Improving Accountability Models by Using Technology-Enabled Knowledge Systems (TEKS)

CSE Report 656

Eva L. Baker
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

July 2005

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University Of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.2 Systems Design and Improvement: Ideal and Practical Models for Accountability and Assessment, Strand 1
Project Directors: Eva L. Baker, CRESST/UCLA, and Robert L. Linn, CRESST/University of Colorado at Boulder

The work reported herein was prepared for the Education Commission of the States and supported in part by a grant from The Bill and Melinda Gates Foundation and in part by the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

IMPROVING ACCOUNTABILITY MODELS BY USING TECHNOLOGY-ENABLED KNOWLEDGE SYSTEMS (TEKS)

Eva L. Baker

UCLA/National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

This paper addresses accountability and the role emerging technologies can play in improving education. The connection between accountability and technology can be best summarized in one term: *feedback.* Technological supports provide ways of designing, collecting and sharing information in order to provide the basis for the improvement of systems and outcomes. Because technological options are so varied and the future of their development unpredictable, the paper argues that the first screen for the review of technology is one addressing core functions the technology serves. Technologies can provide greater efficiency in managing information or serve to improve the quality of the information used to guide the system. At this point, these functions are frequently mutually exclusive, but in the future, wise selection of technology should involve a consideration of how the technology adds value by increasing quality and efficiency in the enterprise. Quality is added when more knowledge is available to the user. For choices that need to be made about investments, a set of criteria is provided to guide decision logic. Criteria include validity, accuracy, and utility, with an emphasis on users directly involved in learning (students and teachers). In addition, criteria are reviewed related to software/hardware operation, including security, backup, compatibility, resilience, and transparency. The assumption is that systematic application of criteria will influence the market to drive development of more useful systems for education at a cost that can be withstood. Throughout, examples of technology uses now and soon to be available are discussed. The paper concludes with two extended educational scenarios constructed to show elements of design and technology in practice in two very different settings.

Part I: Improving Accountability Models by Using Technology-Enabled Knowledge Systems (TEKS)

This paper will analyze the potential use of technology-enabled knowledge systems (TEKS) as integral components in educational accountability systems. The relationship between accountability and knowledge is obvious, for accountability systems depend upon usable information (Baker, 2003) to support their operations. Technology functions, in part, to increase the availability and usability of information. Because there are so many different examples of technology—and more, unimaginable advances left to be developed—it is plainly not practical to give advice on each function or on each system that is presently at hand. Instead, we will provide principles and criteria that will adapt both to the technologies we now have and to those on the horizon. Because our fundamental topic of attention is accountability, we will consider how technology, when applied to certain information functions, can serve to improve both educational processes and outcomes. Even though we provide a number of examples throughout, readers will need to apply the principles we discuss to particular technologies and to their own settings. Essentially, our major question is straightforward—but not easy:

> *How can technology improve accountability and simultaneously improve the efficiency and quality of educational outcomes and supporting processes?*

What specific learning, concepts and ideas should you get out of this paper? Depending upon your background and your present responsibility, you should learn or update your understanding in the following areas:

- basics of accountability;
- functions of technology systems—efficiency and quality;
- functions of technology in education and accountability systems;
- criteria for decision making;
- criteria applied to technology in support of assessment and testing;
- criteria applied to technology in support of information exchange;
- ways to manage system, infrastructure, and capacity demands.

In overall sequence, the paper starts with general considerations, moves to criteria and their applications in accountability functions, and then zooms out again to discuss additional criteria related to systems, context, cost, and capacity—in other words, the requisites to make the system work.

Accountability at the Core

Every accountability system is based on a set of fundamentals that can be summarized in simple 4-part scheme: (1) Valued and attainable goals are identified; (2) responsible parties (RPs) are identified, and feedback on progress toward the goals is given; (3) different levels of performance or rates of progress receive differentiated rewards and sanctions; (4) improvement occurs because RPs modify their actions in order to receive rewards and avoid sanctions. The sum of individual actions determines system progress (Figure 1; Table 1).

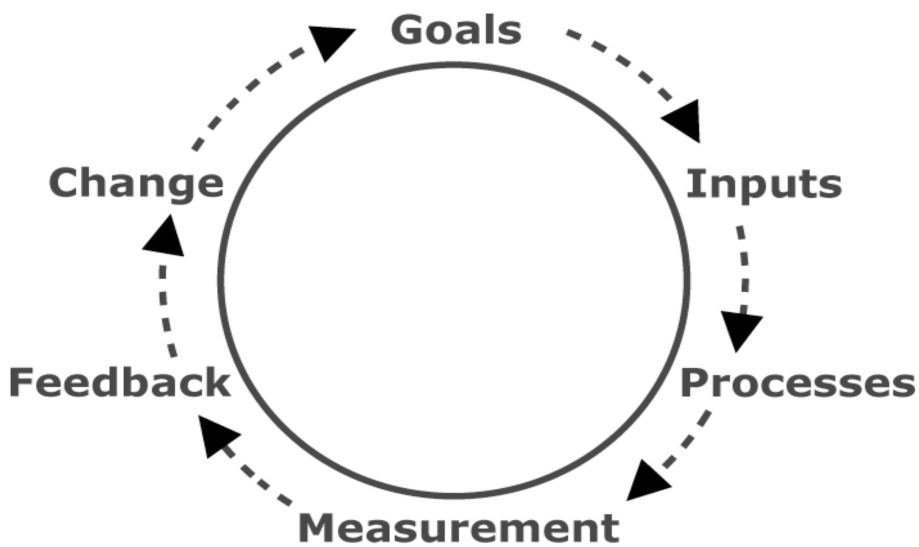

Figure 1. Accountability processes.

Table 1

Fundamentals of Accountability Functions

- Valued and attainable goals—students, teachers, institutions.
- Accurate feedback.
- Rewards and sanctions based on progress and attainment.
- Improvement results from combined efforts by responsible parties (RPs).

There are plenty of examples where such accountability-based schemes have worked—competitive sports, entrepreneurial business, and weight loss. In education and training, health care, and other service sectors, we have been employing variations of this general approach for years with some success. What is new recently is that expectations and the tempo of change have increased just as we move into a less certain fiscal environment.

Spector proposed an activity representation appropriate to understanding accountability (Figure 2).

Representing the activities within an accountability system in a simple, cyclical manner may encourage one to forget one or more of the responsible parties or to focus on one kind of feedback as opposed to another. Moreover, . . . knowledge management system technology may be a key component of a next-generation accountability system. As a consequence, an activity system representation is likely to provide an appropriate design and development framework for such a system since it forces the design team to consider all involved subjects (including responsible parties) working towards shared goals (objects) with a division of labor (tasks and activities) among those subjects. (J. Michael Spector, personal communication, May 24, 2003; see also Nardi, 1996, and www.edu.helsinki.fi/activity/6b.htm)

An Activity System

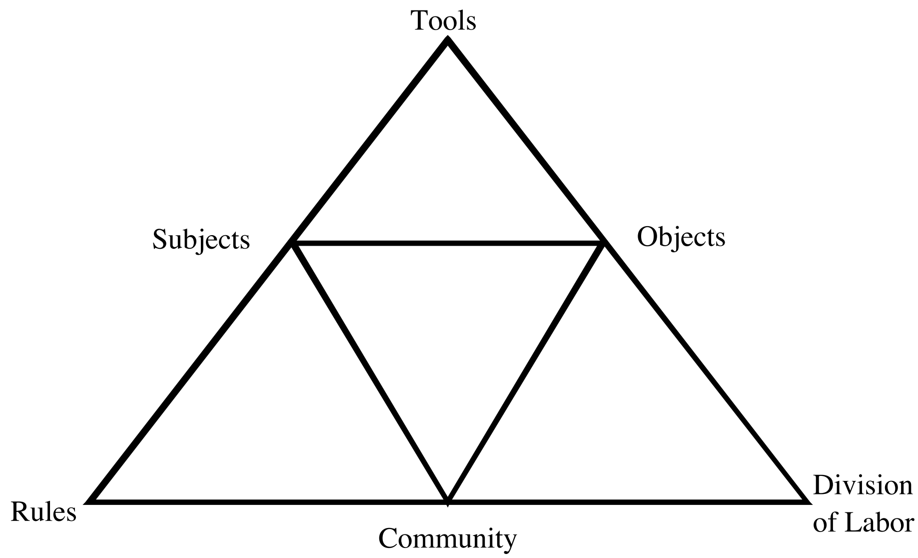

Figure 2. An activity system. Adapted from Bonnie Nardi (Ed.), *Context and Consciousness: Activity Theory and Human-Computer Interaction* (Cambridge, Massachusetts, MIT Press, 1996).

Education itself brings along with it some special provisions for accountability. Participation in an educational accountability system is voluntary neither for students nor for teachers and administrators. Compared to environments where one chooses to compete (in sports) or opts to work at a particular company, participants in education are captives of the system in which they learn, teach, or manage. Furthermore, unlike sports or business, the student participant cannot be cut from the team, except in the most extreme circumstance. The very fact that accountability is mandatory requires special protection for its participants. Next, because the preponderance of education is supported with public money, the design, computation, and consequences of accountability systems must also serve the public trust and meet higher expectations of quality, stewardship, openness, and ethical behavior than those of comparable systems in commercial enterprises. A set of *Standards for Educational Accountability Systems* (Baker, Linn, Herman, & Koretz, 2002) and a chapter explaining their rationale (Baker & Linn, 2004) provide a self-assessment approach to assess the degree to which accountability systems reflect best research and practice. Nonetheless, even well-designed accountability systems, with positive goals and transparent linkages, depend upon technical elements that demand constant scrutiny. It is not sufficient to buy a set of tests, promulgate rules, put the educational accountability system in place, and then watch scores rise or fall, even if the political world would so permit. Accountability systems—to be ultimately credible—must themselves exhibit the features they require: attention to feedback and improvement in their own design and function over time. Leaders must address the degree to which the information produced by the system is valid (its results emphasize and encourage the valuable and reduce negative or unethical behaviors). Vigilance is needed to prevent efforts focused exclusively on the numbers without considering the quality of performance that they are intended to indicate, in other words, the real improvements in learning that we all desire. No one will be fooled for long.

Feedback as a Core Feature of Accountability

In education, as in other areas, the success of an accountability system, not surprisingly, is based on the details of its design and operation. For instance, consider the following guidance: (a) Feedback must be relevant; (b) feedback must be accessible to responsible parties (RPs); (c) feedback must be timely; (d) for productive action to occur, a repertoire of alternative actions must be available to the RPs; (e) progress (and success) measures must be sensitive to different actions; (f) the RPs in the system must be able to decide, control, and intervene on substantive matters and (g) the action plan should be transparent to its users and to the publics it serves. Keywords are summarized in Table 2.

Feedback is one of the most widely researched social science phenomena and often pays off (improves performance) even without additional interventions. The power to know one's strengths and weaknesses, at least for motivated and skilled learners, is sufficient to stimulate self-review of procedures. For schools filled with students with a range of interests, motives and aspirations, interventions will be needed to supplement simple feedback. To be sure, even the abbreviated feedback

Table 2

Feedback Functions in Accountability

- Relevant
- Accurate
- Accessible
- Timely
- Repertoire of options
- Sensitive measures
- Decision, control, intervention
- Transparent action

and action relationships listed above will present a tall order for accountability systems involving participants among the hundreds of thousands or even millions of students, their teachers, and the management leadership of their educational institutions. Especially if feedback is to be relevant, timely, and presumably adaptive, in a system that prizes local control, the task is formidable. Yet, work has been undertaken on dynamic systems that aptly characterize the process (Mandinach & Cline, 1994; The MIT System Dynamics in Education Project, http://sysdyn.clexchange.org; Morecroft & Sterman, 1994; Sterman, 1994).

Before we continue, let's get the matter of common language and definitions out of the way. For the remainder of this paper, terms will be used as listed in Table 3.

There are, in practice, numerous overlaps in the use of these terms. In the public education sphere, "assessment" has been preferred more recently to "testing" but usually stands for much of the same thing, especially in the case of "State

Table 3

A Common Language for Accountability

- Accountability is a system that attaches incentives to progress in meeting goals.
- Evaluation is the act of forming a judgment of the merit or value of an enterprise.
- Measurement is the procedure used to determine formally the degree to which improvement along a dimension has occurred.
- Testing is a formal, often standardized procedure for obtaining and scoring performance in order to place the examinee's results on a comparative scale.
- Assessment is synonymous with testing for some users, whereas others emphasize it as a procedure bringing together many different aspects of performance and making a judgment about overall performance.
- Technology refers to repeatable processes that produce reliable results.

Assessment Programs" and other large-scale uses of tests. In the military, on the other hand, the label "assessment" is used synonymously with the term "evaluation," as in assessing the readiness of a squadron or assessing the quality of training or an after-action review. Even though technology summons up a hardware image, we prefer a broader, more general definition of technology. "Technology is a repeatable process that yields repeatable (reliable) results" (Lumsdaine, 1965). This definition encompasses systematic, validated procedures as well as software systems that depend on hardware, such as computers. So Aunt Bea's proficient sourdough bread making would qualify as a technology (same procedure, same results) as well as would an optical scanner evaluating student papers. Just because hardware is used, there is no guarantee that a technology is functioning as it should, particularly if no one can predict the outcomes (unreliability). Yet, because most educators associate the term "technology" with computer or multimedia support, we will focus primarily on human-machine types of interventions in our discussion.

Technology for What?

The technology story isn't trivial, no matter how familiar it may seem from the glut of advertising, flyers, SPAM or newspapers, each packed with the latest drawing board fantasy, new releases, or 3-month-old marked-down obsoletes. Let us choose a simple strategy. The approach we propose is to think broadly about the functions to be served by the technology, and then focus those functions on key attributes of the accountability system. In reviewing choices, weigh the importance of the function compared to cost: Are you buying efficiency, quality, or both? Where should investments be made?

Core Functions of Technology: Efficiency and Quality

Technology today can do a wide variety of things, and what it will do in 10 years is unfathomable. Nonetheless, even without knowing exactly which technologies will be available, portable, inexpensive and adaptable as they might be, we can sort potential contributions into two separate categories: efficiency and quality. *Efficiency* focuses on "how" to make the operations (in the present topic, of education in an accountability system) work better and cheaper. Efficiency meets existing goals faster, or at less cost, or both. For instance, there is considerable evidence (Fletcher, 2003) that technology-based instructional systems meet the same level of attainment at less cost or in less time compared to a variety of other options. There are systems that can separate (disaggregate) or summarize the performance of different groups in literally one second, simplifying and making more accurate clerical activities that used to require weeks of attention (Baker, 1999; Heritage, 2002; Mitchell & Lee, 2000). A commonly experienced example is the use of computer-assisted scoring of student tests, where hundreds of thousands of tests can be scored and reported using technology, a technology more than 50 years old, and a product of Lindquist and his colleagues at Iowa (1951). You get the idea.

Cost is at the heart of efficiency, and the difficulty of cost-benefit studies on the cost side is to figure what to count and how long to count it. With a short life span, technology can present a real problem here.

A second core function of technology is its ability to expand *quality* or add value to the goals and processes under development, or to add important new goals. Without question, it is easier to develop and apply systems that attempt to make current practices more efficient than those that try to add goals or enhance value, by doing something harder, more challenging, or different. Adding quality means adding new, useful knowledge to the system. Consider the following example of the now-familiar idea of tools to prepare text. Word processing was thought to be an efficiency tool, a time saver, and at the outset, it was, as it greatly sped up the time required for correction and revision of papers (does anyone still remember the powdery white tapes used to correct typographical errors?). But word processing quickly shifted from an efficiency tool to one that promoted improved quality. First, it supported major revisions of structures of texts, rather than only correcting spelling errors. Second, its fonts and graphical approaches for emphasis (font size and attributes, color) required the writer to think about the message more systematically and from the perspective of the reader. The addition of schematics, pictures, and motion, when used to illustrate the text rather than to distract or amuse the reader, can underscore memorable ideas and concepts. Word processing, except for the quick note here and there, has vastly raised the standard of quality expected in written communication and has provided tools to add more subtle goals to basic communicative intentions.

How else does technology, in general, affect quality? One important way is that technology improves information use by changing the rules of *access.* Through the Internet, mobile phones, and simulations, to choose three very different initial technologies, users have greater access to a wider range of information and so are able to act on the information more quickly. Both quality (more information to choose from) and efficiency are served. Efficiency also propels convergence. For instance, mobile phones, originally intended to provide "anywhere" voice access, have become a preferred platform for data sharing, and other information exchange through text and gaming. Technologies almost always converge, integrate and attempt to be seamless, efficiently requiring only one device rather than many to perform multiple functions, and as we have seen, at rapidly reduced costs. The obvious question about the investment in quality is where it should be made. It is clear that many levels (for example, policy, administration, management, teaching, and learning) could profit from quality investments. But in an era of fiscal restraint, investments must be made where they can do the most good, and that means where the distance is short between the technology and the acts of learning and teaching. While efficiency measures can be well argued for all levels of operation, quality (adding knowledge) needs to affect the learning process.

In Figure 3, we have roughly depicted the present distribution of functions of technology in accountability systems. By far the greatest use of technology is to serve the efficiency function: communication by e-mail, Web sites, data storage and statistical manipulations, and distributing record-keeping and analytical functions, using statistical packages and spreadsheets. There are also efficiencies attributable to search and comparison functions, where parents, teachers, and students can access information, especially on topics relevant to their study. While the access to digital material is faster, browser technology also adds value by suggesting options that

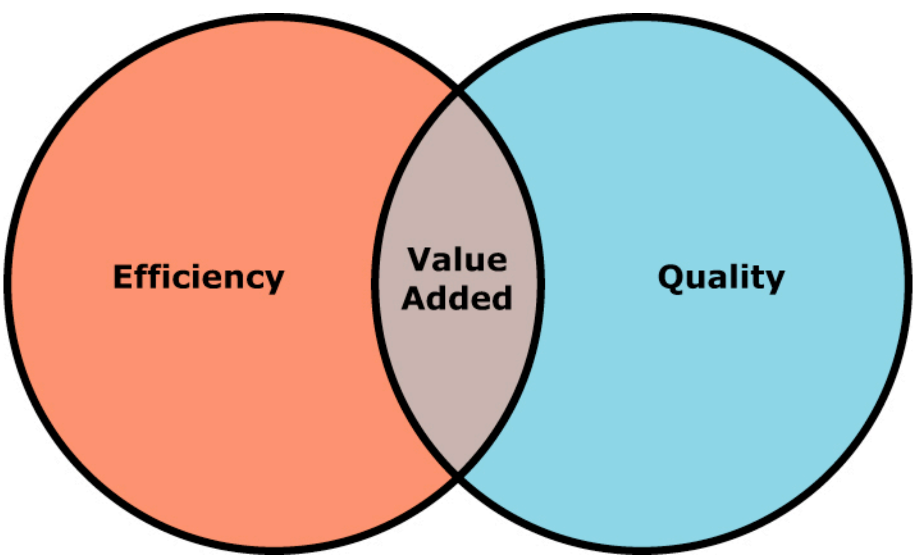

Figure 3. Present relationship of efficiency and quality functions.

might otherwise go unexplored. As efficiency functions become more frequent, quality will begin to improve as a side effect (more time to do a better job, more time to do new or more challenging tasks). Over time, the design of technology itself will focus more on qualitative improvements in performance than merely speed and ease of processing.

Figure 4 shows a relationship where the overlap between quality and efficiency is more extended. Such a future would require technology to be used to improve the quality of expectations, the quality of instructional processes, the quality of professional development, and the quality and credibility of measures used to assess and steer the system. In this vision, technology is not relegated to a mechanical role—its efficiency functions are foregone conclusions—despite the fact that much of education has taken less advantage of even limited use of technology than other sectors have. In the area of achievement testing, for instance, multimedia and other computer testing systems may offer students opportunities to display complex and challenging performance and allow the performance to be recorded and summarized. As a consequence, the substance of educational quality would be affected by raising achievement expectations, through providing tools to permit student exploration and display of challenging performance. One way to think about qualitative improvement is to imagine the *gaps* that might be filled with technology (for instance, practicing important but costly complex problem-solving tasks, like chemistry experiments), or to focus on goals not as yet articulated but

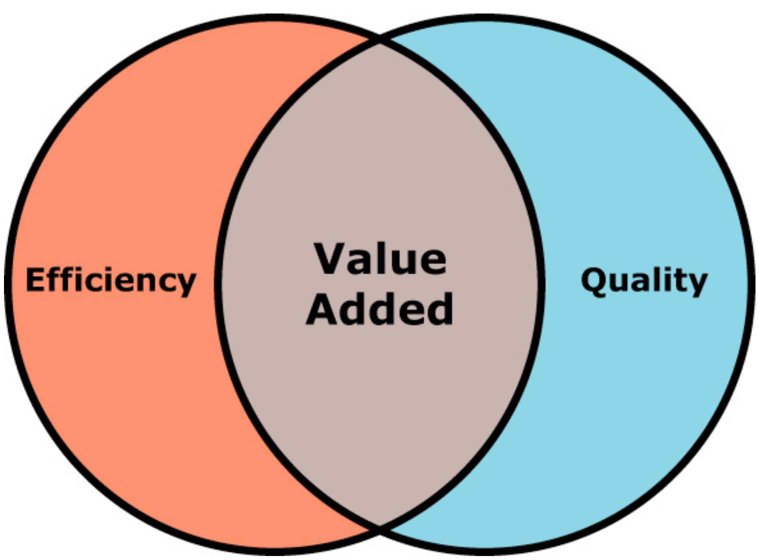

Figure 4. Desired relationship of efficiency and quality functions.

desirable and possible—such as the development of multitasked teamwork focusing on engaging teachers as a team to solve school problems.

So far we have discussed core components of accountability and given a classification system of technology functions involving efficiency and quality improvement. The next section will focus on accountability functions, relevant technological support and state of development.

Criteria for Adding Value to Education Through Technology

Clearly the core criteria for a system of accountability apply, whether the system is technology supported or depends upon papyrus. The three key criteria are validity, accuracy, and utility. *Validity* refers to the degree to which the system produces information that leads to accurate inferences, a standard requirement for measurement systems (see *Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In accountability, however, the use of information is intended to improve the system. Thus, higher test scores achieved by measurement of the intended learning domains would have greater utility for improvement than improved scores yielded by a system that responded largely to test practice or item format preparation. Validity includes a number of concerns (see Linn, Baker, & Dunbar, 1991) for the complexity of what is intended to be improved. For instance, a valid system would not trade off improvement in test scores for less prepared teachers. Nor would a valid system assess or represent performance in a way that was unfair to groups of different economic or linguistic backgrounds. Validity is sometimes assumed to be present for all the purposes a system serves, but it needs to be checked consistently—in particular for technology-based systems that use proprietary algorithms but document the evidence that their practices and cut scores (for instance, differentiating between levels) relate to the predictions or the performance domains intended.

A second criterion, *accuracy*, addresses whether the information is correct and includes the idea of reliability or dependability. As a simple example, scores that vary unpredictably from time to time cannot be accurate. Accuracy may be supported by use of computer adaptive testing (CAT), which provides highly accurate individual scores, usually in a more efficient manner. To date, most systems have been dependent upon selective responses (like multiple choice), but that is not a necessary limitation.

Some might argue that the most important criterion is *utility*. The information generated by the system must be useful to someone. In addition to the validity and accuracy concepts, utility demands that the information be accessible, timely, related to options that can be taken in the system, and have sufficient credibility to be taken seriously by users.

Functions Served by Technology in Accountability Systems

To be successful, accountability systems need help to meet validity, accuracy and utility criteria for the information that they collect and report. And this is where technology can be valuable. The approach taken in this section attempts to illustrate with apt examples how technology (both presently available and on the way) can simplify and improve accountability systems, specifically by addressing the quality of the measures and supporting decision makers' abilities to draw useful inferences and act upon findings.

Technology and accountability can be arrayed by function or by level of use. Our first cut will be functional. In Table 4, we have listed general functions required in accountability systems of any sort.

This set of functions is required by accountability systems but is not unique to them. In fact, Spector (J. Michael Spector, personal communication, May 24, 2003) cites the importance of the new generation of knowledge management systems—including functions such as communication, shared digital workplaces and controls (see also Spector & Edmonds, 2002). These functions emphasize efficiency, but also have the potential to add quality. The list in Table 5 augments general accountability options with those directly focused on educational dimensions, assuming an educational accountability model that represents and iterates targets, inputs, processes, measurement, feedback, and improvement trials.

Table 4

Technology-Aided Accountability Functions

- Store and organize information.
- Generate general reports from databases.
- Perform computations and other analyses to summarize data and make relevant comparisons.
- Provide access to information and means to cross-check its accuracy.
- Encourage interaction and communication to solve problems in common.
- Monitor costs.

The 12 requirements in Table 5 can be combined for discussion. Items 1, 2, 3, 4, and 11 focus on improving the measurement and information base used in decisions by key participants. Items 5, 6, 7, 8, and 12 deal with access to and usefulness of information for users at all levels inside the system and for policymakers and the public in their stewardship roles. Items 9 and 10 use technology to model, provide or manage instruction. Let us consider the measurement cluster and the access and utility cluster for particular examples to explore the efficiency or quality options that are now available and affordable and what is coming on the horizon. We will use a problem-solution format.

Criteria Applied to Technology in Support of Assessment and Testing

Validity is one of the greatest problems in current assessment systems—that is, the direct and clear linkage from what students are learning to what is being measured by the tests that count at the state level. While certain states, for example Nebraska, Wyoming, and Maine (see Marion, 2003; Rochewski, 2003; Rosenblum & Rolfe, 2003), are exploring approaches that use aspects of classroom-level assessments as the data sources for accountability for No Child Left Behind (NCLB; 2002), most assessment systems are designed and administered under state auspices. The key problem addressed by technology is one of alignment, specifically linking large-scale assessment with classroom assessment models. Technology approaches that might serve this problem include the assembly of carefully reviewed (and

Table 5

Specific Educational Requirements Served by Technology

1. Support the design, implementation, and use of testing and instructional systems.
2. Administer and score tests and assessments.
3. Collect and authenticate data on other indicators, such as demographics, processes, resources.
4. Provide multiple, updatable benchmarks for individual and organizational performance.
5. Create individual records of students' or schools' longitudinal performance.
6. Merge databases to allow exploration of relationships.
7. Support individual or collaborative queries of data.
8. Provide fixed or adaptive reports tailored to topics and issues (schools meeting targets for NCLB [No Child Left Behind Act of 2001], for instance).
9. Provide access to strategies and best practices.
10. Provide instruction and learning activities to students and teachers.
11. Provide certification of professional development.
12. Enable learning by all constituencies in a system and monitor and report progress.

technically appropriate) tasks and items into a bank that can be accessed by teachers or by students. The bank can serve to compensate for standards that are not fully tested by the large-scale assessment system (especially those that might take longer than the usual time-constrained test administration period). On the other hand, the bank could provide comparable items intended to measure the key outcomes assessed on the large-scale measure, serving to provide relevant practice in important knowledge and skills (see the argument for key standards by the Commission on Instructionally Supportive Assessments, 2002).

Another approach based on technology is one undertaken at CRESST, which operationalizes models of learning in templates that assist less skilled assessment designers (or those with limited time, like teachers) to create assessments that tap standards and are useful to provide both large-scale and classroom-level information (Baker, 2003; Baker & Niemi, 2001; Niemi & Baker, 2002). Explorations of the use of templates to design either paper-based assessments (Niemi & Baker, 1998; Niemi, Sylvester, & Baker, 1998; Waltman & Baker, 1997) or computer-administered tests (Chung, O'Neil, & Herl, 1999; O'Neil, Chung, & Brown, 1997) have been undertaken. The models that link different levels of assessment are based on learning and include skills such as problem solving, metacognition (planning, using feedback), teamwork, communication, and knowledge comprehension. Each of these models is then connected to the specific content or subject matter of interest. The benefits of this approach include the alignment of assessments at different levels, the support of transfer, and cost savings. The hope is that common models will allow, in the future, the aggregation of classroom assessments to provide an additional indicator of system performance. On the horizon is an approach that will provide greater flexibility to designers, and which is specifically being tried in a distance learning environment, in an attempt to create a set of computer objects that can be used to design complex problem-solving tasks (Baker & Chung, 2002). Such systems attempt to address validity, accuracy and utility simultaneously. In addition, authoring systems are under development to assist in the design of complex simulations of cognitive and psychomotor tasks (Munro, Breaux, Patrey, & Sheldon, 2002).

In another testing area, the benefits of diagnostic testing (Black, Harrison, Lee, Marshall, & Wiliam, 2002) are being developed in a number of venues. One of the most interesting, by Minstrell (2003), emphasizes the presentation of partial learning, as well as misconceptions, as a basis of diagnosis and improvement. For such systems to be useful, we believe they should focus specifically on learning, through their design and through the feedback they provide. They also need to connect to teachers' ways of thinking about assessment, and model how to understand student performance more deeply, as well as provide means for filling any subject matter gaps teachers themselves might have.

Finally, the use of computer support to aid in the scoring of open-ended performance, such as essays, is now available, using various models based on human raters, linguistic rules or a combination of factors (Chung & Baker, 2003). These approaches are feasible and in widespread use now, but may unfortunately lack the credibility needed for rapid implementation in precollegiate education. Nonetheless, the use of expert models for automatically scoring graphical representations of student work can be inexpensively and confidently employed at this point (Chung, Baker, & Cheak, 2001; Herl, O'Neil, Chung, & Schacter, 1999). The future of measurement will depend largely on how quickly and efficiently improvements can be implemented and adopted. In any case, online assessment will allow greater understanding of different learning trajectories of individual students. When instruction and testing occur, online or offline or in some combination, the key validity question to be solved is whether the assessment performance can be shown to detect differences in instruction and to support transfer of learning to new situations (Baker, 2003).

Criteria Applied to Technology in Support of Information Exchange

A second important area of technology support for accountability is providing access to data for those interested in improving the system or monitoring progress. Of great value is the creation of systems that can array an individual student's record over a number of data points, including systems that can help keep track of progress within a classroom, as well as across a student's career. Decision support systems, built on various software models, have been available for more than 30 years. However, they have only recently been created so that they support large numbers, very large sets of variables and flexible reporting. The rapid advance of hardware and software in support of such information systems has made it possible for individual organizations, schools, and classroom-level teachers to merge data from various sources, and more importantly, add to databases so that the data set reflects both external (such as state) and local (such as school or classroom) interests. Digitally supported grade books—archives of student work, from essays to mathematics problems, research studies or piano performance—can be stored and used to draw inferences about performance or to help students and their families to understand progress and focus on areas to improve.

One example that we know well, the Quality School Portfolio (QSP © 2002), permits flexible questions and immediate answers about the performance of individuals and subgroups, and various comparisons. QSP, among other systems, supports the disaggregation and progress monitoring of significant subgroups as required by NCLB.

These systems have four areas of difficulty to address. One is the shift in culture that makes the use of information a seamless part of everyday reality (the notion of learning organizations, or six sigma improvement). When teachers are used to thinking about instructional sequence from their own, chronological point of view, it is difficult to find a way to bring in the need for continuous attention to performance and required modifications of the instructional sequence. The second area of difficulty is to prevent inappropriate inferences from being made from the data, especially because most questions will be addressed not by carefully controlled studies but by drawing inferences from a stream of school-based experiences and out-of-school experiences. Thus, it might be easy to attribute growth to a particular practice, but not have a procedure for verifying that the inference is correct. The best option, in addition to training, is to support information networks, where individuals can share inferences and ask their colleagues with particular expertise for advice. Technology-supported networks have long been a feature of higher education and business and provide the core of social capital needed to move an organization forward (Hargreaves, 2003).

Visual display is the third problem area for decision support systems, where, in fact, considerable progress has been made, both in research and in technology supports. Graphics and pictures can be used to show progress, and tools are available to allow teachers, students, and administrators easily to make their own pictures or other displays, following the metaphor that they prefer. Such approaches improve the utility of data, as their presentation is not encumbered by the usual tables, statistical information, and so on. It is important, however, that any display show clearly the degree of confidence (or, on the other hand, uncertainty) around group performance. All data users must realize that measurement and other accountability indicators can never be fully accurate and will vary from administration to administration. Oddly enough, some technological advances, such as animation, have been shown to have negative rather than positive impacts on performance (Mayer, 2001).

The fourth area, of course, is the utility of the data in addressing the problem at hand. Dexter Fletcher (personal communication, May 28, 2003) reminds us that if data systems are "garbage-in," then no matter how nifty their interface, we must be resigned to garbage, or garbage plus no outputs.

A number of providers have Web-based systems (for example, Standard and Poor's School Evaluation Services [www.ses.standardandpoors.com]; JustfortheKids [www.just4kids.org/us/us_home.asp]; The Grow Network [www.grownetwork.com]), some of which provide limited query capacity to systems, but focus on the comparison of individual schools, systems, and programs.

These and other systems provide advice or prescriptions about ways to modify and improve instruction with different levels of confidence that such suggestions will result in greater performance. Some systems (for example, some state and school district report cards) provide comparative data and display progress against goals on important indicators. Access to such information is usually available on the Web, and given the merging of technologies (cell phones and networks) it is likely that greater access will be available to groups heretofore on the wrong side of the technology divide.

Ways to Manage System, Infrastructure, and Capacity Demands

What are the practical matters that control the advance of technology in support of accountability? What gets in the way of improving efficiency and quality and adding value? Four criteria will be discussed, although the reader should be aware that multiple technology organizations have created well-designed technical standards (for examples, use your browser to search on ISTE, IEEE or SCORM). The criteria to be briefly addressed here are security and privacy, compatibility, resilience, and cost benefit.

Security and *privacy* are key concerns in data systems of any sort. Security needs to be maintained to ensure that records remain accurate and are not accessed by unauthorized personnel. Security provisions are constantly under development and respond, as they must, to ingenious and malevolent incursions by unintended users. One important aspect of security for multi-user data systems, of the sort that are used by schools, is to find ways to provide security for the core data systems, so they remain uncontaminated, while allowing local users to add or modify data to support their own inquiry and improvement. Security relates to validity, accuracy, and utility directly. Privacy is a related and important issue, protected by the law and ethics about individuals' rights to control access to personal information. Privacy is enhanced by security but not completely controlled by such approaches. Questions about informed consent now related to research may very well be more prevalent in the future, to ensure that the use of information about individuals meets our society's expectations and that we have not created either Big Brother or The Matrix.

The second criterion, *compatibility*, sometimes called by terms such as interoperability, addresses the requirement that systems work well with various hardware platforms and software systems. Remember the oft-used but still scary example of VHS and Beta formats in video tape and equipment. Compatibility means that the predictability of one system is not influenced by the involvement of any other. In schools, there is a standard called SIF (School Interoperability Framework), which represents an expectation that software should currently meet. Compatibility is an efficiency concern, as well as one of utility. There are alternative positions on how to achieve compatibility and interoperability at low cost. One approach is to use a particular software/equipment platform. Another is to support open-source systems, where software is not proprietary. Approaches to improve interoperability mean that software must meet particular standards, and those standards may deal with only the way variables are coded or might involve substantive requirements intended to support the sharing of instructional content and other material.

The term *resilience* is a criterion of great importance where systems may need to be used for a longer period because of cost constraints. In any case, it is a good idea. Resilience means that changes in hardware and software do not undermine the functioning of the information. One way to think about the concept is to imagine having a system and then upgrading a component. Does the new component have downward compatibility? Will the parts of the system not upgraded still work? Without a doubt, the cost of replacing equipment and software is affected by such features. A second aspect of resilience is the degree to which the system and its components experience errors that either were never addressed in the software or emerge as a function of the situations of use. Errors in software obviously affect validity, accuracy, and utility. They also add operational costs, downtime, and user frustration. While there are never protections against errors in software, the kind of technical support available makes a difference. If permanent staff are needed to maintain systems, as they well might be, it is important to calculate their costs in acquisition. If the vendor is to supply maintenance, contracts need to specify, to the degree possible, the type of service expected and the speed with which it will be provided.

Resilience also refers to the degree to which a system (or equipment) can withstand and support everyday use by the range of users in the system. In education, that very well may mean children or others who may not be aware of particular system vulnerabilities. Thus, part of resilience is simply how sturdy a system is—does it keep on ticking when "wrong" moves are made? Do parts break and need replacing? Another aspect relates to the utility of design for the purposes to be served. Are interfaces intuitive and sensible? Are training requirements reasonable? The good news on system hardiness is that the industry is making generally good progress, although there are still enough horror stories around (personal digital assistants that lose data when a battery fails—thus, if one is used by teachers to collect classroom data, be sure battery warnings are heeded).

The last criterion will not take much time. The issue of *cost* must be raised, and technology needs to be treated as an ongoing expense rather than a capital acquisition. In general, investing in one, large, all-purpose system has a number of risks. If systems can be distributed and compatible, risk can be contained. The message about technology and cost is straightforward, however. Technology is not an option that should be compared and evaluated. It is a utility, like telephones and lighting. It can be designed to support important educational functions, if the market is given clear signals by educators. At this point, however, the education market has not been given clear signals, and available options for testing, accountability, and instruction may be put together by entrepreneurs with little understanding of the settings or people that they must work with. Educational policy leaders need to provide a strong signal to the marketplace, so that technology tomorrow will creatively support teaching, learning and performance.

Summary

This section did not touch on teaching and learning systems designed to help students acquire skills. It did not discuss distance learning. It was mute on digital records and libraries. It skipped high-end stuff like models of nanotechnology, expert systems to schedule and relieve routine tasks, sensor-based performance monitoring, and intelligent tutoring systems. It said nothing about game and motivational environments or the appetite created by the use of games (now exceeding television and school in some locales in students' hours per week). The field of robotics was omitted. The fact is that all of these weird things will play a part sooner than we think.

The main message here is to use technology to add value, to make school better, learning deeper, and teachers more effective. Its core recommendation is to invest closer to the action, to the end user (the teacher and the student), rather than think of assessment and accountability systems as top-down management and monitoring systems. Technology supports individuals and teams, it supports decentralization and initiative. It needs to have enough smarts so that validity, alignment and sensible data use are enabled substantively. It needs to add real and new knowledge to the system. Educational policy leaders may need to redefine some of their role as knowledge managers and seekers. Quality happens in no other way.

Part II: Technology in Schools: Approaching Real Life

The following section provides examples of how to encourage the use of technology to support accountability in school district settings. Two different scenarios will be presented—one in which technology is used principally to optimize efficiency, and another in which its use is directed toward efficiency as well as quality improvement. Then, as a final reprise, rules of thumb will be outlined for expanding or redesigning the use of technology in schools to support reform.

To address the practical considerations of using technology as an approach to support accountability, for each scenario we begin with certain assumptions and move to set a limited group of plausible, measurable goals. Because readers will manage on a continuum of large to small districts, with a range from being well-equipped to having barely serviceable technology, we use the assumptions as a means to provide context for some sort of a practical baseline. Readers can use the baseline for general guidance, with the recognition that the scenarios attempt to show districts with different sets of strengths and weaknesses.

Scenario—Assumptions at Lincoln School District

A. Technology at Lincoln School District (70 schools)

- Technology is present but not omnipresent—there is a 1:12 ratio of computers to students, but not all equipment is up-to-date; computers do not uniformly connect with one another, and some use slow technology for Internet connection.

- Some classrooms do not have Internet connections.
- Technology is routinely used for bureaucratic jobs by clerks (attendance, download of state assessment data, bus schedules).
- There is technology in instructional use for students, but the degree to which it is actually integrated in curriculum varies.
- Teachers differ in the degree to which they know how or are willing to use technology in school, although their out-of-school use may be comparatively high, for example, games, e-mail, e-commerce.
- The school provides minimum technology support to individual teachers.
- Technology is treated as a capital acquisition by the board.
- A large proportion of the technology is aging.

B. The policy context at Lincoln District

- Accountability and testing are high on the policy agenda, with state standards and assessments, target specification, and annual performance receiving strong media attention.
- An achievement gap exists among subgroups related in part to the economic status of students.
- The diversity of the student population is increasing, but not equally at every school.
- Substantial numbers of senior teachers are retiring soon.
- Budget restrictions are serious and look to be getting worse.
- Vendors (stimulated by "underperforming school" dollars) claim to have effective packages to sell to relieve achievement problems.
- Politics are more apparent in school management than before.

C. Goal framework for Lincoln District. In the very basic situation we have posited, the investment in technology must be one that shows an obvious return on investment, in terms of reduced time, increased effectiveness, and strengthened morale. The goal of technology use under these conditions should be predominately to (a) sensitize and encourage technology use by teachers in a productive manner and (b) improve teachers' efficiency in understanding student performance, with an early spill-over effect on performance quality. Specifically, in order to gain efficiency outcomes, equipment and support software for teachers and staff should allow them regularly to

1. monitor students' progress on external and internal standards of performance;
2. identify subgroups of students in need of special attention (either help or enrichment);
3. provide a strategy for communicating with parents at least four times a year and in understandable detail about their child's performance;
4. communicate among teachers and leadership within a school and among schools (for example, feeder patterns), with continual refresher training for new personnel; and
5. focus student attention on routine (word processing) and nonroutine (science Web projects) uses of technology in order to yield convincing examples of student work. This is necessary to communicate to the public about the District's currency in the technology world.

D. Measures. Measures of efficiency should be planned, communicated, and obtained. They include time saved in performing routine analyses; number of opportunities to review with students, either independently or with their parents, their performance; number of communications home; and e-mail traffic with other faculty. Emerging measures of quality include teachers' perceptions of a cooperative and sharing environment.

E. Phasing acquisition, implementation, and training. Phasing is a key element of any technology use plan. First, there is always something desirable and new. Second, new software and platforms require acquisition, installation, maintenance, and training costs. Depending upon budget sources, such as categorical funds or the extant technology renewal plan, it would be ideal to select something like 10% of the schools for targeted upgrading of technology directed to the accountability and efficiency goals above. Within the 10%, it is possible to select a subset of teachers for technology use. A criterion should be teachers that feed and receive students in sequence (Grades 2-3-4, or 3-4-5) in order to foster discussion and a collaborative learning group. A second question is what technology is needed? For the purpose of understanding and analyzing student work, minimum requirements suggested are an Internet-connected computer for every three teachers, although separate access on a common server would be desirable.

The computer for the teacher should have graphics capability and a printer, and be connected to a local area network within the school. Minimum software includes a calendar, e-mail accounts, word processing and spreadsheet software, video capability and state-of-the-art browsers. The computer should accept CD-ROM inputs, be able to create CDs, and probably have an auxiliary such as a Zip drive. Connection should be as fast as possible, T-1, satellite or digital cable. Desirable, additional equipment could include infra-red enabled personal data assistants for downloading information into the computer, large displays and robust notebooks for out-of-school data collection. Teacher computers will need to interface with both administrative and student computers. If mixing PCs and Macs, this can be accomplished through the use of appropriate software. Technical assistance will require one full-time person for approximately 100 computers, but can be augmented by using proficient students.

Specialized software should include a decision support system (like QSP) and graphical report card options.

- First, acquisition and training should begin with e-mail (although most teachers may have personal accounts already).
- Categories are to be devised by local teacher teams for e-mail correspondence sorting and processing—instructional queries, student advising, etc.
- Some time will need to be set aside, either voluntary or compensated, to allow teachers from the same and comparable schools to plan for the use of software to understand student performance.
- Questions about how to improve performance of students should start with a comprehensive analysis of where students are, within grade ranges and within the individual schools. The work should be conducted on real student datasets that are relevant to the teachers in training. This work will be facilitated with decision support software, where individual student records will be available, including all tests and computerized records.

These records can be summarized, combined and accessed using simple interfaces. Teacher teams should develop questions about how participation in particular programs, experience in certain courses, or attendance patterns relate to or predict performance on within-school and external examinations (such as state examinations). Through meetings and e-mail follow-up, teachers should create and try out plans and compare results for improving one targeted underperforming group.

- Collaboration across classes would be desirable. Incentives by the administration should be provided for teacher leadership as a way of widening the circle of participants.
- University or industry speakers could be invited at little or no cost to keep teachers abreast of new developments they may wish to try.

Following the ability to manipulate e-mail and the decision support software, teachers should identify those areas in which they have fewest resources (for example, life science), or most interest (for example, celebrities) and devise projects requiring the use of Internet access for students. Whether access is through computers in their own classroom (8-10 most desirable) or in a lab, students will need to create plans, manage their access, work in teams and produce a product to be shared with classmates and parents. These products can be stored and recalled for use with the decision support software.

F. Management. The administration must make a plan for the expansion, upgrading and replacement of equipment and software. The plan needs to include the extension of support to schools and classrooms not presently involved in the plan as well as the replacement and upgrading of equipment and connectivity. The acquisition of gee-whiz tools needs to be supported by external grant support or by a well-conceived plan that links use to improved learning. Standards of acquisition should require a link to an instructional plan tied to learning reform.

A similar plan needs to be developed, systemwide, for the acquisition and updating of student computer support. Transition costs to full Web-environments will need to be calculated (for cost-benefit analyses, see The Mellon Foundation, Cost-Effective Use of Technology in Teaching [CEUTT] Program at www.ceutt.org). Reviews of current protections (viruses and encryption) will need to be regularly conducted. Target and real costs need to be computed and managed.

Scenario—Assumptions at Matthew Elementary School District (10 Schools)

A. Technology assumptions

- The district has been involved in technology use for a considerable period.
- Technology support in the schools involves computers in each classroom and stand-alone computers for the teachers. Each classroom has one relatively fast (ISDN) connection.
- Most technology use has been focused on classroom projects (student exploration of tools) and the use of packages designed to teach young students pattern recognition, preliminary drawing and skills practice.
- Teachers are eager, in general, to use technology, and there is a cluster (2-3) of specially interested teachers at each school who often serve to review software packages for the school.
- A few teachers have asked for decision support software, after seeing demonstrations of the software.

B. Policy assumptions

- The district is relatively small, but new developments and economic upheavals are changing the population of students.
- The schools do reasonably well on external examinations, and there has been reasonable stability in the staff and leadership.
- Some teachers and a number of parents are unhappy with the emphasis on external tests and wish to have more curriculum-related measures developed.
- The school leadership would like to see how learning proceeds longitudinally, as well as the outcomes that are reached.
- The district hopes to use "scientifically valid" approaches to make instructional decisions regarding programs and practices.

C. Goal framework for Matthew Elementary District. The district wants to implement technology to assist in the improvement of quality, but wants to retain a substantial focus on teacher decision making rather than adopting turn-key programs. The district is interested in supporting teachers' and parents' inquiry into the impact of top-down accountability, as well as creating additional capacity to conduct formative, in-class assessments. In general, this district is at least on the midpoint of the continuum of technology use. The major goals of the technology system will be

1. to advance the utility of decision support software showing students' progress;
2. to develop teachers' capacity to create, evaluate, and interpret measures of student progress in areas important to standards;
3. to conduct action research on the effectiveness of alternative instructional approaches, particularly for less well performing students; and
4. to involve students in the management of their own growth.

D. Measures. Measures of the effectiveness of the technology interventions will include (a) the numbers of teachers participating, frequency of use, and satisfaction of teachers and parents with efforts to communicate student progress; (b) that 50% of the teachers will attempt to use software designed to assist them in the development of formative and summative assessments of students; (c) that teachers will compare the findings on their measures with those from external examinations; (d) that teachers will begin to plan in terms of learning trajectories rather than simple targets; (e) that efforts will be made in each school to conduct one study using student performance outcomes to decide on the utility of an instructional intervention; (f) that at the highest grade (5th or 6th) students will use assessment authoring systems to create peer and self-tests, with the outcome that students will report being more responsible for their own learning.

E. Phasing, acquisition, and training. Acquisition of technology platforms will be made according to a renewal plan that will include basic services, connectivity, decision support software, test authoring software, and Web-based training. Teachers will meet quarterly to present their analyses of student progress to one another and to seek input about their interpretations. Teachers will provide parents with hard-copy or e-mail reports showing student progress on multiple measures (some teacher-developed), as well as an electronic or hard-copy version of student work, against an exemplar. These may be Web-based, and computers could be made available for parents' use, with appropriate privacy protections, to compare an individual student with various comparison groups.

Teachers will add in to the decision support system their own grades and evaluations of students, including exemplars and scoring schemes. These exemplars (with names deleted) and scoring schemes will be published on a teacher-network Web site for commentary by other teachers. Teachers will use the system to plot growth of each student against various standards or targets.

Teachers who volunteer will be given access to cognitively oriented assessment authoring systems. The system selected will first require the teacher choose a kind of learning (for example, problem solving), and will present the teacher with "smart" menus for designing complex tasks involving problem identification, strategy, verification, and so on across different subject matters (for example, sciences, history, mathematics). The system will allow teachers to create paper-and-pencil or mediated assessments, and scoring will be either by hand, using explicit or locally modified rubrics, or by computer automated systems. Students' performance on a series of such tasks will be compared with their external examination scores, and different patterns of growth will suggest alternative instructional options. Teachers will give students the opportunity to design measures using the authoring system. Students (and some teachers) may need additional subject matter content support, which they may obtain by (a) communicating electronically with other teachers or experts or (b) investigating Web resources. Students' tests will also be administered and discussed. School teams will review options for dealing with recalcitrant instructional problems and will experimentally consider (teachers or students assigned at random) alternative approaches. Teachers will share and report publicly the effects of alternatives on progress and outcomes of students.

F. Management. In addition to acquisition concerns, management will need to assure that assessments developed by the teachers in schools receive adequate review and support. The technology transition is moving from the use of an occasional tool to the full internalization of the importance of progress and outcomes—a culture shift. It is likely that some teachers will adopt this approach with particular ease. Such teachers should be encouraged to set up partnerships with teachers with varying skills and propensities for accountability. The collaborative work of deciding on instructional options, from type of homework to purchase of products, will provide a source of power and expertise to teachers and enable them to share, across grades, schools and districts, what they have learned. The involvement of students will need to be carefully managed to assure that sensitivity to performance is maintained and security and privacy are protected, and that students begin to acknowledge their own responsibility for managing learning.

Rules of Thumb for Designing and Using Technology in Support of Accountability

This section is presented to extract and reprise the principles underlying the scenarios. First, the scenarios are much less ambitious than the range of options discussed in the attached paper. They illustrate a set of principles.

- Do only one or two things and focus on them.
  This approach contrasts with some technology strategies that involve exploring alternatives simultaneously.

- Start with outcomes and understanding their meaning.
  If one starts with instructional options, using technology, it is possible never to reach the outcome stage.

- Make sure technology is available for needs, but don't overload the thinking systems of teachers.
  Teachers already have lots to do, so keeping it simple is a good idea.

- Find natural leaders and use the technology to build on their expertise.
  E-mail and bulletin boards are inexpensive and useful ways to build communication and sharing.

- Build public reporting into the system (including costs and return on investment) as an early strategy.
  Responsible management behavior may well yield unexpected sources of community support, in addition to the guidance it will provide in acquisition, maintenance, and training needs.

- Move teachers into a position of acknowledged expertise and accountability in the measurement of outcomes.
  Make sure teachers feel like actors, and not acted upon, by expanding their capacity to perform central tasks and building the culture of important results.

- Use teacher-made measures seriously to understand growth patterns and to select alternative instructional options.
  Although technical quality may vary (depending upon the design of test authoring systems), some data-based decisions will be better than a good deal of typical evaluation of materials or progress.

- Plan to increment and expand.
  The technology world will be throwing magical looking objects at schooling. We need to plan how we are going to manage the push of technology and the pull of accountability requirements.

Summary

This paper addresses accountability and the role emerging technologies can play in improving education. The connection between accountability and technology can be best summarized in one term: *feedback.* Technological supports provide ways of designing, collecting and sharing information in order to provide the basis for the improvement of systems and outcomes. Because technological options are so varied and the future of their development unpredictable, this paper argues that the first screen for the review of technology is one addressing core functions the technology serves. Technologies can provide greater efficiency in managing information or serve to improve the quality of the information used to guide the system. At this point, these functions are frequently mutually exclusive, but in the future, wise selection of technology should involve a consideration of how the technology adds value by increasing quality and efficiency in the enterprise. Quality is added when more knowledge is available to the user. For choices that need to be made about investments, a set of criteria is provided to guide decision logic. Criteria include validity, accuracy, and utility, with an emphasis on users directly involved in learning (students and teachers). In addition, criteria are reviewed related to software/hardware operation, including security, backup, compatibility, resilience, and transparency. The assumption is that systematic application of criteria will influence the market to drive development of more useful systems for education at a cost that can be withstood. Examples of technology uses now and soon to be available are discussed, and two extended educational scenarios are provided to show elements of design and technology in practice in two very different settings.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Baker, E. L. (1999, Summer). *Technology: Something's coming—something good* (CRESST Policy Brief 2). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L. (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues & Practice, 22*(2), 13-17.

Baker, E. L. (2003, January). *From usable to useful assessment knowledge: A design problem*. Paper presented at the Keynote Session: "Usable Assessment Knowledge: A Design Problem" for the International Congress for School Effectiveness and Improvement, Sydney, Australia.

Baker, E. L. (2003, January). *Model-based assessment: Why, what, how, how good, and what next?* Presented at the National Research Council Workshop "Bridging the Gap Between Classroom and Large-Scale Assessment," Washington, DC.

Baker, E. L., & Chung, G. K. W. K. (2002, September). *Model-based assessment.* Presentation to the Office of Naval Research, ONR-NETC Meeting, University of California, Los Angeles.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability* (pp. 47-72). New York: Teachers College Press.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). From the directors: Standards for educational accountability systems. *The CRESST Line*, 1-4.

Baker, E. L., & Niemi, D. (2001). *Assessments to support the transition to complex learning in science* (Proposal to the National Science Foundation). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom.* London, UK: King's College London Department of Education and Professional Studies. (Copies available for order from the Publications Secretary, Department of Education & Professional Studies, Franklin-Wilkins Building, Waterloo Rd, London, SE1 9NN Tel: 020 7848 3189.)

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein, (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23-40). Mahwah. NJ: Erlbaum.

Chung, G. K. W. K., Baker, E. L., & Cheak, A. M. (2001). *Knowledge mapper authoring system prototype* (Final deliverable to OERI). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., O'Neil, H. F., Jr., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15*, 463-494.

Commission on Instructionally Supportive Assessments. (2002, October). *Building tests to support instruction and accountability. A guide for policymakers* (Panel convened by the American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, and National Middle School Association). Retrieved January 30, 2003, from www.aasa.org/issues_and_insights/assessment/Building_Tests.pdf

Fletcher, J. D. (2003) Evidence for learning from technology-assisted instruction. In H. F. O'Neil, Jr., & R. Perez (Eds.), *Technology applications in education: A learning view* (pp. 79-99). Mahwah, NJ: Erlbaum.

Hargreaves, D. (2003, January). *From improvement to transformation.* Keynote address to the International Congress for School Effectiveness and Improvement 2003 conference "Schooling the Knowledge Society," Sydney, Australia.

Heritage, M. (2002, Winter). Quality School Portfolio—Web version. *The CRESST Line*, 3.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15*, 315-334.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, DC: American Council on Education.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21. (ERIC Document Reproduction Service No. EJ 436 999)

Lumsdaine, A. A. (1965). Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), *Teaching machines and programmed learning. II: Data and directions* (pp. 267-320). Washington, DC: National Education Association of the United States.

Mandinach, E., & Cline, H. (1994). *Classroom dynamics: Implementing a technology-based learning environment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Marion, S. (2003, January). *Wyoming Assessment System*. Presentation at the National Research Council Workshop "Bridging the Gap Between Classroom and Large-Scale Assessment," Washington, DC.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge: Cambridge University Press.

Minstrell, J. (2003, January). *Facets based assessment*. Presentation at the National Research Council Workshop "Bridging the Gap Between Classroom and Large-Scale Assessment," Washington, DC.

Mitchell, D., & Lee, J. (2000, April). QSP software and school data driven decision-making. In J. L. Herman (Chair), *School inquiry: From rhetoric to reality*. Symposium presented at the annual meeting of the American Educational Research Association, New Orleans.

Morecroft, J., & Sterman, J. (Eds.). (1994). *Modeling for learning organizations.* Portland, OR: Productivity Press.

Munro, A., Breaux, R., Patrey, J., & Sheldon, B. (2002). Cognitive aspects of virtual environments design. In K. Stanney (Ed.), *Handbook of virtual environments* (pp. 415-434). Mahwah, NJ: Lawrence Erlbaum Associates.

Nardi, B. (Ed.). (1996). *Context and consciousness: Activity theory and human-computer interaction.* Cambridge, MA: MIT Press.

Niemi, D., & Baker, E. L. (1998). *Design and development of a comprehensive assessment system: Pilot testing, scoring, and refinement of mathematics and language arts performance assessments* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Niemi, D., & Baker, E. L. (2002, November). *Assessments to support the transition to complex learning in science*. Poster presentation at IERI Principal Investigator Meeting, Alexandria, VA.

Niemi, D. M., Sylvester, R. M., & Baker, E. L. (1998). *Design and development of a comprehensive assessment system: Identification and pilot testing of performance assessments and validity studies development* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil, Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Rochewski, P. (2003, January). *Nebraska: STARS Program*. Presentation at the National Research Council Workshop "Bridging the Gap Between Classroom and Large-Scale Assessment," Washington, DC.

Rosenblum, J., & Rolfe, P. (2003, January). *Maine's Comprehensive Assessment System*. Presentation at the National Research Council Workshop "Bridging the Gap Between Classroom and Large-Scale Assessment," Washington, DC.

Spector, J. M., & Edmonds, G. (2003). Knowledge management in instructional design. *ERIC Digest EDO-IR-2002-02* (Syracuse, New York: ERIC Information Technology Clearinghouse, 2002, accessed 30 May 2003). Available 18 May 2005 from: http://www.ericdigests.org/2003-1/design.htm

Sterman, J. (1994). Learning in and about complex systems. *Systems Dynamics Review, 10,* 291-300.

Waltman, K. K., & Baker, E. L. (1997). *Design and development of a comprehensive assessment system: Summary of the 1996-1997 large-scale pilot tests, scoring sessions, and teacher feedback* (Final Deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.