Annotated Bibliography: Select Articles Regarding Learning, Performance Assessment & Educational Technology

By Laura Blasi
The University of Virginia
Curry School of Education
Center for Technology and Teacher Education
August, 2000

This annotated bibliography was written with teachers in mind, specifically for those who are trying to implement performance assessment while using educational technology.

Performance assessment is being developed at a time when the limitations of standardized, norm-referenced tests are being felt. This bibliography attempts to convey some of the findings and cautions in the current literature regarding performance assessment. We have not exhausted the literature, but instead have arranged key pieces thematically, moving from practice-based issues to policy-based issues.

The first half of this bibliography should also be of interest to instructional designers (including developers of software-based assessment and educational multimedia programs). The first eight citations outline learning strategies, explaining how they can be taught and how they can be evaluated.

Problem-based learning, reflective problem solving, and learner self-regulation, can be used within performance assessment, while integrating technology into teaching and learning. Moving from technology in teaching to technology in assessment, Kelly-Benjamin (1995) and McNabb and Smith (1998) offer insights into ways technology could help teachers develop and record performance assessments.

The second half of this bibliography offers insights into the more technical aspects of performance assessment. Linn (1994), Messick (1994), and Herman (1997) offer comprehensive overviews of performance assessment in relation to issues of validity and replicability. The authors emphasize the need for careful specification procedures, multiple samples of student performance, and attention toward discriminant evidence toward establishing validity.

The need for developmental sensitivity is raised, as authors caution against basing the performance assessments upon the student goal of mimicking adult roles and capabilities. Authors also caution against teaching students to perform, urging cosmetic expertise at the expense of the development of habits of learning, which include taking risks, acknowledging limitations, mistakes, and areas of further development. Assessments structured around “tasks,” instead of “constructs” (such as student comprehension), risk returning to behavioralistic educational approaches.
Several of the authors urge the documentation of unintended consequences stemming from the specific tasks, as processes and outcomes may vary from the construct the teacher is seeking to measure. Unintended consequences also need to be documented as they arise from the approach to assessment. A teacher using a writing prompt to indirectly measure student editing skills, for example, may begin to emphasize student practice with prompts, without teaching specific editing skills or the process of revision.

This bibliography takes a critical approach to performance assessment. Burdens felt by teachers and discomfort felt by parents are described by several of the authors. Performance assessment often calls for k-12 schools to compensate for the lack of training available for preservice teachers (in post-secondary education schools) regarding performance-related instructional methods and approaches to assessment.

In-service teachers, currently in the classroom, also work within conditions that can be adverse to performance assessments. Teachers using both standardized testing and performance assessments are often left to negotiate the potentially conflicting teaching methods and educational goals. Students are left to resolve the repercussions from the conflicts on their own. As teachers face a lack of available time, among other challenges, the k-12 school system stands to be held back by the mismatch between the norms of the U.S. teaching culture and the collaboration demanded by increasing use of performance assessment and portfolios. This mismatch will make it difficult to implement some of the approaches that increase reliability and generalizability described within this bibliography.

Within this bibliography, we have taken care to keep in mind that, as Baker and O'Neil (1996) have eloquently stated, “performance-based assessment is obviously grounded in a different instructional model, one for which the majority of teachers of disadvantaged children may be unprepared” (p. 185). As teachers at Mantua Elementary seek to differentiate instruction and assessment, to meet the needs of all of their students, they do so with the intent of using technology to meet individual learner needs.

Problem-Based Learning (PBL)


Hsu discusses the methodology of assessment for student learning within a problem-based learning approach (PBL), after reviewing evaluation theories and techniques in both medical and educational fields. The author’s theory articulates the condition variables, the method variables, and the outcome variables for student assessment within the PBL approach. First implemented in medical schools in the 1970s, this approach features students in small groups “working on a problem presented by the teacher. They need to internalize and reason through the problem to identify the learning issues and define the objectives” (Gallagher, 1997; Savery and Duffy, 1996; West, 1992; cited in Hsu, 1999, p. 199). Hsu defines the process in steps:

- Defining the problem and identifying the issues
- Framing learning objectives
- Determining work schedule
- Gathering and analyzing data
- Presenting and discussing learning findings related to learning issues
- Synthesizing and summarizing findings into solutions
- Justifying solution and evaluating performance

The PBL approach seeks to foster the following student learning goals that are elaborated upon in this paper:

- Metacognition skills
- Self-directed learning skills
- Critical thinking and problem solving skills
- Learner’s knowledge acquisition and use
- Collaborative learning
- Higher motivation and positive attitude

To evaluate student progress towards these goals, strategies have been proposed focusing on process and outcome, with several also emphasizing content (Barrow and Tamblyn, 1980; Swanson, Case, and Vlerten, 1991; Glasgow, 1996). Bridge and Hallinger (1995) are also cited, delineating four evaluation types: 1) instructor-structured and student-judged; 2) instructor-structured and instructor-judged; 3) student-structured and instructor-judged; and 4) student-structured and student-judged. Hsu draws upon
Reigeluth’s model of “methods, conditions, and outcomes” (1983) when synthesizing the research into process-oriented and outcome-oriented assessment strategies. Condition variables are threefold: 1) environment/learner characteristics; 2) subject content; and 3) time constraints. With an understanding of conditions, process-oriented evaluation methods include:

- Tutor/peer ratings
- Self-evaluation
- Unobtrusive measure (such as student journal entries, records of resources used)
- Oral examination
- Observation
- Problem simulations
- Review of case record (which is also an outcome-oriented method)

Hsu also outlines authentic assessment (Herman, Aschbacher, and Winters, 1992; Barrows, 1994) measuring student skill development focused on individual pacing and growth, without single correct answers. Performance assessment concludes the list of process-oriented evaluation methods, assessing oral presentation, written work, audiovisual materials, portfolios, etc. “Students demonstrate their knowledge as they report their thinking and problem-solving process. The skills to be assessed are creativity and communication skills” (Barrows and Myers, 1993). Outcome-oriented evaluation methods include:

- Review of case record
- Student-judged evaluation
- Multiple-choice examinations
- Short-answer or word completion examinations
- Essay examinations
- Problem simulations
- Portfolio assessment

Process-oriented methods have beneficial effects on student learning, asserts the author, while outcome-oriented methods have the ease of conducting and scoring. The author concludes by pointing out that the epistemology underlying assessment methods has shifted from behaviorism to cognitivism to constructivism, with heightened emphasis on self-reflection and on assessment within the context of learning. (Contains 25 references.)

**Evaluating Metacognition in Multimedia Instruction**


“Learners’ reported inabilities to form and test hypotheses, draw conclusions, and evaluate . . . materials to which they have been exposed” have been of concern for many
years (Franklin, 1779; Monroe, 1907; Flexner, 1917; Judd, 1933, referenced by Cates, 1992, p. 1). Cates addresses how hypermedia/multimedia instruction might enhance students’ metacognitive abilities (Flavin, 1976). He does not attempt to prove or disprove whether metacognitive activities produce higher levels of content material learning (p. 4).

He synthesizes a broad body of literature to generate specific metacognitive categories:

- Task analysis
- Goal setting
- Strategic action (regarding learning strategies)
- Cognitive load (particularly toward minimizing the memory load.)
- Persistence/ responsibility (recognizing scope and distributing effort accordingly; taking responsibility for learning) and
- Growth (metacognitive) (p. 2).

The author details specific skills within each heading, linked to learners’ use of hypermedia/multimedia products.

Instruments for measuring students’ metacognitive engagement include: 1) progress logs; 2) journal entries; 3) transaction shell data; 4) student discussion; and 5) learner-created materials. The author discusses approaches for developing each skill, and types of data that would support evidence of operation of each skill.

Noting that cognitive load is based on the limits of short-term memory (Miller, 1956; Fleming & Le Vie, 1978), overload becomes more probable with students absorbing frequently revised and restructured information. Developers of multimedia packages threaten to overload students at the same time they are seeking to “empow 1980; Morariu, 1988; Gygi, 1990; Heller, 1990; and Oren, Salomon, Kreitman, & Donn, 1990). Students, therefore, need efficient and efficacious strategies (Charp, 1986; White, 1988).

Distinguishing between the internal structure (the organization of information within) and external structure (the accessibility of the information) of hypermedia/multimedia products, Cates asserts that teachers need to assess the levels of internal and external structure (high or low) and keep in mind that the level of structure may influence the student outcome. Low levels of structure may lead to “fruitless searches, learner disorientations, and . . . disjointed outcomes” (Marchionini, 1988; Jonassen & Grabinger, 1990; White, Cates, & Fontana, 1991). Teachers may supply external structure by helping students utilize the static/directive support features and the dynamic/interactive support features found within the programs (p. 3).

Drawing from the literature, Cates offers support strategies and possible evidence within each metacognitive category (a total of six, summarized here):

- 1) Task analysis
Key issue/ Concerns: Helping the learner understand internal program structure while focusing on the task at hand. Meta-cognition is not a comparison-based activity, speed does not indicate excellence or quality of metacognition. Students need to be aware of their own work habits and rate of progress to self-gauge against teacher advice.

Static/ directive support: Outline of steps; content maps; schedules; questions prompting learners to create schemata. Dynamic/ interactive support: Help learners monitor their progress; clarify scope; create models (of beginners, intermediate, & expert) against which to level advice for present learners.

Process evidence: Advance organizers created or used by students; discussion with/ among students. Product evidence: Journal; learner-created materials.

   2) Goal setting

Key issue/ Concerns: Before students can set goals they must understand what goals and subgoals are, and how they related to each other to accomplish the task; students may need to be taught how to represent these goals in hierarchies and networks.

Static/ directive support: The program could suggest goals and subgoals. Dynamic/ interactive support: The program could allow students to choose among goals and subgoals, tailoring advice to the selections.

Process evidence: Access to listed goals, or selections; discussion of these. Product evidence: Journal entries/ learner-created materials including hierarchical layouts and network illustration.

   3) Strategic action

Key issue/ concerns: Selecting appropriate learning strategies/ determining the effectiveness of the strategy/ revising the strategy as needed. Exercises may need to be followed by debriefing, students need experience in judging the goals and the teacher, rather than the program, can provide this. Learners may not believe they can change their goals at any time.

Static/ directive support: The program might provide the ability to view goals and subgoals while working; a “coach” may intervene on screen to externally evaluate goals and suggest revisions. Instruction on evaluating strategy could be provided. Dynamic/ interactive support: Learner progress could be monitored by the program with intermittent suggestions based on comparison to three learner models; questions could be posed during the process. Modelling of strategic evaluation could be offered.

Process evidence: Repeated access to goal/ subgoal statements (possibly tracked by the program); comments made by students regarding strategy, evaluation, and revision.
**Product evidence:** Journal entries would address evaluation/revision of goals; goals would have been revised.

- **4) Cognitive load**

**Key issue/Concerns:** The skill is in minimizing the load, particularly the memory load. Learners may need to become accustomed to program demands; they may need to learn “memory tricks” determining the optimum combination of their own memory aids and support mechanisms.

**Static/directive support:** Key information can be maintained on screen and through online help. **Dynamic/interactive support:** Advice can be offered on sequences or paths, second requests for advice would cause the tips to be rephrased; advice would be multi-modal – using text, graphics, and sound as possible.

**Process evidence:** Little evidence of recursive backtracking; few expressions of confusion or requests for help. **Product evidence:** Cognitive overload might be evident when learners express concerns about confidence or confusion over program features.

- **5) Persistence/responsibility**

**Key issue/Concerns:** Recognizing the scope of mental effort required and distributing effort across the task as appropriate. This includes taking responsibility for learning, acknowledging personal control. Low-ability or low-esteem learners may be more easily discouraged. Computers cannot provide warm, sensitive support for students, so the teachers role is crucial. The program also cannot use the language of “empowerment” without providing for it.

**Static/directive support:** Information could be provided regarding time demands, relative difficulty of material to be covered. Phases used within the program can emphasize the learner as an active participant (i.e., using “your choice” within directions, using an active rather than a passive voice).

**Dynamic/interactive support:** Encouraging comments could be provided as learners work through tasks; program could offer an experienced model describing how he or she worked through the tasks. Adaptive difficulty levels could assist learners encountering troubles, either learner controlled or triggered by symptoms suggesting flagging experience. The program could offer choices instead of imposing decisions; it would be non-intrusive.

**Process evidence:** Few breaks, focused discussion with little sidetracking; screen times comparable to learners who work persistently. First person pronouns and possessive would be used in referring to progress, evincing a sense of control and responsibility.

**Product evidence:** Entry dates and times in journals would reflect distribution of task across effort. Progress log would indicate steady progress. Entries would not express feelings of helplessness, nor would they attribute outcomes to the behaviors/control of others.
6) Growth (metacognitive)

**Key issue/ Concerns:** Analyzing the success of a learning outcome and generalizing from one learning episode to another. Some learners may define success as simple completion or escape from the task; others may believe they failed because they compare themselves to more expert models or the person doing the debriefing. Common definitions of success must be agreed upon in-advance. We cannot know for sure if learners are generalizing from one learning episode to another unless we have opportunities to observe learners across multiple learning episodes.

**Static/ directive support:** Program could maintain accessible records of learner’s previous performances. **Dynamic/ interactive support:** The program could comment on learner success; teacher and peer debriefing could be assisted by a debriefing print out.

**Process evidence:** Learner would refer to records of previous performances. Learners might participate in a discussion of the strengths and weakness of their task performance and their definitions of success. **Product evidence:** Entries would refer to evaluation of performance in completing the task.

Cates emphasizes that it is unclear “whether external support structure actually stimulates metacognition or merely substitutes for it” (p. 11). Researchers caution that learners may become too dependent upon external support structures (Derry, 1985; Day, et al., 1985; and Wellman, 1985) or that too much support might inhibit student development of their own strategies (Yore, 1986; Whitener, 1989). Kozma (1987) disagrees, asserting that external structure helps students develop their own skills. Even as control is -- increasingly -- transferred to the learner, more research is needed regarding metacognition and external support. (Contains 118 references.)

**Learning Strategies & Interactive Multimedia Instruction**


Reviewing research on note taking and regarding strategies for outlining and concept mapping, Kenny and Schroeder seek to integrate these strategies into multimedia instruction. Unless the strategies are used and understood, they will be ignored by the students.

Drawing from broad definitions of “multimedia” (computer-driven devices and/or computer programs, capable of producing sound, graphics, animation, and video) and “hypermedia” (an alternate term for non-linear, interactive, multimedia instruction) the
authors note that prior research has asserted that “structure in hypermedia may actually inhibit user interest and, hence, exploration” (p. 962). This view supports the constructivist view of learning that puts the learner at the center of meaning-making (Jonassen, 1991). The structure, then, is intrinsic to the devices.

As learners interpret and encode new information, they integrate it into their memory, and it is assimilated into existing schemata, or accommodated as they develop new schemata (p. 963). Encoding may take place according to four processes:

- Selection of information
- Acquisition (transferred between working and long term memory)
- Construction (building connections with the information in long term memory)
- Integration (prior knowledge transfers to working memory to construct external connections to the new information) (based on Cook and Mayer, 1983).

“Learning strategies (underlining, repetition, etc.) are used to rehearse, organize, and elaborate the information to make it more meaningful” (p. 963). Three kinds of learning strategies are: 1) rehearsal; 2) organization; and 3) elaboration.

Note taking has been found to be beneficial, independent of review, because it increases attention during lectures while facilitating the encoding of ideas into long-term memory (Hartley & Davies, 1978; Kierwa, 1985 and 1987). Peper and Mayer (1986) advance three sub-hypotheses: 1) attention (note taking asks the writer to pay attention); 2) distraction (the writer focuses on writing instead of listening; 3) generative (meaning is generated, with more integrated learning outcomes.)

In addition to influencing encoding, note taking is the product function or external storage of the notes. Outlining is a skill resulting in: 1) focus on important points; 2) familiarity with text structure; 3) retention; 4) generation of useful alternative texts; 5) active participation in learning (Bianco and McCormick, 1989). Findings from high school and college age students are mixed regarding the effectiveness of this strategy and its impact upon recall and paper writing. The authors cite Taylor and Beach’s (1984) finding that “the use of a hierarchical summarization strategy… improved comprehension and recall in middle school students” (p. 966). Slater, Graves, and Piche (1985) found that the 9th graders they studied did better on multiple choice tests after outlining.

“Outlining, like strategies such as imaging and paraphrasing, requires a major intrusion in the reading processes and also necessitates a significant amount of training” (Anderson, 1980, referenced in Kenny and Schroeder, 1994, p. 966). The authors note that it may be taught in middle school, if at all, but not always in ways students understand conceptually.

Based on Ausubel’s “Theory of Meaningful Learning,” holding that knowledge in memory is hierarchical, concept mapping is another researched spatial learning strategy (Novak, Gowin, & Johansen, 1983; Novak & Gowin, 1984; Heinze-Fry & Novak, 1990;
Novak, 1990). While maps can be used as tools for the negotiation of meaning (constructed by the learner and instructor together), they can also be used as an advance/ graphic organizer or as note taking techniques for material from texts or lectures (p. 967). This strategy has been shown to have medium positive effects on achievement and large positive effects on attitude (Horton, McConney, Gallo, Senn & Hamelin, 1993).

The authors note that many multimedia programs include built-in note pads or notebooks, but little to no guidance is provided regarding note taking, outlining, and concept- mapping strategies. As opposed to paper and pen notes, computerized notes offer the possibility of: 1) integrating and restructuring older information; 2) providing verbal and visual information as well as external links; 3) including prompts, cues, and models; and 4) offering the option of expanding or compressing outlines. The authors elaborate upon these advantages noting the positive and negative challenges inherent in each.

They conclude with design considerations, noting that -- in addition to knowing both the learners and the computers -- the research on note taking strategies needs to be considered. Just copying and pasting is not likely to be any more effective than highlighting (p. 975). Issues of pacing and organization also change in the move from paper to screen. (Contains 62 references.)

**Reflective Problem Solving**
**Rather than Acquisition and Application**


The authors assert in this discussion that students should be resolving problems rather than mastering skills and applying them. They emphasize the student inquiry process rather than teaching activities (such as scaffolding, modeling). “The basis for engaging a task is not the task itself but the prior knowledge of the student and the conditions under which the task is completed” (Hatano, 1988). They propose that reflective inquiry and the approach of problematizing a subject “depends more on the student and the culture of the classroom than on the task” (p. 16).

Without prescribing a specific curriculum or instructional methods, the authors illustrate their belief using and example focused on mathematics, while asserting that the approach can be transferred across subject areas. Problem-based learning, they assert is a response to the gap between acquiring and applying knowledge, but “although these approaches have been widely endorsed, we believe they do not resolve the difficulties that are inherent in the distinction between acquiring knowledge and applying it” (p. 14).

Drawing from Dewey’s notion of reflective inquiry (1933) they describe three stages: 1) problems are identified; 2) problems are studied through active engagement; 3) conclusions are reached as problems are (at least partially) resolved. The authors then extend the range of tasks that can be considered problematic (in the sense of being a
problem that can be solved by students), mindful of both “functional” and “structural” views. They then link reflective inquiry with understanding.

According to the “functional” view, understanding means “participating in a community of people who practice mathematics” (Lave, Smith, & Butler, 1988; Schoenfeld, 1988; Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991). Within this view, the focus is on the classroom activity including: the role of the teacher (providing information and setting tasks); the role of the students (including responsibility for sharing results and for explaining / justifying their methods; listening / engaging to learn from others).

Within the “structural” view, understanding means “representing and organizing knowledge internally in ways that highlight relationships between pieces of information” (Hiebert & Carpenter, 1992). Rather than focusing on the activity of the classroom (the “functional” view), the focus is on what students take from the classroom. After activities are completed, three types of “residue” may remain as knowledge retained: 1) insights into structure; 2) strategies for problem solving; and 3) dispositions toward mathematics. The value of a problem “depends on two things: whether students problematize the situation and whether it offers a chance of leaving behind important residue” (p. 18).

As students think about problems that are mathematically fertile, four other dichotomies, in addition to acquisition/application, are disintegrated by the authors: 1) that of teacher telling as opposed to student discovery (as students treat tasks as genuine problems, and teachers are not restricted exclusively to one or the other response); 2) that of “real-life” problems as opposed to “school” problems (as all problems are dependent upon whether the student engages, and the “residue” that remains); 3) that of cognition as opposed to affect (as both should be addressed within instruction); and 4) that of control of the curriculum: “bottom up” from the child, versus “top down” from the discipline (as residue is possible from problems originating from both places) (pp. 19-20).

The authors caution against modeling student activities after professional roles, explaining that: “children are different than mathematicians in their experiences, immediate ambitions, cognitive processing power, representational tools, and so on. If these differences are minimized or ignored, children can be thought of as small adults and education can become a matter of training children to think and behave like older adults” (p. 19). (Contains 67 references).

Learner Self-Regulation:
Self-Evaluation and Learning Goals


This article describes two studies conducted in sequence involving fourth-grade students learning about fractions. In both studies “skill” refers to “cognitive skill learning.” A pretest was administered measuring goal-orientation, self-efficacy, skill, and persistence.
In the two studies some of the students were motivated by the goal of learning how to solve problems (a learning goal) and others were motivated by the goal of simply solving the problem (a performance goal) (Dweck & Leggett, 1988). The two types of goals (learning and performance) do not have to be mutually exclusive (Nicholls, 1983) but were distinguished for the purposes of this study.

Focused on 20 African-American and 24 White students, study 1 found that the learning goal (with or without self-evaluation) and the performance goal, with self-evaluation, led to higher self-efficacy, skill, motivation, and task orientation than did the performance goal without self-evaluation. Study 2 consisted of 19 African-American and 21 White students. In study 2 all of the students self-evaluated their progress in skill acquisition, and the learning goal led to higher motivation and achievement outcomes than did the performance goal.

The author cautions against the belief that self-evaluation may always have desirable effects: “asking students to periodically assess their capabilities on a task they have repeatedly failed to master might lower, rather than raise, self efficacy and motivation, because after many negative attempts students might conclude they are incapable of learning” (p. 378).

Across the two studies presented, Schunck notes that learning goals and self-evaluation raise task orientation and lower ego orientation. Students less often compare their progress against their peers, while more often comparing their present performance against past performance. Self-regulation in both studies was influenced by perceptions of self-efficacy, or “personal beliefs about one’s capabilities to learn or perform skills at designated levels” (Bandura, 1986; Zimmerman, 1989; Schunk, 1991, et al.).

The effects of goals on performance depend on three properties: 1) specificity; 2) proximity; and 3) difficulty (Locke, Shaw, Saari, & Latham, 1981, Bandura, 1988). Goals that “incorporate specific performance standards, are close at hand, and are moderately difficult are more likely to enhance performance than goals that are general, extend into the distant future, or are perceived as overly easy or difficult” (p. 360).

Research on goal setting had focused on rate or quantity of performance, asserts the author, but the emphasis is shifting toward measurement of learning processes and strategies. Prior research links self-regulatory activities to more active cognitive engagement (Meece, Blumenfeld, and Hoyle, 1988). Citing Elliott and Dweck (1988), the author explains that, when given instructions focused on a learning goal (of developing competence) or a performance goal (of appearing competent),

“children given the performance goal and high-ability feedback persisted at the task but also avoided challenging tasks that might have entailed public errors. Performance-goal children who received low-ability feedback selected easier tasks, did not persist to overcome mistakes, and displayed negative affect” (p. 362).
Due to differences in type of subjects, experimental content, and teaching method, inconsistencies among studies in this area are difficult to resolve. The author concludes noting that “to be effective, self-evaluation must be linked with instruction so students learn and perceive they are making progress” (p. 378). (Contains 40 references.)


This study consisted of 15 elementary and middle school mathematics teachers, using technology to collect and deliver student achievement information. Tools used included: computers and software, video cameras, the Newton, scanners, and the Learner Profile. Assessment methods included interviews, portfolios, and scoring rubrics.

Noting that new models for assessment allow glimpses into the ways learners construct new knowledge (Cobb, 1990; Greeno, 1989) the authors asserts that new techniques require new tools. Even teachers who use technology in their classrooms overlook the uses of technology for assessment (Ginsberg, Sebastian, Underwood, et al., 1994). Prior research has focused on test-items (scoring, interpreting, reporting) (Bunderson, Inouye, & Olsen, 1989). Currently, video is one of several approaches to capturing student performance, providing evidence which would not be evident in formal testing (Barrett, 1990; Kaput, 1992). The three goals of this project were:

- To determine expert and novice teachers’ capabilities to develop comprehensive integrated assessment materials in mathematics;
- To investigate the role that technology plays in helping teachers present integrated assessment situations, as well as in recording and in analyzing student achievement;
- To investigate changes in teachers’ pedagogical reasoning, pedagogical content, and schemata as they develop and use integrated instructional assessment tasks (p. 49)

Data was collected through observations, interviews, teacher field notes and the assessment materials developed. Phase one consisted of the 15 teachers proposing curriculum modules and receiving video cameras. In phase two a training program was offered for teachers, regarding technologies, assessment, and techniques for being a teacher-researcher, including meetings with content area specialists and cognitive scientists.
The study found use of technology within the curriculum (relevant use) increased along with the use of technology for assessment. All but one of the fifteen incorporated the camera into their daily routine from the beginning of the program for: capturing classroom events, delivering information to students, parents, and colleagues, and for documenting changes.

Describing several features of technology that enhance teacher assessments, the author noted teacher-identified limitations in technologies for student observations (which proved to be either too cumbersome, or too expensive). This aspect of the project provided “informed feedback about the usefulness of new tools for the educational

Teachers found incorporating technology dramatically changed their assessment practices. As delivery of content changed (including use of simulations, audio-taping tasks; Hyperstudio and Powerpoint) and data collection also changed. Teachers used the video to gauge their students thinking processes and accumulate evidence of achievement indicators. This report was based on an NSF grant. (Contains 14 references)

**Technology-Based Assessment Tools for Managing Performance-Based Learning**


Reporting on two research studies focusing on the role of the teacher within computerized learning contexts, McNabb and Smith underscore the need for technological solutions to provide efficient, reliable, and valid information to teachers and learners, enhancing the instructional processes that accompany engaged learning experiences.

Study 1, developed within a grounded theory approach, investigated teacher monitoring strategies during computer-assisted composition instruction. Nine instructors and 180 students participated. Findings included four principles underlying process:

- Guided practice;
- Collaborative assessment;
- Instructional branching (preceded by diagnostic monitoring);
- And development of learner self-monitoring skills.

As the teacher played coach, facilitator, or co-evaluator, prescriptive monitoring strategies (such as goal-setting, explaining) and descriptive monitoring strategies (such as prompting, or mirroring) were evident.
Prescriptive monitoring strategies seemed to lead to skill development associated with a lower level of student autonomy whereas descriptive monitoring strategies seemed to lead to those associated with a higher level of student autonomy. Exceptions were evident in instances where the instructor misdiagnosed the student monitoring needs in relation to the task and their skill level (p. 264). Findings suggest that the computer facilitated a shift from teacher focus on process-writing pedagogy (focusing on the analysis of student products) toward students’ writing process.

Noting that constructivist theories of learning call for the use of alternative assessment measures (Wiggins, 1993; Boyer, 1995; Cole, Struyk, Kinder, et al., 1997) the authors highlight the trend towards “learning to learn” that has become part of the national reform agenda (p. 270). Study 2 concerned the design of electronic performance assessment systems, using inductive analysis, and offers guidelines for instructional designers and educators. Research included interviews with designers and consultants representing 5 performance assessment software programs.

Citing Marzano, Pickering, & McTighe (1993, p. 13) McNabb and Smith define three forms of assessment:

- Authentic: students apply knowledge and skills in the same way they would in the “real world” (Wiggins, 1989);
- Alternative: a broad term encompassing approaches that differ from traditional standardized testing;
- Performance: encompasses authentic and alternative (Marzano, Pickering, & McTighe, 1993) “it requires students to actively accomplish complex and significant tasks based on experiential learning and relevant skills to solve realistic . . . problems” (p. 271).

Within performance assessments, data collection includes: anecdotal records, journals, rubrics, portfolios, projects, presentations, and/or interviews. Gathering data is complex, time consuming, costly, and difficult to manage. If performance assessment is to succeed, “support for the classroom teachers to effectively collect and manage performance-based assessment is a strategic management issue facing the educational system” (p. 271). Later, they warn, “many designers have only a surface level understanding of the intricacies involved in performance assessment as a component in the total schema of a classroom environment” (p. 271).

In addition to developing guidelines, and questioning how technology could bring about cost-efficient, time-effective, and standardized approaches to assessment, the authors ask: “How can technologies improve reliability and validity in the performance-based assessment process?”

According to the researchers, standardization must be a facet of technologies for performance assessment. To ensure a valid assessment process, issues of accuracy,
multiple samples, data gathering flexibility, and consistency must be addressed. To ensure reliability, data entry, replication, historical review, and training must be addressed. Reliability also requires the establishment of the criteria and standards to be used through the recording of assessment data using: rubrics, anecdotal records, journal notations, video clips, hand drawn pictures, and/or written reports.

Student progress can be captured from multiple sources, and aid teachers working with students who may be struggling, exhibiting changes, or may have recently moved. Preparing teachers to become “more accurate and consistent in gathering and documenting assessment data is critical . . . a carefully and systematically designed training program . . . helps to ensure that the same rating can be replicated for a given student by another teacher at another time, thus supporting the reliability of the assessment program” (p. 273). The training plan must have an appropriate learning curve. As teacher confidence grows so does efficiency, allowing more time for teacher reflection on delivery of information or content; this allows for timely modification of educational goals.

Cost structures need to address single users, and small or large districts, with operational features that are practical and futuristic (drawing upon the possibilities inherent within the connectivity of E-mail, the Web, etc., especially between home and school).

The negative impact of connectivity in terms of communicating with parents could come from: inhibitions about writing; the loss of face-to-face / non-verbal communication; and the loss of spontaneous reactions / attention (with asynchronous communication). (Contains 49 references; includes table of guidelines for the management of performance-based assessment and an appendix of questions used to develop assertions regarding the monitoring of students.)

**An Evaluation of Performance Task Instruction in the Classroom**


Focusing on one Mississippi school district that has implemented comprehensive performance task instruction from grades 2 to 9. The Iowa Test of Basic Skills (ITBS, a norm-referenced test) and Tests of Achievement and Proficiency (TAP, also a norm-referenced test) were administered, in addition to the PREPS Integrated Assessment Model developed specifically for schools in Mississippi. Both qualitative and quantitative findings are offered.

The quantitative findings are limited by a small sample size and unclear study specifications. Researchers pose the hypothesis that performance task instruction would increase student achievement scores. The hypothesis is not supported when math performance assessment scores and language performance assessment scores are
analyzed separately (a total sample of 12 scores was used). The hypothesis is supported when the combined scores are analyzed, comparing 1999 to 1998. The researchers assert that no relationship is evident between performances assessment scores and core battery scores (a total sample of 18 scores was used.) Both calculations used one-way ANOVA, which indicates whether or not there is a difference but is limited to simply detecting that there is a variation. The authors note that “future studies should include individual student scores for data analysis” and the source of the scores being analyzed in this study is unclear (p. 14).

In the context of this project, performance assessment was defined as using “items [that] are multi-stepped tasks which require students to structure their own learning activity. To create this structure, students are required to apply information, knowledge, and previous learning. All performance tasks require student-constructed responses. . . students are required to describe the process by which they arrived at an answer” (p. 5).

Referencing the “New Standards Project,” the “Coalition of Essential Schools,” and the College Board’s “Pacesetter” program, performance assessment is at the focal point of these endeavors (Hakel, 1998). The authors note that “Goals 2000 gave, and is still giving, states the ability to develop a standards-based education by granting money to fund changes. Title I, another national program, is pushing for performance-based assessment to be an optional measure beyond the norm-referenced, multiple-choice testing” (Howell, et al., 1999, referencing Borko, Mayfield, Marion, Flexner, and Cumbo, 1997; Hakel, 1998).

Thirty-eight states as of 1994 were using writing samples (measuring writing proficiency) while twenty-five used performance-based items within their state-wide assessments (p. 4). Vermont, Maryland, and Kentucky, have fully implemented performance assessments in writing and math, at times to parental criticism regarding content (Khattri, Reeve, and Kane, 1998). When reform is adopted at the district-level, training is often needed as preservice programs do not train teachers in these teaching methods (Khattri, Reeve, and Kane, 1998).

After mentioning positive and negative beliefs regarding performance assessment, and reviewing the quantitative findings mentioned above, teachers indicate through the qualitative findings culled through a post training questionnaire that performance-based instruction and assessment had improved their own performance (they used checklists and rubrics more often, they used cooperative groups more often). The units within this PREPS Integrated Assessment Model were developed to emphasize the following components:

- Multi-stepped tasks
- Use of manipulatives
- Presence of literature/ text
- Students’ practice of returning to text
- Student constructed responses / student writing
- Checklists for student self evaluation
Rubrics for external evaluation
Checklists correlating to rubric (p. 12).

Student confidence sharing ideas seemed to have risen, students used strategies taught during their exams, and overall seemed to increase in their development of higher order thinking skills (stating opinions, self-assessing). In addition to offering qualitative insights, authors conclude that “a larger sample size would probably yield significant differences between uncombined scores” (p. 19). (Contains 17 references.)

Lessons Learned: Large-Scale Assessment and Alternative Measures


Outlining the history of alternative assessment in the U.S., the author identifies qualities needed for valid approaches to assessment. Research results are reviewed regarding the technical quality and consequences of using this new form of assessment for large-scale accountability purposes. Implications for future practice are included.

The 1965 Elementary and Secondary School Act first brought the requirement that schools receiving federal funding administer standardized tests to determine eligibility and effects, as the government entered local schools as advocates for disadvantaged students. In 1994 Goals 2000 encouraged states to set rigorous standards for student performance and to assess student progress.

The author uses authentic assessment, alternative assessment, and performance assessment interchangeably, all requiring students to generate rather than choose a response. Direct writing assessment was a first example of the positive effect assessments modelling skills could have on instruction (p. 5). Herman further notes: “change in assessment practices was one of several important factors that were likely responsible for changes in teachers’ practices and students’ performances” (p. 6). Today, learning is defined as reflective, constructive, and self-regulated (Marzano, Brandt, & Hughes, 1988; Bransford & Vye, 1989; Davis, Maher, & Noddings, 1990; Wittrock, 1991; Glaser & Silver, 1994).

Researchers have found evidence that traditional standardized tests, however, have a negative impact upon the curriculum and upon classroom learning (Dorr-Bremme & Herman, 1986; Corbett & Wilson, 1991; Herman & Golan, 1991; Kellaghan & Madaus, 1991; Shepard, 1991; Smith & Rottenberg, 1991).

Drawing on criteria for validity outlined by Linn, Baker, and Dunbar (1991), [see Messick, 1994, in this bibliography] Herman also emphasizes reliability drawing an available data to point out that problems achieving reliability decrease over time (as
rubrics are fine tuned, and teachers/raters become more experienced) (p. 15). She also notes difficulties in generalizability.

Sugrue (1996) has found that when tasks are tightly defined and the questions parallel except for format, results are more consistent. This approach is opposed to more loosely correlated measures, such as the results of a direct writing prompt versus those measured through a portfolio of writing. The measures from one assessment are not always comparable to the measures of another, as in one study students were found less capable when assessed by a prompt than they were when assessed by portfolio (Herman, Gearhart, & Baker, 1993). Although raters are an important source of error, variability due to task sampling is far greater (p. 18).

Cautioning that results vary, Herman notes that when items are not tightly specified 8-20 tasks are needed to obtain reliable individual estimates, with most studies pointing to the task range of 15-17 (Dunbar, et al., 1991; Shavelson, et al., 1990, 1991, 1993; Linn, et al., 1995). Even at these levels only a minimum level of reliability is reached (about .8) (Rogosa, 1994). Careful specification procedures can reduce the number of items needed to achieve minimum reliability.

She cites Glaser and Baxter’s framework and methodology for examining the match between an assessment’s intentions and the nature of cognition that is actually addressed (1996). Guided by expert-novice literature they highlight 4 types of cognitive activity that differentiate levels of competence: 1) problem representation; 2) solution strategies; 3) self-monitoring; and 4) explanation (p. 20). This framework has been applied to large, statewide performance assessments and in two of the three groups measured their approach was found to be invalid (p. 21).

Further describing issues of fairness, objectivity, bias in performance, and bias in opportunity to learn, Herman calls for comparability and equity in portfolio assessment. Reviewing the potential impact, she notes that some teachers are concerned about the time taken up by performance assessment, while others are concerned about the shift away from basic skills. Burdens are felt in terms of profession development time and resources as well as overall costs. Examining NC and KY performance assessments, teachers took the assessments very seriously and were generally supportive of the reform goals embodied by the change (McDonnell & Choisser, ND). In some cases parents misunderstand the tests, and disagree with underlying values (p. 31). (Examples of performance assessments are included p. 3; Contains 113 references.)

Impact of the Maryland School Performance Assessment Program (MSPAP)

In the 1996-97 school year researchers studied the extent to which the instruction, assessment, and Maryland School Performance Assessment Program (MSPAP) test preparation activities of language arts teachers aligned with the Maryland Learning Outcomes (MLOs) and the reading and writing tasks on the MSPAP. MSPAP is a performance assessment program for grades 3, 5, and 8 implemented in the early 1990s to assess student achievement and school performance in regard to the MLOs.

Elementary and middle school language arts teachers (280 from 44 schools) contributed reading and writing activities in the fall and spring (including those reflecting instruction, assessment, and MSPAP test preparation).

In this study the majority of the language teachers' classroom activities generally reflected some of the MLOs. The extent to which the classroom instruction / assessment activities reflected characteristics of the MSPAP tasks, however, was more limited. Approximately one-third of the classroom instruction / assessment tasks were classified at one of the two highest score levels on the MSPAP scale. 49% of the reading and 66% of the writing MSPAP test preparation activities were classified at one of the two highest levels.

The MLOs include reading and writing activities that are defined with a purpose (“tasks” with purposes explicitly described, such as “reading comprehension”) and those without definition (“skill exercises” such as a vocabulary exercise) (p. 8). MLO activities are also paired with learning stances. For reading the four stances include: 1) global understanding; 2) developing interpretation; 3) personal reflection/ response; and 4) critical interpretation (p. 12). The need to provide students with opportunities to work collaboratively is emphasized in the MLOs, while the MSPAP features tasks that emphasize interdisciplinary language arts skills.

As noted earlier, a majority of the classroom activities reflected the MLO priorities, while evidence of the MSPAP priorities was more limited. Teachers seem to have made changes in their instruction based on the MLOs and the MSPAP, according to results from language arts questionnaires, though these reported changes were not as evident in the sampled classroom activities (p. 31). (This project was supported by a grant from the US DOE; Contains 18 tables.)

Revising Technical Measurement Standards to Address Performance Assessment


Discussing the implications of efforts to revise the Standards for Educational and Psychological Testing, Linn notes that the expanded role of the federal government and increased emphasis on standards and performance assessment are changing the nature and context of assessment. Linn explains, “negative consequences of high-stakes uses of
standardized tests and presumed positive consequences of performance-based assessments have clearly been an important part of the movement toward the latter type of measures” (p. 7).

Goals 2000 has mandated that assessment contribute to improved student achievement, and Linn emphasizes: “it is not sufficient to provide evidence that the assessments are measuring the intended constructs. Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and, at the same time, not producing unintended negative outcomes” (p. 8). As a result of the changes within the nature and context of assessment, technical measurement standards are also shifting.

Opposition to performance-based assessment, and outcome-based education in general often has little to do with “questions of validity, reliability, or other related technical issues” (p. 6). Instead, Linn points to controversies regarding subject matter and how it is taught. When opponents criticize technical short-comings of performance assessment, defenders respond that the assessments should be judged in terms of other criteria, instead of existing technical measurement standards.

Noting the difference between the primacy of validity and of practice, the performance-based assessment movement is, in part, “a response to a perceived overemphasis on reliability that has come at the expense of validity” (Gipps, 1994). Linn lists the four guiding principles of validity, which are:

- Validity is a unitary concept;
- The uses and interpretations of the results are validated, rather than the assessment itself;
- Validity is a matter of degree rather than an all-or-none judgment;
- Multiple types of evidence are needed in arriving at an integrated judgment regarding the validity of a particular use or interpretation (p. 6).

He then challenges the specification standards for carrying out the four guiding principles of validity. Rather than considering content, criterion, and construct (Guion, 1980), the principle that validity is “a unitary concept” has still led to interpretations of the measurement standards that were not unified. Content-related evidence would be considered, for example, while disregarding criterion-related evidence.

The technical measurement standards are “fragmented and enable test developers to pick and choose the standards they will consider and how rigorously they will meet them” (Shepard, 1993, p. 444). Indeed, the standards at the time this article was written “stopped short of asserting that construct validity is the unifying force” (Messick, 1989, p. 92). Even with validity as a unifying force, “there will be questions about its limits, particularly with regard to the consequences of uses and interpretations of assessment results” (p. 7). And there is a need to focus on “what is intended to be measured, rather than the use to which measurements are put” (Wiley, 1991).
The author emphasizes that technical measurement standards need to: 1) contain explicit guidance for performance standards; 2) acknowledge opportunities to learn; and 3) include issues concerning LEP (limited English proficiency) and IEP (individual educational programs) students. He adds that “technical standards clearly need to be higher for high-stakes student certification decisions than low-stakes uses of results” such as those that provide information about the status of the educational system. For low-stakes uses, lower levels of generalizability can be accepted (p. 11). More context-dependent interpretations focused on the characteristics of the particular tasks used demand less generalizability across tasks (Messick, 1994). “If taken to the extreme where the performance on a particular task becomes what Messick (1994) calls the target rather than the vehicle of assessment, without inferences about likely performance on other tasks or about domains, then cross-task generalizability is not relevant” (p. 11).

Linn points out that the language of reform invites generalizations to much broader domains and constructs. He asserts “in my view, they lead to the need to meet standards that require the reporting of estimates of variability, due to the sampling of tasks and the likelihood of misclassification” (p. 11). In this climate, the temptation in performance assessment is to focus on task-based performance as the target, ignoring generalizability, but often the reporting of the results switches to broad constructs (Messick, 1994). Task-centered and construct-centered interpretations need to be distinguished.

Issues of generalizability (across raters, across time) and reliability (generalizability across tasks) challenge proponents of performance assessment. This challenge had been standing in discussions of tests and measurements long before the issue of performance assessment was raised (Tyler, 1934). Recently, reliability has been challenged as a viable approach to scoring student performance (Wolf, Bixby, Glenn, and Gardner, 1991; Moss, 1994).

“Confirmation of the original judgment rather than independent replication” can drive the assessment task performance (Delandshere and Petrosky, 1994). This calls for the judgment of diverse experts or the ‘professional confirmation approach’ (p. 12). These labor intensive limitations may also produce results that do not transfer across schools. Linn introduces “moderation” as a way to address comparability without the use of an external exam, for:

- quality control (including inspection of samples of local performance by external moderators – called auditing or verification)
- quality assurance (including group moderation -- developing shared understanding of assessment criterion -- called consensus or social moderation)

Linn concludes that while this moderated approach will be needed within the current move towards portfolios or teacher-controlled, instructionally-embedded assessments, experiences of this approach within the US are too limited to allow for the codification of acceptable practices into standards.
He concludes cautioning that “it is critical . . . that the standards provide a better way of establishing priorities for obtaining evidence based on the stated claims that are made for an assessment, the specific uses and interpretations intended, and the plausible unintended consequences that need to be investigated” (p. 13). (Contains 59 references.)

**Consequences without Evidence?: Performance Assessment and Validity**


Examining authentic and direct assessments, Mesick, describes their contrasting functions and the associated implications for validity. He emphasizes that “validity and social values are intertwined and that evaluation of intended and unintended consequences of any testing is integral to the validation of test interpretation and use” (p. 13). He then questions the embrace of a “consequential basis expense of the “evidential basis,” particularly in regard to performance assessment.

The issue is not for or against authenticity or directness, he asserts, but instead the issue is the nature of the evidence countering two major threats to construct validity: 1) construct underrepresentation (which jeopardizes authenticity); and 2) construct-irrelevant variance (which jeopardizes directness) (p. 14). (Examples of constructs are: student “competence,” “skills,” and “knowledge.”) The author asserts that authentic assessment must be sought in the quest for “complete construct representation.”

Considering issues of validity, Messick cites specialized criteria proposed by Linn, Baker, and Dunbar (1991) tailored for performance assessment:

- Content quality
- Content coverage
- Cognitive complexity
- Meaningfulness
- Cost and efficiency
- Transfer and generalizability
- Fairness and
- Consequences.

Distinguishing between task-driven and construct-driven performance assessment, Mesick articulates several distinctions used when conceptualizing the contrasts and trade-offs associated with performance assessments, including those between:

- Performance and products;
- Assessment of performance per se and that of competence (or other constructs);
- Structured problems/ response modes and unstructured problems/ response modes; and
Breadth / depth of domain coverage.

Given these distinctions, Messick asserts that performance assessment traditionally refers to performance-and-product assessment (Fitzpatrick and Morrison, 1971). Rather than using any product, performance, or student-constructed answer, a principled approach or guiding rationale is needed, akin to test specifications, tying the assessment of products/performances to the purposes of the testing (p. 14).

He notes that in dance or art, performance and product are the same. Citing Fitzpatrick and Morrison (1971) he explains, “In general one should consider assessing performances if task procedures have been explicitly taught and deviations from practice can be detected, whereas assessment of products should be considered if proper task procedures are diverse or indeterminate or have not been explicitly taught” (p. 14).

If evaluation of the performance or product per se is the focus of evaluation – the target rather than the vehicle – then judgment is of quality without concern for generalizability or replicability. “Inferences are not to be made about the competencies . . . from observed behavior to constructs such as knowledge and skill” (p. 14). When evaluation of the performance/product is the vehicle of assessment, with concern for generalizability and replicability, problems arise when task-driven assessment is paired with construct-driven inference. Attention must be paid to convergent and discriminant evidence in order to sustain the inferences.

Recalling that “construct” can refer to student competence, knowledge, skill, etc., generalizability also becomes problematic within performance assessment when considering three aspects of the task, or the “domain coverage;” 1) the content (represented within the construct measure); 2) the process (representing the construct); and 3) the degree to which the processes are reflected in the construct measure (Embretson, 1983). Messick asserts that construct-related evidence must be collected from sampled processes that students engage in task performance.

“A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society” (p. 16). Messick then questions behaviors or performances revealing the constructs, and the tasks that elicit the behaviors. This approach may heighten awareness of construct irrelevant variance, for example – “as when scores on essay tests are influenced by . . . quality of handwriting.” When in doubt regarding the relevance, both should be assessed separately (Breland, 1991) so that “their differential effects can be taken into account in scoring” (p. 16). Impact upon gender and racial/ethnic groups must also be considered (p. 22).

He cautions that “the preemptive emphasis on tasks and performances in the task-centered approach to performance assessment may not only bring behaviorism back into education by the rear door but, in effect, also . . . the operational definition” (p. 17). At the same time construct-centered scoring criteria and rubrics may become too generic,
they must be able to address different levels of performance quality, while not becoming too specific to the task. Considering authentic assessment, decontextualized tasks (such as multiple-choice tests) can provide additional information about students connected to diagnosing and remediating performance deficiencies, while possibly tapping skills based in abstraction (p. 18).

Assessments may be: 1) contextualized; 2) decontextualized; or 3) cross-contextualized (finding correlations across task types or contexts). “Whether one focuses on context-dependent measurement (or instruction) or on increasingly context-generalizable measurement or on . . . context independent measurement may depend upon the student’s level of developing expertise in the subject matter” (p. 19).

Messick puts richly contextualized assessments in perspective, stating: “features that engage and motivate one student and facilitate his or her effective task performance may alienate and confuse another student and bias or distort task performance” (p. 19). Such unintended consequences would need to be evaluated when establishing the consequential-basis for assessment validity. One approach to address differential student responses to context is to construct tests for comprehension, or reasoning, within several subject areas.

While writing about direct assessment of task behaviors and the indirect assessment of skills (inferred from performances and products), Mesick cautions that the term “direct” is a misnomer but is often used in education. “Direct assessment” is often defined by open-ended tasks and judgmental scoring, with the intent of minimizing constraints upon students implicit within formal testing conditions (i.e. tendencies towards guessing). But open-ended tasks have their own conditional constraints.

Again, attention must be paid to convergent and discriminant evidence to establish validity. The claim that a performance assessment is authentic and direct is, in essence, a claim of construct validity and, accordingly, needs to be supported by empirical evidence. Messick is VP for research at ETS, and has been working with issues of test validity and social values over a span of 30 years. (Contains 34 references).

A Practice-Based Model to Develop Alternative Assessments


Researchers using mixed methods, including a grounded theory approach, describe alternative assessments being used at two elementary magnet schools in North Carolina. They refer to case studies of the assessment development process. Assessments, grade levels, school years, and school districts in the case studies differed, but commonalities in developing the assessments enabled the researchers to adapt a seven-step practice-based model from the SouthEastern Regional Vision for Education (SERVE), the federal
education laboratory in their region. Findings from this study indicate that the assessment model derived from the initial case studies is applicable to other schools.

The authors discuss similarities and differences between the schools including their approach to the assessment, with one school developing a 14-page assessment and the other opting for most costly electronic portfolios. Grants and professional development fueled the process at both sites. At one school the principal led the charge, with a teacher development group receiving informal information about assessment, while at the other school leadership came from the district office, with all teachers offered development opportunities.

Over the three years of this study, in one school the assessment system was a “continuous progress” method of evaluating students and communicating information to parents. In the other, the assessment drew upon the multiple intelligences and arts integration aspects of their curriculum and sought to influence the local preservice teacher curriculum.

The SERVE model’s seven areas of assessment were employed by the authors, including:

- Student impact;
- Support;
- Curriculum and instruction;
- Resources;
- Staff development;
- Leadership; and
- Faculty buy-in (p. 7).

The study sought to test this model’s validity for assessing assessment development at other schools. In addition to the two schools in this study, the authors tested the model against 10 schools attending a summer workshop, synthesizing all of the commonalities exhibited across the sample as organized by the model. Discrepancies are defined as “new understandings” that embody teacher experience and contribute to an expansion of the model to accommodate short-comings in the current model (i.e. students may not realize that assessment is part of an on-going plan, this then needs to be explicitly addressed) (p. 17). All seven areas of assessment from the SERVE model are thought to be influenced by planning for quality alternative assessments and, in turn, influence the assessments themselves. (Contains 7 references.)