



EDUCATION

What Have We Learned from Pioneers in Innovative Assessment?

Brian Stecher and Laura Hamilton

What Do We Mean by Innovative Assessment?

- **Different from the norm, i.e., not multiple-choice testing**
- **Innovation in terms of:**
 - **Prompts**
 - **Response options**
 - **Entails innovative scoring procedures**
 - **Delivery systems**
- **Other terms “performance assessment” and “alternative assessment”**

What Aspects of Assessment Are of Interest?

Structure	Format, stimulus materials, response demands, administration, etc.
Technical Quality	Reliability (ratings, scores), fairness, validity for use in accountability
Impact on Practice	Changes in instruction: reallocation of time, reallocation among topics, test preparation
Cost	Expenditures, burden on staff, students
Reactions of Stakeholders	Knowledge, opinions: support, concern, opposition

Part I: State Initiatives in the 1990s

- **24 states “interested in, developing, or using” performance assessments in 1991**
- **Focus on a handful of “vanguard” efforts**
 - **Vermont Portfolio Assessment**
 - **Kentucky Instructional Results Information System (KIRIS)**
 - **Maryland School Performance Assessment Program (MSPAP)**
 - **Washington Assessment of Student Learning (WASL)**
 - **California Learning Assessment System (CLAS)**
 - **NAEP Higher Order Thinking Skills Pilot**

Vermont Portfolios

- Implemented 1991-92, replaced by late 1990s
- Promote instructional change and provide comparable school scores for accountability purposes
- Ended largely due to low quality of scores

Structure	Math and writing portfolios in grades 4 and 8. Varying amounts of student work and “best pieces” On demand “uniform” tests.
Quality	Inter-rater correlations on pieces 0.35-0.50, on total scores 0.50 - 0.80. High student by task variance. Low convergent/divergent validity with Uniform test scores.
Impact	Big effect on instruction; more problem solving and communication. “Worthwhile burden.” Concern about “rubric-driven” instruction.
Cost	\$13 per portfolio to score; classroom time.
Stakeholders	Value performance assessment: New standards reference exam, then NECAP.

KIRIS

- Implemented 1992; replaced by CATS in 1998, then KCCT
- School-level accountability, cognitive and non-cognitive domains, seven KERA content areas
- Quality mattered; writing portfolio still used as part of KCCT

Structure	Grades 4, 8, and 12 initially. MC, short essay, performance “events” and portfolios. Biennial accountability cycles; proficiency in 20 years.
Quality	Single score per portfolio; inter-rater ~0.70. Own teacher’s scores > neutral teachers’ scores. “Gaming” amount of class time per subject. Lack of concurrent validity with NAEP and ACT scores.
Impact	Changes in classroom practice to promote problem solving, communication, writing. Open response more than MC. Lots of “preparation” and “KIRIS like” tasks.
Cost	\$26 million bonuses (\$2000 per teacher) in 1994-95.
Stakeholders	Parents/educators raise questions. Still dispelling “myths.”

MSPAP

- Implemented 1991; replaced in 2002
- Progress towards state's educational goals; accountability
- Final blow was NCLB need for individual student scores

Structure	Grade 3, 5 and 8. All performance based, ranging from short answer to multi-stage. Cross-subject. Matrix-sampled, supports school level reporting.
Quality	Internal consistency alphas ~0.85 in four subjects , 0.70 in writing. Expert panel endorsed. School scores fluctuated annually.
Impact	Push instruction toward more complex and authentic tasks. Instructional changes associated with increased scores. Did not influence classroom assessment very much.
Cost	Rewards and sanctions at school level
Stakeholders	Abell foundation, district superintendents

CLAS

- Implemented 1993, modified in 1994, discontinued 1995
- Align with state curriculum, use performance assessments
- Public controversy over content led to demise

Structure	Reading, writing and mathematics, grades 4, 8 and 10. Group activities, essays, short-stories (ELA); show work and explain reasoning (math). Portfolios of student work.
Quality	State committee report critical of sampling and objectivity of scoring.
Impact	
Cost	No statewide test until 1998, then strictly MC.
Stakeholders	Parents objected to test content. Controversy increased when items were not made public.

WASL

- Implemented 1996-4th grade, then other grades, modified in 2004, postponed in 2008, replaced by MSP in 2009
- Measures learning in four subjects, listening dropped in 2004
- New tests MC and short-answer only

Structure	MC, short answer, essay and problems in four subjects. Includes classroom-based assessment options in other subjects. Individual and school scores.
Quality	Reliability / generalizability coefficients ~0.60 (listening), 0.70-0.88 (other subjects). Exact agreement on open-ended items 76%-97%. Variation in school percent passing year to year, larger variation at strand level.
Impact	Teachers align classroom instruction to perceived emphasis of WASL; changes weakly related to scores.
Cost	
Stakeholders	Low 10th grade pass rate as exit exam generated controversy. Governor delayed use.

NAEP Higher Order Pilot

- Pilot test in 1985-86
- Higher order thinking in math and science; 30 tasks adapted from UK
- Not implemented in NAEP science or math assessments

Structure	Pencil and paper, demonstrations, hands-on performance, computer administered. 1,000 students, grades 3, 7 and 11
Quality	Responded positively, some did well, older did better than younger, better on sorting and classifying than determining relationships and conducting experiments.
Impact	
Cost	“Feasible and worthwhile” but “costly, time-consuming and demanding”
Stakeholders	

Summary

- **Bold experiments did not survive. Why?**
 - **Moved too quickly, too large a scale**
 - **Too innovative, ahead of the science**
 - **Relatively costly and burdensome**
 - **Advantages not obvious (standards-aligned?)**
 - **Inattention to stakeholders and politicians**
 - **Conflicts over role of assessment**
 - **NCLB rules added further constraints**

Part II: Innovative Assessment Continues to be in Widespread Use

- Applications outside the K-12 sector and outside the U.S. are common
- Three approaches are particularly prominent
 - On-demand performance assessment
 - Essays
 - Hands-on tasks
 - Portfolios
 - Technology-supported assessment
 - Computerized adaptive testing
 - Simulations
 - Automated essay scoring

On-Demand Performance Assessments

- **Essays are widely used in K-12 assessment programs, especially in writing**
- **Essays are common in licensure, certification, and admissions testing**
- **Other types of performance assessment are less common in K-12 but there are prominent examples outside K-12**

Examples

- **Multi-state Essay Exam (MEE), part of the Bar Exam**
- **Queensland Studies Authority student assessment system**
- **Collegiate Learning Assessment (CLA)**
- **US Medical Licensing Examination Step II Clinical Skills Examination (standardized patient)**

Queensland Studies Authority (QSA) Assessments

- **Broad set of assessments, each with a specific purpose, linked to common course syllabi**
 - **Includes national assessments at selected grade levels to compare across states and territories**
 - **Supplemented with performance-based tasks (Queensland Comparable Assessment Tasks, or QCAT) in other years**
 - **High school includes projects, essays, oral recitations, and performances**
- **Because comparability is not needed for all assessments, system relies partly on local scoring and task selection**
- **System design was motivated in part by desire to avoid problems encountered in U.S. state testing systems**

Collegiate Learning Assessment (CLA)

- **Designed to assess student learning in colleges and universities**
- **Administered online, with two task types—Performance Tasks and Analytic Writing Tasks**
 - **Stimulus materials are varied and may include text-based documents, maps, photographs, charts, letters, etc.**
- **Responses are centrally scored; system is experimenting with automated grading**
- **Matrix sampling is used to limit testing time and promote breadth of coverage**

On-Demand Performance Assessment: Key Lessons

- **Scores on performance-based tasks are often combined with scores on a multiple-choice portion**
- **Feasibility depends on level at which inferences are being made**
 - **If individual examinee scores are important, a long testing time is required**
 - **If individual examinee scores are not important, matrix sampling is often used**
- **Some systems allow local autonomy in task selection or scoring, but this diminishes comparability**

Portfolios

- **Rely on products created outside a formal testing context (contrast with on-demand assessments)**
- **May be stand-alone portfolio or combined with on-demand assessments**
- **Inclusion of portfolio-type products is common in national assessments outside U.S.**

Example

- **National Board for Professional Teaching Standards (NBPTS)**
 - **Portfolio component includes 4 exercises**
 - **Aligned with content standards that include “soft” skills**
 - **Require submission of videos, student work, and other artifacts plus reflections**

Portfolios: Key Lessons

- **Typically requires extensive time commitment by examinee**
- **Scoring is expensive, with often insufficient reliability**
- **Task sampling is another source of error**
- **Might be most useful when combined with other formats**

Technology-Supported Assessment

- **Advances in information technology have affected test development, administration, scoring, and reporting**
- **For the most part, technology has improved feasibility of assessment but has not fundamentally changed it**
- **Recent developments have potential for altering the nature of assessment**

Three Applications of Technology are Increasingly Prominent

- **Computerized Adaptive Testing (CAT)**
- **Simulations**
- **Automated Essay Scoring**

Computerized Adaptive Testing

- **CAT is used in admissions, licensure, certification, and other testing contexts**
- **Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) is a K-12 example**
 - **Customized to state standards**
 - **Can be given multiple times in a year**
- **For the most part, CAT facilitates continued reliance on multiple-choice tests of broad academic areas**

Computer-Based Simulations

- **Allow students to interact with people or objects in ways that mimic real-world activity**
- **Reduce logistical burdens associated with hands-on performance tasks**
- **Can track problem-solving strategies and errors**
- **Most common in science**
 - **Example: Minnesota Comprehensive Assessment Series II**

Automated Essay Scoring (AES)

- **Various approaches to AES have been developed and fielded**
- **Typically require a number of essays to be scored by expert humans; software identifies criteria and weights that predict human scores**
- **Empirically derived criteria don't always align with expert rater criteria**

Technology Offers Potential to Improve Large-Scale Assessment

- **Expansion of constructs that can be measured**
- **Increased availability of detailed data to facilitate improved decision making**
- **Opportunities to address needs of students with disabilities and ELL students more easily**
- **Better integration of classroom-based and large-scale assessment**

But Several Concerns Remain

- **Effects of examinee familiarity with technology (source of construct-irrelevant variance)**
- **Uneven availability of technology in schools**
- **Need for educator training in technology use and data use**
- **Validity concerns stemming from limitations of automated scoring**

Part III: Implications for State Assessment Systems and Policies

- **Efforts to broaden accountability systems may benefit from advances in testing technology**
 - **Include new subjects and skills (e.g., foreign language production, artistic products)**
 - **Create richer and more complete data files to facilitate reporting and analysis**
- **But limitations of innovative assessments still remain**
 - **Cost, technical, and political concerns must be addressed**
 - **Effects on curriculum and instruction require monitoring**

Implications, cont. (2)

- **Revision of standards to describe both content and ways that mastery might be demonstrated could guide development of performance assessment**
- **Inclusion of multiple formats in a single testing program may address some concerns**
- **Scoring systems should receive as much scrutiny as the test items**

Implications, cont. (3)

- **Professional development is needed to promote appropriate use of information from new assessments**
- **Advantages of innovative assessments must be made clearer to stakeholders**
- **Classroom assessment should be considered part of a comprehensive assessment *system***



EDUCATION