



Performance-Based Assessments: Examining the Issues and the Progress

Ronald H. Heck

During the past decade, states have attempted to upgrade their expectations for student learning and school accountability. Forty states now have curriculum standards in all four core subjects (math, language arts, science, social studies), and an additional eight have adopted standards in at least one subject.ⁱ States are making similar changes in adopting new assessments. In recent years, performance-based tests have evolved in large-scale testing as an alternative format for assessing student learning and school improvement. Performance assessments rely on samples of students' work or judgments about students' performances that are used to evaluate thinking skills. In contrast to the more narrow focus on the accumulation of facts assessed through multiple-choice tests (i.e., where the student chooses the correct answer), cognitive approaches to learning have encouraged the development of hands-on assessments that require students to demonstrate their acquisition of problem solving, critical thinking, and application skills that are integral to conceptual understandings of core subjects.ⁱⁱ

The use of performance-based assessments in large-scale testing has dramatically increased in the past two years. In 1998, 21 states used tests that included some type of performance tasks. By 1999, this number had increased to 34 states. Across the states currently utilizing some type of performance assessment, the tasks range from composing a sentence to completing scientific experiments and writing up the results.ⁱⁱⁱ At present, more time and energy have gone into the development and validation of math and science performance-based tests than the other core areas.^{iv}

Despite this interest in alternative formats for math and science assessment, a 1999 issue of *Education Week* devoted to examining assessment and accountability practices across the 50 states suggested that writing assessments

are the most common type of performance assessments used across the K-12 curriculum. Most often, the writing assessments are in addition to standardized tests used in the core areas.^v

In a few states, however, performance-based measures are now used exclusively in all core areas. This can be a problem, however, if a state sets its standards' bar too low. It becomes more difficult to determine how students stack up against any type of national norms. For a large number of states multiple-choice, standardized tests remain the primary format for monitoring student progress and school accountability (i.e., with 13 states still relying exclusively on multiple-choice tests).^{vi} Whether performance-based tests will be even more widely used in large-scale testing will depend heavily upon the quality of the tests, their demonstrated relationship to curricular and instructional goals, and costs associated with their development, implementation, and scoring.

While performance tests can help improve instruction, for high-stakes testing, there are still issues to be resolved in their construction and use. It is important to ensure that performance assessments used for monitoring student learning and school accountability demonstrate valid and reliable measurement of designated learning tasks that are related to the school's curricular and instructional practices and are fair and nondiscriminatory to the fullest extent practical and possible.^{vii} In other words, measures should be chosen that are sensitive to what schools are trying to teach and minimally affected by those factors that they cannot control.

Outlining A Number of Issues

Performance assessments emphasize the integration of learning through tasks that require students to use multiple types of skills (e.g., critical thinking, problem solving, applying

knowledge) and response settings (e.g., calculating, conducting experiments, writing) to demonstrate what they can do.^{viii} A number of studies have shown that performance assessments and standardized tests measure different aspects of learning. For example, multiple-choice tests and performance tests purported to measure similar domains correlate only in the neighborhood of about .5 or .6. Unfortunately, performance tasks measuring the same student abilities do not seem to correlate any higher with each other than they do with standardized tests that measure the same general core area.^{ix} These findings suggest that further research is necessary in clarifying exactly what performance tests and standardized tests do, in fact, measure.

Performance assessments are thought to have greater utility than multiple-choice tests for helping school personnel improve their curriculum and instructional practices. They provide teachers with a means of ongoing evaluation of student progress that is more closely linked to what is actually taught. Proponents argue that performance-based measures broaden teachers' curriculum responsibilities, as opposed to narrowing their focus to "teach to the test."^x Educational reformers suggest that alternative assessments can provide an impetus for changing educational practices, as teachers must change classroom practices to teach to new curriculum standards that address a wider range of student skills.^{xi} Most standards-based assessments have only recently been implemented, however, so it is still too early to determine their effects on improving classroom instruction.

Performance assessments may also turn out to be a more equitable testing format, if they reduce differences in scores associated with student composition (e.g., gender, ethnicity, socioeconomic status) that are normally observed on standardized tests. The difference in achievement for certain groups of students remains an enduring dilemma in American education. Ensuring equity in the assessment of student learning is based on the belief that all students should have access to knowledge that emphasizes a conceptual understanding in the core subjects and develops their abilities to use that knowledge to reason and solve problems--

regardless of their socioeconomic status, ethnicity, gender, geographic location, and need for special services.^{xii}

In the past, the inappropriate use of standardized tests to make educational decisions has posed a substantial threat to the equality of educational opportunity for significant numbers of students. For example, standardized tests have been challenged in the courts on the basis of their content, uses, and disproportionate impact on minority students.^{xiii} Biases in standardized tests have led to unreliable results for minority students due to inappropriate norms and subsequent inappropriate placements of minority students in special assistance and remedial programs.^{xiv} In contrast, courts have held that disproportionate impact on a particular ethnic group, in and of itself, does not violate the equal protection clause of the 14th Amendment, if the quality of the test and its relationship to a legitimate educational purpose can be demonstrated.

Current student assessment practices vary considerably across the 50 states.^{xv} Only a few states test students in grades K-2. Most states begin testing students in the middle elementary grades. To look at some of these practices a bit more closely, 28 states currently have third grade student assessment. Of these states, 16 rely exclusively on standardized tests to assess student learning, while 11 states provide some types of performance tasks, and one state uses both multiple-choice and performance-based formats. More specifically, 11 states use performance tasks for language arts, or have a writing performance assessment, seven states include some types of performance tasks in math, two states provide science performance tasks, and one state provides performance tasks in history/social studies. A greater number of states test students in the fourth and fifth grades. For example, 27 states use performance tasks at the fourth-grade level. Twenty-three of these states use writing assessments, 19 states include math performance tasks, 8 states have science performance tasks, and 5 states have history/social studies performance tasks. For secondary grades, this pattern is generally repeated, with the greatest number of states testing in grades 8 and 10 (with only minimal testing now conducted at grade 12).

What the Research Says

Because performance assessments are relatively new in large-scale testing, their utility for assessing student learning and for promoting school accountability has not been fully determined yet in all core areas.^{xvi} To examine one core area more closely, while multiple-choice tests have been used in the past to measure aspects of language arts and writing indirectly (e.g., knowledge of writing mechanics, sentence structure, syntax, and grammar), more recently, test developers have also included the “holistic” aspects (e.g., merit, organization of ideas) of the writing process. Often, however, the writing samples used in state testing are limited (e.g., one 30 minute writing prompt).

The interest in writing performance assessment has led to several studies exploring the relationship between direct and indirect measures.^{xvii} Some researchers view students’ construction of writing responses as complementary with the multiple-choice type of information,^{xviii} while the argument against a comparison of direct and indirect assessments of writing is primarily that they measure different constructs. Others have found some correspondence between formats, suggesting that strength of the relationship depends on the ability levels of the students.^{xix}

A number of empirical studies have examined possible differences in writing and other core curricular outcomes for multiple-choice and performance-based formats among students of various backgrounds.^{xx} Several explanations have been offered for students’ observed differences in performance. One explanation involves the construction of the tests themselves and particular types of formats and test items that may be more difficult for some groups of students. While performance tests are supposed to reduce achievement differences among groups of students by reinforcing appropriate classroom curriculum that requires students to demonstrate their knowledge, some difference in performance among groups of students seems to persist regardless of the method of assessment used.

Another explanation for performance differences among groups of students focuses on

the home and school environments and their effects on learning. For example, family resources and learning opportunities in the home are related to high achievement in science and math, and Caucasian students tend to have more of these advantages than African-American and Hispanic students.^{xxi} These home inequities appear to translate into inequities in learning opportunities. Minority students are more likely to attend schools with less-qualified teachers, poor school climate, and less press for achievement.^{xxii} They may also attend schools that are more racially segregated.

Researchers at Harvard^{xxiii} recently examined differences in language arts outcomes across student groups. They compared first and second grade students’ portfolio assessment scores with their scores on multiple-choice tests in language (language mechanics and expression) and reading comprehension. The findings for students’ background variables were mixed. For example, portfolio assessment was found to be more equitable for African Americans. The achievement gap for the portfolio assessment was about one half of the gap on the standardized test; however, the scores for African Americans were still significantly lower than their Caucasian counterparts. They also found a smaller gap in performance for students receiving federal free lunch assistance on the portfolio assessment compared with their performance on the multiple-choice test. The achievement gap on the portfolio assessment was also smaller for limited-English speaking students compared with their scores on the multiple-choice test. In contrast, however, portfolio assessment produced a wider gap in performance favoring females over males than the advantage for females over males on the multiple-choice test. From these results, the researchers concluded that, at least for some groups of students, portfolio assessment reduced performance differences generally observed on standardized tests.

Similar results were also found in a study comparing constructed-response writing and indirect writing formats among elementary school students in Hawai‘i.^{xxiv} The good news is that the performance-based writing assessment reduced gaps in achievement commonly observed on standardized tests. For example, the

achievement difference was smaller for low socioeconomic students and some ethnic groups on the direct writing assessment relative to the corresponding standardized test.

While the evidence for overall greater equity favors performance-based measures, it is clear that these newer assessment formats will not entirely remove the types of gaps in student achievement observed on standardized tests.^{xxv} Achievement differences among groups of students are likely due to a variety of factors such as home environment, the quality of schools different groups of students attend, and the construction of the tests themselves. Performance assessments appear to be a step in the right direction, however.

Performance-Based Assessment and School Progress

Pressure to change educational practices can come from a mix of policy supports and sanctions that result from student test scores. There is, however, little research yet on the effects of school reforms linked to performance-based tests. A number of reforms over the past decade have focused on improving the school's instructional practices and developing appropriate measures for examining student learning. The effects of school improvement efforts have been difficult to demonstrate, however, partly because of the greater impact of student composition and school context on student performance on standardized tests and the mismatch between what is taught and what is measured.^{xxvi}

A number of studies have found that a considerably larger proportion of the variance in achievement scores between schools is explained by student composition variables on standardized tests, as opposed to variance explained by these variables on performance tests.^{xxvii} In the Hawai'i study, for example, student composition variables explained only about half as much between-school variance on the direct writing performance assessment as compared with the indirect, multiple-choice format. This means that observed differences in schools' scores on standardized tests are largely due to the students who attend the schools rather than the schools' instructional programs. These differences across assessment formats are

important to note because they call into question the use of standardized tests for school accountability purposes, since much of the variation in school scores is due to factors that are outside of the school's control.

Preliminary research on performance-based measures suggests that they may be sensitive to the school's intended improvement processes. For example, in one study of elementary schools, schools' efforts focusing on improving student writing were positively related to students' performance-based writing achievement levels.^{xxviii} From a practical perspective, this finding is encouraging because it provides initial evidence demonstrating that schools can undergo a specific, purposeful improvement process that helps them become more effective. Further study may reveal what schools actually did on a day-to-day basis to strengthen their curriculum. It may be that a focus on writing in a school introduces a completely new element in the curriculum. This may result in different resource allocations, staff development activities, and increased opportunities for students to write. Another possibility is that a district or state policy initiative, like the introduction of a new test, may result in curricular changes.

What Is the Status in Hawai'i?

Hawai'i is in the middle of efforts to develop performance-based assessments that measure student progress and hold schools accountable for meeting curriculum standards. Currently, the Department of Education is developing performance-based measures to be used in connection with standardized tests as part of the state's comprehensive assessment and accountability program. This provides both a way to look at the quality of classroom instruction (e.g., how well students achieve learning benchmarks) and to compare students against national norms. Similarly, efforts are underway to pilot test school-level performance indicators that describe the quality of school processes and help interpret levels of school outcomes.

In addition to monitoring quality, in developing these assessments, it will be also important to examine equity on several levels. Equity in terms of individual comparison should

focus on the performance outcomes achieved for various groups of students and the extent to which students' backgrounds influence their performance on various types of assessment. Similarly, ensuring equity in terms of school comparison should focus on how student composition (e.g., student socioeconomic status, students receiving special services) and context may influence the school outcomes produced. As research indicates, student composition appears to influence school performance differently on the two contrasting assessment formats.

Recommendations

Encourage the use of performance assessments for monitoring student learning.

Performance tests are an important means for monitoring student learning and comparing schools. Some evidence suggests that students' scores on performance tests may be less affected by their backgrounds. Performance-based assessments may also provide a closer linkage between the learning skills emphasized in the classroom and the actual assessment of those skills. As assessment experts are looking for measures that are less sensitive to what schools cannot control, this is an important point to keep in mind when considering the use of various assessment formats in making comparisons among schools.

On the other hand, performance tests have not yet undergone rigid empirical scrutiny. While it is important to examine how different test formats are related to students' socioeconomic status and other background variables in relative terms, one would not accept or reject a test that accurately measured a designated instructional concept solely on the basis of how it correlated with student composition variables--unless another measure of the same construct were available that displayed greater fairness across groups.

The effectiveness of performance assessments in monitoring student learning and school accountability depends on the extent to which states have implemented strong standards. Although initial comparisons with standardized tests are promising, it is likely that some of the reduced gaps in achievement observed on

performance assessments result from the specific test formats used in the research studies^{xxix} and from other student abilities that may interact with the formats. Some of these variables may include reading ability, home environment, and the quality of the schools students attend (e.g., press for achievement, culture, resources, teacher quality). One way to minimize the influence of these background influences is to track the progress of students longitudinally.

What Do We Do About Students Receiving Special Services?

There is a need to examine more closely the difficulties these special groups of students face in demonstrating their learning on writing performance assessments.

Students receiving special services may have difficulty with performance tasks such as writing. In some studies, students receiving special education or limited-English language services had large achievement gaps, regardless of assessment format. For example, it is likely that special education students' low scores on writing performance assessments is not a measurement problem, but rather, is associated with the difficulties of teaching a relatively complex performance task such as writing to these students. For special education students, these gaps are important to note because states are currently struggling to create alternative assessment systems. Under the 1997 amendment to the Individuals with Disabilities Act, states are required to include as many special education students as possible in regular assessments (with appropriate accommodations) and to create alternative assessments for students who are judged to be unable to take the regular tests.^{xxx} Further research is clearly needed on appropriate assessment formats for students who receive special services.

What About Cost?

Performance-based tests cost more to develop, administer and score.

In part, these higher costs associated with performance assessments are related to the number of tasks required to generalize about the student's performance in the domain.^{xxxi} There is a need to balance these increased costs against

the political and social costs of continued over-reliance on standardized tests. Implementing performance-based assessment in general requires a substantial commitment from policymakers in terms resource support.

Summary

Increasingly, states are adopting performance-based measures as part of their large-scale student testing and school accountability programs. Initial comparisons of multiple choice and performance assessments support the further refinement and use of these latter assessments for school accountability purposes. Future research should be directed at determining how student ability and background may interact with the content and format of the assessments developed, the correspondence and divergence in cognitive domains that each assessment format assesses, and the optimum number of content tasks that might be needed to generalize about the overall performance in different domains.

While there is great promise in these new measures for monitoring student learning and school accountability, there are increased costs associated with their use. Research suggests, however, that performance assessments are a useful format in large-scale testing because they are capable of assessing a wider range of student knowledge and skills. Moreover, when performance assessments are developed to be closely related to the school's curriculum and instruction, they should be more useful than standardized tests in monitoring the school's deliberate focus and energy directed toward curricular and instructional improvement. This would be welcome news for parents, educators, policymakers.

REFERENCES

ⁱ Jerald, C., Curran, B., & Boser, U. (January 11, 1999). State of the states. *Education Week*, 18(17), 106.

ⁱⁱ Baxter, G., Shavelson, R., Goldman, S. & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29, 1-17; Klein, S., Jovanovic, J., Stecher, B., McCaffrey, D., Shavelson, R., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis*, 19(2), 83-97;

Resnick, L. & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction*. Boston: Kluwer, 37-75; Supovitz, J. A., & Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equitability of portfolio assessment relative to standardized tests. *Harvard Educational Review*, 67(3), 472-506.

ⁱⁱⁱ Jerald, Curran, & Boser, 1999.

^{iv} Hamilton, L. (1998). Gender differences on high school science achievement tests: Do format and content matter. *Educational Evaluation and Policy Analysis*, 20(3), 179-196; Klein, 1995; Klein et al., 1997; Baxter et al., 1992; Stecher, B. & Klein, S. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, 19(1), 1-14.

^v Jerald, C. & Boser, U. (January 11, 1999). Taking stock. *Education Week*, 18(17), 81-99; Olsen, L. (January 11, 1999). Making every test count. *Education Week*, 18(17), 11, 15-16, 18-20.

^{vi} Jerald & Boser, 1999; Olsen, 1999.

^{vii} Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64(1), 5-29; LaMorte, M. (1996). *School law: Cases and concepts*. Boston: Allyn & Bacon; Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.

^{viii} Baxter et al., 1992; Resnick & Resnick, 1992; Stecher & Mitchell, 1995.

^{ix} Heck, R. & Crislip, M. (2001). Direct and indirect writing assessments: Examining their relationship to issues of equity and school improvement, *Educational Evaluation and Policy Analysis*, 23(1); Klein, S. (1995). Racial/ethnic differences and the relationships among measures: Do performance assessments narrow or widen the gap between population groups; do performance assessments measure something different? Paper presented at the RAND/NSF Conference on Performance Assessment in Science and Mathematics, Washington, DC; Klein et al., 1997.

^x Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York: Teachers College Press; Garcia, G. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. *Reading Research Quarterly*, 26, 371-392.

^{xi} Darling-Hammond, 1994; Klein et al., 1997; Linn, 1994; Resnick & Resnick, 1992; Smith, M. L. (1996). *Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program*. Tempe: Southwest Educational Policy Studies, Arizona State University.

^{xii} Garcia, G. & Pearson, P. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of research in education*, 20. Washington, DC: American Educational Research Association, 337-383; Klein et al., 1997; Porter, A. (1995). The uses and misuses of opportunity to learn standards. *Educational Researcher*, 24(1), 21-27; Wiley, D. & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California learning assessment system (CLAS). *Educational Evaluation and Policy Analysis*, 17(3), 355-370.

^{xiii} See, for example, Debra P v. Turlington, 564 F. Supp. 177 (M.D. Fla. 1983), affirmed, 730 F.2d 1405 (11th Cir. 1984); Hobson v. Hansen, 269 F. Supp. 401(D.D.C. 1967); Bond v. Keck, 616 F. Supp. 565 (E.D. Mo. 1985); and Larry P. v. Riles, 495 F. Supp. 565 (N.D.Cal.1979), affirmed 793 F.2d 969 (9th Cir. 1984).

^{xiv} Darling-Hammond, 1994; Garcia & Pearson, 1994; McCarthy, M., Cambron-McCabe, N., & Thomas, S. (1998). *Public school*

law: *Teachers' and students' rights* (4th edition). Boston: Allyn & Bacon.

^{xv} Jerald & Boser, 1999.

^{xvi} Darling-Hammond et al., 1995; Linn, 1994; Manzo, K. (2000) NAEP drops long-term writing data. *Education Week*, 29(27), 1, 17; Supovitz & Brennan, 1997.

^{xvii} See, for example, Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28(1), 77-92; Haladyna, T. M. (1998). Fidelity and proximity to criterion: When should we use multiple-choice? Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA; Hennings, S. S. (1996). A Comparison of equating methods applied to performance-Based Assessments. Paper presented at the Annual Meeting of National Council on Measurement in Education, New York, NY; Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction*. New York, London: Longman; White, E. M. (1994). *Teaching and Assessing Writing: Recent Advances in Understanding, Evaluating and Improving Student performance*. San Francisco, CA: Jossey-Bass.

^{xviii} Bennett et al., 1991; Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessment of writing skill. *Journal of Educational Measurement*, 16(2), 119-128; Culpepper, M., & Ramsdell, R. (1982). A Comparison of multiple-choice and an essay test of writing skills. *Research in the Teaching of English*, 16, 295-297.

^{xix} Hennings, 1996.

^{xx} Hamilton, 1998; Heck & Crislip, 2001; Klein, 1995; Klein et al., 1997; Linn et al., 1991; Supovitz & Brennan, 1997.

^{xxi} Peng, S., Wright, D., & Hill, S. (1995). *Understanding racial-ethnic differences in secondary school science and math achievement* (NCES 95-710). Washington, DC: U.S. Department of Education.

^{xxii} Finley, M. K. (1984). Teachers and tracking in a comprehensive high school. *Sociology of Education*, 57, 233-243; Lee, V. & Bryk, A. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, 62, 172-192.

^{xxiii} Supovitz & Brennan, 1997.

^{xxiv} Heck & Crislip, 2001.

^{xxv} For example, see, Cavazos, L., Finn, C. E. J., Goldberg, M., & Fields, R. (1988). *Creating responsible and responsive accountability systems*: Office of Educational Research and Improvement (OERI); Oakes, J. (1989). What educational indicators? The case for assessing the school context. *Educational Evaluation and Policy Analysis*, 11(2), 181-199.

^{xxvi} Wiley & Yoon, 1995.

^{xxvii} Heck & Crislip, 2001; Supovitz & Brennan, 1997.

^{xxviii} Heck & Crislip, 2001.

^{xxix} Klein et al., 1997; Resnick & Resnick, 1992; Supovitz & Brennan, 1997.

^{xxx} Sack, J. (2000). States report trouble with special ed. testing. *Education Week*, 29(27), 24.

^{xxxi} Stecher & Klein, 1997.

Ronald H. Heck is Professor of Educational Administration at the University of Hawai'i at Mānoa. He can be reached at 808-956-4117 or rheck@hawaii.edu.