/her technique until it reflects that of the expert. As an example, figure 2 depicts the motion trajectories of novice, trainee and experienced surgeons performing a simulated sinus surgery (note that this example is not from the chest tube simulator). As expertise increases, random motion decreases and movements become more precise and focused on each of the four specific areas of work.

Figure 2: Motion trajectories of novice, trainee and experienced surgeons

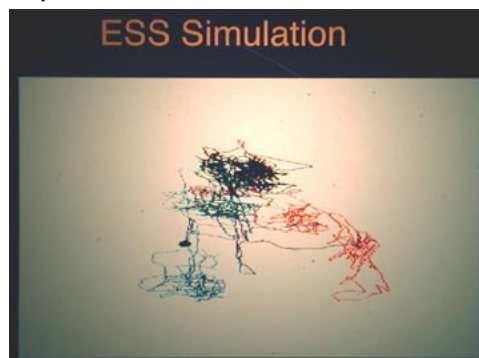
Novice



Trainee



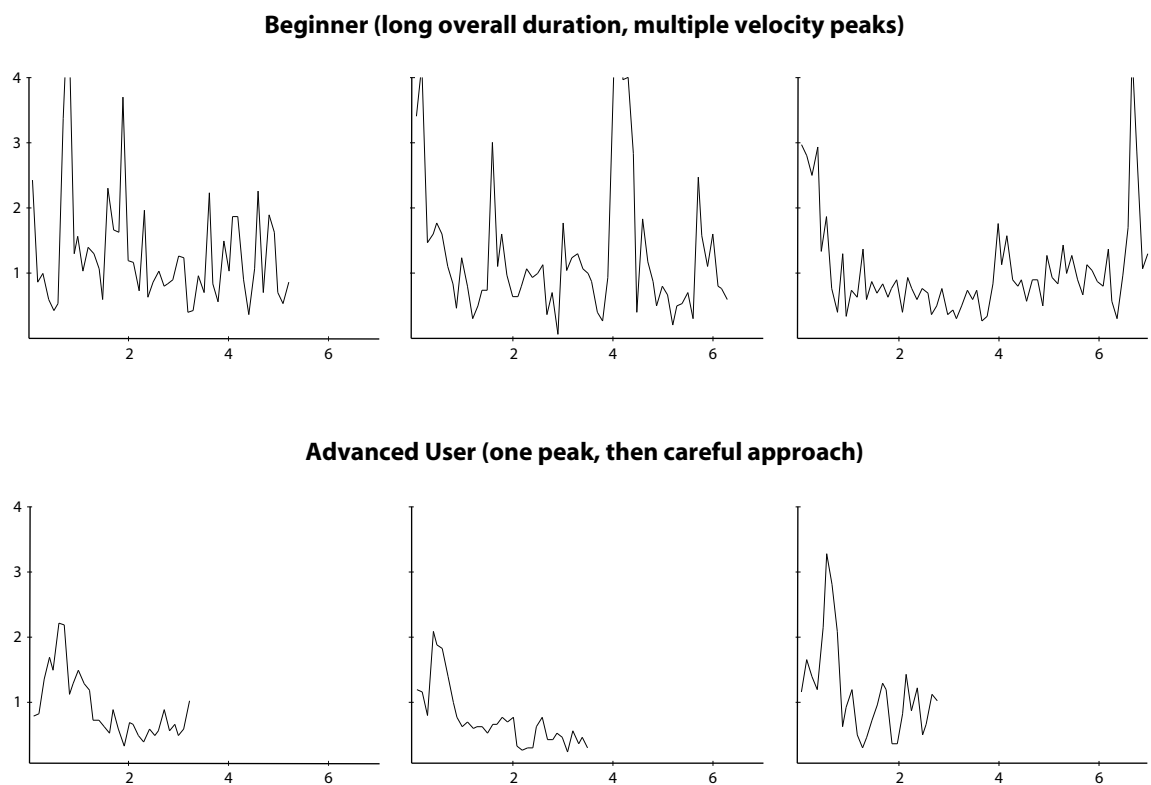
Expert



(Images taken from Metrics for Objective Assessment of Surgical Skills Workshop Draft Report)

At the micro level, individual metrics such as changes in velocity, angle of insertion or depth of insertion, can be compared between the recruit and experts. As an example, figure 3 compares both the amount of time required to complete a task and the velocity of movements while performing the task for advanced and beginning surgeons. In all three trials, the advanced surgeon completed the task in about half the time. In addition, the advanced surgeon executed the tasks with one initial burst of speed and then deliberately slowed down whereas the beginner made several bursts of speed and inefficiently narrowed in on the target.

Figure 3: Results of motion tracking of novice and expert surgeons acquires the target.



(Images taken from Metrics for Objective Assessment of Surgical Skills Workshop Draft Report)

Whether focusing on the macro or micro images of performance, comparisons between the performance of the recruits and experts may be a useful way to assess mastery. In addition, by examining how many practice sessions are required before a recruit's performance reflects that of an expert, it is possible to identify recruits who seem to possess innate kinesthetic surgical skills and/or who are rapid learners – recruits that could be very valuable to the Army when training time is cut short by military engagement.

By presenting recruits with scenarios that involve a range of complications and then examining how the recruit responds, the learning system provides opportunities to examine how well the candidate is able to integrate his/her content and conceptual

knowledge with his/her kinesthetic skills. The realism of the scenario could be further enhanced by placing the chest tube simulator in a simulated war environment and/or by introducing the scenario after extended physical exercise and/or sleep deprivation. The recruit's ability to respond to complications and conduct the necessary physical movements can be examined in a real-life context. Finally, by providing the recruit access to reference materials that might be available in the field (either during initial training or during future training), the recruit's ability to rapidly access and apply information to resolve a problem could also be assessed.

K–12 Learning Environments

At first brush, medical simulators may seem far removed from K–12 education. However, the approaches used to collect a diverse set of information about recruits and the challenge of figuring out how to make use of this set of information are directly applicable to learning systems currently in place and under-development for K–12 schools.

Recently, CSTEPP and the Concord Consortium have begun to brainstorm ways in which assessments can be built into learning systems. To date, our discussions have been limited to *Biologica*, a learning system developed by the Concord Consortium that focuses on genetics. The system is intended to help students learn about genetics through guided exploration. In its current form, *Biologica* comprises 13 modules each of which focuses on a different and increasingly more complex aspect of genetics. In most cases, the modules begin by asking students to explore a specific topic by manipulating genetic traits of a fictitious species of dragons. As an example, figure 4 depicts the first exploration students encounter in the second module. In this exploration, students manipulate the dragon's chromosomes to determine how many different ways they can produce a dragon with horns. As each module progresses, new concepts are revealed through guided exploration. For example, the first set of explorations during lesson two culminate by asking students to describe how traits are produced in dragons (Figure 5). At times, the learning system presents textual or graphical information to explain concepts and provides students with access to various tools and pieces of information via menu selections. In addition, the system often asks students to demonstrate their understanding via written responses to specific questions, multiple-choice questions, and, most often, modifying different aspects of genetic codes to create dragons with specific traits or to determine how a trait suddenly appeared in a generation of dragons. Throughout the students' interaction with the system, all interactions with the system are recorded.

Figure 4: First Exploration During Second Module of Biologica

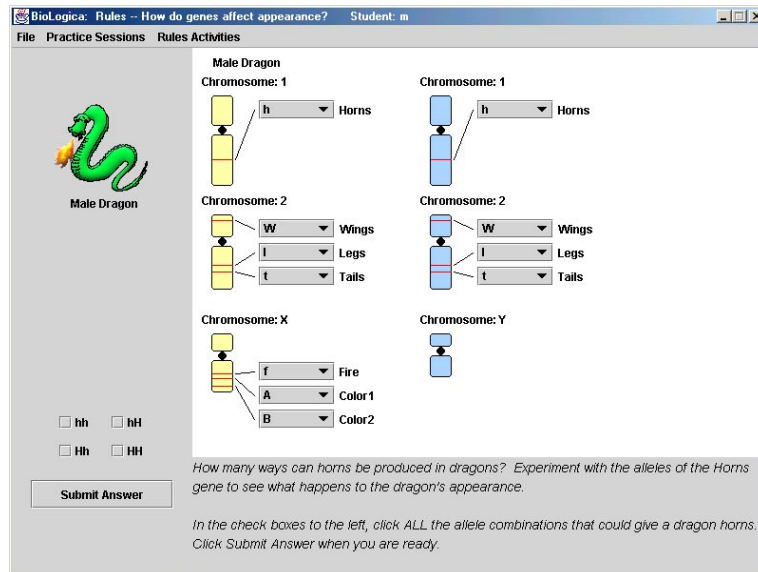
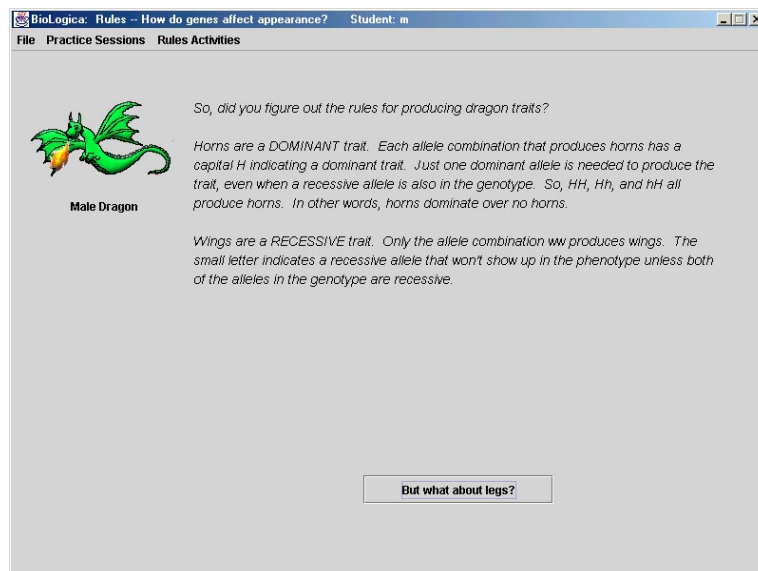


Figure 5: Generalizing from Guided Explorations to Rules of Genetics in Biologica



From an instructional perspective, Biologica enables students to explore a complex topic via a variety of media and enables teachers to work individually or with small groups of students as questions arise. From an assessment perspective, the learning system provides a number of opportunities to assess student learning. Beyond examining students' understanding via their responses to the multiple-choice and open-ended questions (which could be analyzed by computer), the guided explorations and the problems posed to students present opportunities to:

1. Examine students' conceptual understanding by examining the tools and information they opt to use, the amount of time required to solve problems, the type of strategies they employ (e.g., randomly changing chromosomes versus making initial changes on the appropriate chromosomes), as well as their success with the problem;
2. Compare the students' pattern of interactions with those of "experts"; and
3. Probe apparent misconceptions by presenting additional problems that focus on the specific misconception.

In addition, insight into students' learning styles might be gained by beginning modules and sub-modules in different ways. For example, a module might begin with a textual explanation of a concept followed by an opportunity to demonstrate understanding. If understanding is not demonstrated, subsequent "instruction" might employ a guided exploration of the same concept. If understanding is still not demonstrated, a visual presentation of the concept might follow. Across multiple concepts, the order of presentation could be altered and the efficiency with which the student mastered the concept recorded. After several iterations, the system might identify the preferred order of instructional strategy and utilize that order for that student during subsequent modules.

Finally, and perhaps most importantly, since the learning system provides multiple opportunities for students to demonstrate conceptual understanding, the need to administer a separate test on the material mastered could be eliminated. Moreover, since records could be sent electronically to any location, it would be possible to maintain a database that indicates which students have mastered each concept. This information could be used by the teacher to identify common misconceptions and help inform instruction. In addition, this information could be used to assess achievement at the student level or at a higher level of aggregation. While there might not be much value in recording achievement data at an aggregate level for a single learning system, the value would increase rapidly as more learning systems are used within a school, district and state. And, again, if this information proves to be redundant to information provided by on-demand, external tests, the need for external standardized tests might be eliminated.

Moving from Virtual Possibilities to Reality

While the possibilities are enticing, several challenges must first be overcome. These challenges fall into three broad categories: A) Technical; B) Political; and C) Practical.

Technical Challenges

The first major challenge involves figuring out which information collected by these systems is most useful for a given purpose and how to combine this information so that it is interpretable. By no means is this an easy challenge to overcome. Unlike a traditional multiple-choice test that may contain fifty to hundred pieces of binary information, the amount of data produced by these systems can span several pages and include everything from the amount of time between actions, quantity of changes made before a solution was found, materials and tools accessed, textual responses, and long lists of items clicked, alterations made, or items moved. While current psychometric models should not be abandoned altogether, new models will need to be created to make use of these multiple pieces of information.

Given the potential to map actions, whether they be physical in the case of surgical simulators or cognitive in the case of K–12 learning systems, methods of analyzing graphical representations of processes should also be explored. As an example, advances in image recognition now make it possible to quickly identify people by comparing video images of their faces with digital photographs stored in large databases. Adapting this technology to compare the paths of learners and experts may prove a useful way to assess level of expertise.

If comparisons between learners and “experts” are to be made, a significant investment must also be made in capturing the strategies and processes that experts employ. While this may be a relatively easy task in the case of physical skills such as those employed during surgery, it is a significantly greater challenge for K–12 learning systems. This challenge is compounded at lower grade levels for which the definition of “expertise” may be radically different than for high school students. While settling on an appropriate definition of expertise may be more political than empirical, acceptable definitions will need to be reached before such comparisons will be broadly embraced.

Much work will also be needed to validate the decisions made as students work with such learning systems. This is particularly true for decisions about academic achievement. While these systems have the potential to greatly reduce or eliminate external testing, these radical changes will not occur unless it can be demonstrated that the information gleaned from these systems are redundant with that information provided by external tests. Moreover, given the current climate of high-stakes testing, it will be necessary to develop methods of verifying the identity of the student working with the learning system.

Political Challenges

Currently, political and educational leaders strongly embrace large-scale and high-stakes testing and educational accountability appears to be the top priority shaping our educational system. Despite calls for the incorporation of multiple-measures into these school accountability systems, political and educational leaders appear deaf to these calls. One reason for the resistance to broaden the types of measures (be they grades, teachers judgements, portfolios or work samples, or “performance-based” tests) likely relates to a belief that standardized tests provide more objective, reliable and accurate measures of student achievement. In part, the failure to expand the measures used for accountability purposes results from the failure of critics to convince leaders of the utility and validity of these other measures. Although several years of research, development, validation and disseminations are required before integrated learning and assessment systems could be widely available, efforts should begin now to familiarize political and educational leaders with these methods of assessment. To increase buy-in, roles in the development process should also be created for political and educational leaders.

Additionally, efforts are needed to help leaders see the potential role computer-based technology can play in expanding notions of accountability. As Haney and Razcek (1994) argue, current notions of accountability in education are narrowly defined as examining the performance of schools via changes in their test scores. Under this definition, the iterative process of reflecting on programs and strategies, providing accounts of the successes and shortcomings of those programs, and setting goals in response to those shortcomings is, at best, an informal and secondary component of school accountability. While computer-based learning and assessment systems have the potential to make information provided by current achievement tests redundant and thus eliminate the need for such external tests, computer-based technologies could also be applied today to disrupt current notions of school accountability by providing a forum for schools to account for their practices and to learn from the practices of other schools. Rather than simply transferring achievement testing from paper to a web-based delivery system (as is currently occurring in Virginia, Oregon, Georgia and South Dakota), the internet could be used to collect information about classroom performance (e.g., electronic portfolios or work samples), more closely scrutinize the reliability of scores given to such work, return data from multiple measures in more useful formats, share information and student work with a wider base of constituents, and provide a forum for schools to account for their programs and strategies. Investing now in developing web-based accountability systems that broaden the definition of educational accountability will better set the stage for replacing external state-mandated achievement tests with assessments that are integrated with learning systems.

Practical Challenges

If these disruptive approaches to assessment are to become a regular practice within schools, learning systems like *Biologica* will need to be developed in a wide range of topic areas. Anticipating the potential growth of these types of learning systems, the Concord Consortium has developed a scripting language that allows users to easily create new modules for current learning systems or develop new learning systems. In a sense, this scripting language is analogous to HTML in that it is easy to learn and has the potential to standardize learning systems.¹ Not only will this scripting language be useful for those who want to develop new learning systems, it also provides an easy way to alter current systems so that assessment components can be added or modified.

The high initial costs required to develop a learning system coupled with the need to have a learning system (or at least a prototype) in use with students before much of the technical work described above can be performed poses a major obstacle. Not long ago, the National Board on Educational Testing and Public Policy worked with a coalition of schools, political and educational leaders, and internet-based database developers to develop a proposal to design a comprehensive web-based accountability system that builds on the Massachusetts current MCAS. While the proposal dedicated substantial resources to piloting and validating the system, the high costs associated with developing database engines and interfaces resulted in a research and development budget that was too large to be attractive to funders. The same potential exists for learning and assessment systems. One strategy is to focus first on those systems that are already in use or already have funding to support development. By collaborating with the developers of these existing systems, the resources required to support the development and validation of new approaches to assessment are greatly reduced. In addition, by starting with those systems that are already in use in schools, sets of data are more immediately available. As an example, *Biologica* is currently being used by some 10,000 students across the nation. In addition, because it is delivered via the web, its modules can be easily updated and student data can be sent to a central database. Thus, rather than investing two to three years developing a learning system, working with the highest quality systems that are currently in use or will soon be in use provides opportunities today to begin exploring some of the technical challenges outlined above.

A third practical challenge involves tapping expertise from a range of fields. As the NRC's *Knowing What Students Know* notes, collaboration among cognitive scientists, assessment experts, and content experts is needed to better inform the development of new approaches to assessment. But in addition, input from instructional leaders and developers of technology are also needed to better anticipate how these systems might be used within classrooms and how emerging computer-based technologies might impact these systems. Finally, as noted above, political and educational leaders must be brought into the research and development process to better assure that these systems will be accepted as valid means of measuring student achievement.

¹ CIMIT is developing a similar language called CAML that could create a common structure for describing tissue, tools, interactions and other features of medical simulators.

Clearly, there is a tremendous amount of work that must be performed before these learning systems can adequately meet assessment needs. As the way students – whether they be in the K–12 classroom or surgeons – learn changes, there are important opportunities to acquire a more thorough and useful understanding of how and what students learn. Without question, as current and future computer-based technologies are used regularly in the classroom, they will continue to pressure changes in testing. While small-scale studies may be required to initially demonstrate the need to incorporate these technologies into testing, the primary responsibility for examining and implementing changes falls on the testing programs themselves. Similarly, given the financial rewards the testing industry will realize, it is likely that it will continue to take on the challenge of developing ways to apply computer-based technologies to increase the efficiency of testing. However, given the potential of computer-based technologies to seriously disrupt the current technology of testing, it is unlikely that the testing industry itself will invest in researching and developing disruptive uses of computer-based technology. Given the potential positive impacts integrated learning and assessment systems could have on teaching and learning coupled with the vast amount of technical work that must be done to develop these new methodologies, the educational community needs to follow the military's lead by investing now in developing disruptive applications of computer-based technology to the technology of testing and assessment.



References

- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practices*, 18(3), 5–12.
- Bennett, R. E. (2001). How the internet will help large-scale assessment reinvent itself. *Educational Policy Analysis Archives*, 9(5), [Available online: <http://epaa.asu.edu/epaa/v9n5.html>].
- Clarke, M., Madaus, G., Horn, C., & Ramos, M. (2001). The marketplace for educational testing. *National Board on Educational Testing and Public Policy Statements*, 2(3).
- Dion, G., Harvey, A., Jackson, C., Klag, P., Liu, J. & Wright, C. (2000). SAT program calculator use survey. Paper issued by the Educational Testing Service, Princeton, NJ.
- Dunham, P. H. & Dick, T. P. (1994). Research on graphing calculators. *The Mathematics Teacher*, 87(6).
- Forster, P. A. & Mueller, U. (2001). Outcomes and implications of students' use of graphics calculators in the Public Examination of Calculus. *International Journal of Mathematical Education in Science and Technology*, 32(1), 37-52.
- Gould, S. J. (1996). *The Mismeasure of Man*. New York, NY: W.W. Norton & Company.
- Haney, W. & Raczek, A. (1994). Surmounting Outcomes Accountability in Education. Paper prepared for the US Congress Office of Technology Assessment.
- Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston, MA: Kluwer Academic Publishers.
- Kenelly, J. (1990). Using calculators in the standardized testing of mathematics. *Mathematics Teacher*, December.
- Madaus, G. F. (2001). Educational testing as a technology. *National Board on Educational Testing and Public Policy Statements*, 2(1).
- Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (1993). Program evaluation: A historical overview. In *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*. Boston, MA: Kluwer-Nijhoff Publishing.
- Metrics for Objective Assessment of Surgical Skills Workshop: Developing Quantitative Measurements through Surgical Simulation. Summary (Draft) Report for conference held in Scottsdale, AZ, July 9–10, 2001.
- Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47, 238–243.
- Page, E. (1968). The use of computers in analyzing student essays. *International Review of Education*, 14(2), 210–221.
- Rudner, L. (2001). Bayesian Essay Test Scoring sYstem—BETSY, [Available online: <http://ericae.net/betsy/>].

- Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), <http://olam.ed.asu.edu/epaa/v5n3.html>.
- Russell, M. & Plati, T. (2001). Mode of Administration Effects on MCAS Composition Performance for Grades Eight and Ten. *Teachers College Record*, [Available online: <http://www.tcrecord.org/Content.asp?ContentID=10709>].
- Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20).
- Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.

