Interview on Assessment Issues With Lorrie Shepard

The News and Comment section will feature a number of formats for examining research issues. This is the first use of the interview format for eliciting opinions about controversial areas. The *ER* interviewer is News and Comment editor, Michael Kirst of Stanford University. He first talks with Lorrie Shepard, who is professor at the School of Education at the University of Colorado, Boulder, and then with James Popham, who is professor in the UCLA Graduate School of Education and director of IOX Assessment Associates.

ER: **What are the reasons for the movement towards authentic testing, and what does this concept mean to you?**

Shepard: Use of the term *authentic assessment* is intended to convey that the assessment tasks themselves are real instances of extended criterion performances, rather than proxies or estimators of actual learning goals. Other synonyms are *direct* or *performance assessments*. The intense interest we are seeing in these alternative measures is a response to some of the deadly effects of multiple-choice tests, which are, in turn, the result of the inordinate weight given to traditional standardized tests in the past decade as a key feature of educational reform. Under pressure to raise test scores, the known limitations of multiple-choice tests have become greatly exaggerated. They become less valid indicators of what students know (because scores can go up without a commensurate gain in achievement); and more seriously, when multiple-choice tests become the focus of instructional effort, they have a negative effect on teaching and learning.

ER: **What are your concerns about using multiple-choice tests to drive classroom instruction?**

Shepard: When important standardized tests become the curriculum guides in a school or classroom, the quality of instruction is reduced in several respects. First, as many critics warned in advance, the curriculum is narrowed to only those topics that are tested. This often means that writing, social studies, and science are driven out of the instructional day, as well as "frills" such as art and music.

In addition to the predicted distortion of curricular frameworks, we now have evidence of unanticipated effects on the way that even basic skills subjects are taught. For example, in many cases teachers teach reading and math using worksheets and practice materials that closely resemble test materials. The behavioristic decomposibility and decontextualization assumptions—which the Resnicks identified as the faulty learning-theory assumptions underlying standardized tests—then shape the daily mode of instruction, leading to repeated drill on isolated skills. Even if well-crafted multiple-choice tests can assess higher order thinking skills, measurement specialists should recognize that the classroom tests created by teachers to mimic accountability tests are much more likely to elicit rote learning. Emphasis on raising test scores above all else reinforces other behaviorist principles widely held in schools, like the idea that thinking and reasoning should be postponed until after basic skills have been mastered. Instead of instruction being improved as intended, poor test performers get more drill, while only high scorers are provided with instruction aimed at teaching comprehension and problem solving.

Lastly, conceiving of instruction in the format of tests also affects children's attitudes and the inferences they draw about the purpose of learning. They learn, for example, that there is one right answer to every question, that the right answer resides in the head of the teacher or test maker, and that their job is to get that answer by guessing if necessary—hardly a perspective consistent with the goal of having children construct their own understandings.

ER: **How would authentic assessments help with the problem of postponing instruction that teaches thinking?**

Shepard: The tasks and problems used in authentic assessments are complex, integrated, and challenging instructional tasks. They require children to think to be able to arrive at answers or explanations. Thus performance assessments mirror good instruction, which engages children in thinking from the very beginning. For example, in first grade good teaching would not sort children into readers and nonreaders, letting readers do comprehension work because they had passed the decoding threshold while denying to nonreaders a chance to think about comprehension from text. Instruction aimed at thinking and the construction of meaning would instead focus on listening comprehension and ask all of the children to do

some of the things to understand a story line and remember some of the important elements of the story, whether or not they were decoding. These expectations would then situate decoding instruction properly in the context of why we do it, which is to be able to read and get meaning from texts.

Authentic assessment supports good teaching by not requiring teachers to redirect attention away from important concepts, in-depth projects, and the like. To the extent that performance assessments merely replace standardized tests as a different external demand, then at least when classroom instruction imitates these types of tasks and gives children practice with solving these kinds of problems, the focus is more likely to be on thinking rather than eliminating wrong answers.

ER: **What are some of the problems with implementing authentic assessment in the next two years or so?**

Shepard: My answer to that question depends on the purpose of the intended assessment. If the idea is to provide better classroom assessments in support of instruction and learning, then the problem is inadequate education of teachers, and the remedy is to extend to a wider group of professionals the insights that the best teachers have about how to construct their own assessment tasks and conduct systematic observations to inform instruction.

If the purpose is, however, to conduct a large-scale survey for accountability purposes, then the technical problems to produce reliable and representative scores are potentially much greater. We have many admirable examples of authentic assessments, but they are invariably judge or observer intensive compared to paper-and-pencil devices run through optical scanning machines. Therefore cost is a big factor, both for development and scoring. It is possible to have sufficient funds to conduct authentic assessments well without raising the total price-tag, at the state level for example, if legislators could be convinced to test less. Rather than testing every pupil in every grade in every subject, policymakers should be willing to invest in a few exemplary assessments in key subject areas by using a sampling of students and grade levels. The trade-off between quantity and quality of data should seem worthwhile once one recognizes both the corruptibility of standardized tests as indicators and their distorting effect on the teaching of challenging content.

ER: **Now in the sampling procedure, would authentic assessment be similar to current procedures where we gather teachers to judge writing samples? You get a consensus among two or three judges as to what the score is. Is that part of it?**

Shepard: Yes, current writing assessments and the College Board's Advanced Placement (AP) exams are examples of performance assessments. Although there are quarrels about the content of some AP exams (breadth over depth) and some writing assessments as currently administered, these examples demonstrate that we know how to solve problems of scoring standards and interjudge reliability. The general strategies for ensuring reliable and valid scores from subjective judgments can be applied whether judges are asked to evaluate written products, video-tapes of performances, oral interviews, or observations during science experiments.

ER: **Do you see this likely to happen in major ways at the local or state level in the next three to five years? Jim Popham, whom I interviewed earlier, was somewhat skeptical that this was going to happen very soon.**

Shepard: I don't think you'll see the 35 states now using norm-referenced tests all chucking them in the next three years and replacing them with authentic assessment. But I think you'll be surprised at the enthusiasm for these ideas. Some legislators are still absolutely convinced that holding schools accountable with mandatory basic-skills tests will make education better; I submit that they reside disproportionately in states that have just recently instituted such tests. In states that began high-stakes testing in 1984, however, proponents are now not so sure. As negative evidence accumulates—such as poor performance on higher order tasks on National Assessments of reading and mathematics—policymakers are becoming increasingly interested in alternatives to standardized tests.

ER: **States are sometimes going in one direction on testing towards authentic and performance testing, but the local districts still use standardized norm-referenced tests like Iowa, Stanford, Metropolitan, or the California to test basic skills, which don't really have the same concept as the state assessment. What will happen if we have these two different concepts implemented, one at the state level and another at the local level?**

Shepard: Well, making a prediction about that really depends on which of the tests has the greatest power. When OERI (Office of Educational Research and Improvement) commissioned CRESST (Center for Research on Evaluation, Standards, and Student Testing) to do a follow-up study of Cannell's report (that all 50 states are above average), we interviewed a nationally representative sample of 50 local testing directors, as well as all 50 state testing directors. Based on those data, we know that there is variation from state to state as to whether the state or district tests have the greatest political clout. We also learned that the stakes associated with a given test are based as much on the public visibility of test scores as on important decisions or sanctions that follow from test results. The test that leads to ranking of schools in the local paper is the one that is more likely to drive instruction. Therefore, it is possible that multiple-choice tests will continue to have a deadly effect if district-level standardized tests receive the greatest media attention.

There is even the danger that the advice I gave earlier about using sampling to make performance assessments feasible will unwittingly yield greater power to local standardized tests because they will be the only ones that continue to produce rankings of schools. While I think this problem has to be thought through, I remain convinced that impressively different authentic assessments can help to redirect effort toward important learning goals. In the case of science and social studies, for example, a state-level assessment would not be upstaged by local standardized test scores. To command attention for more ambitious assessments of reading and mathematics, it might be effective to extend the state-level sampling to provide district comparisons, as an external check on local claims made on the basis of standardized tests.

ER: **Let's shift to another subject now. You have been concerned about the effects of testing on various public policies. Let's start with your impression on where we are on readiness testing for kindergartners and first graders in terms of holding them back or starting them late. What has been the recent policy trend in terms of using tests? Now some policymakers seem to be removing them. Why is this?**

Shepard: I agree that at the state level readiness testing has been mandated and then withdrawn or greatly modified—the most infamous example being the Georgia kindergarten exit test. Policymakers were simply embarrassed by the public outcry. Whatever the public's understanding is about the fallibility and potential bias of tests, it's just much more believable that asking a 5-year-old or 6-year-old to take a test may lead to invalid results.

However, I do not think that there has been a diminution in the local use of readiness testing where it remains largely unscrutinized by the public. In a recent survey sponsored by the National Research Council, only three states did not report the use of readiness tests, at the state or district level, to delay school entry, to deny entrance to first grade, or to make special placements such as developmental kindergarten or pre-first grade. Recently we have begun to see a new use of screening and readiness tests which is to place "at-risk" children into kindergarten classes tracked by ability.

ER: **Your view is that the technology and validity and reliability of these preschool and first-grade tests are not adequate to do the job they're intended and that the locals want them to do?**

Shepard: The reliabilities of these instruments typically do not meet the standards of accuracy expected when making important life decisions for individual children, and their construct validity is questionable. Anne Stallman and David Pearson have done an illuminating analysis of academic readiness tests. They look pretty much like the first reading readiness tests given in the 1930s and are wholly incompatible with recent research on emergent literacy. And screening measures, often used as readiness measures, are basically short IQ tests.

The more serious problem, however, is that the treatments that follow as a consequence of the tests are themselves inadequate, even harmful.

ER: **You mean the educational program that follows low test results?**

Shepard: That's right. It is acceptable to give a treatment based on a fallible diagnosis if the treatment is unambiguously a benefit and has no side effects. But in this case, the treatments in the form of various two-year kindergarten programs are demonstrably ineffective based on controlled studies. And kindergarten retention and transitional grades often have negative social and emotional consequences for children. Therefore, the tests lack validity for these types of placements because the placements themselves are invalid.

ER: **Let me shift now to tests which are being used by localities or states for promotion purposes to hold kids back from grade to grade, and your view of both the validity and reliability of those tests, plus the impact of the educational prescriptions and treatments that come from nonpromotion.**

Shepard: Once again, I think the issue should be the efficacy of the treatments that follow from low test scores, not just the reliability coefficient associated with the test instrument. The research on retention is overwhelmingly negative. Out of 63 controlled studies identified by C. Thomas Holmes at Georgia, only 9 showed positive effects for retention. The average effect size was quite negative and did not improve when only the studies with the most extensive controls were aggregated. What's more, in the years following retention, retained children were further behind promoted controls on achievement measures than on self-esteem measures, which contradicts popular wisdom about the benefits of retention.

ER: **Do you have concerns about the large amount of testing used for placing special education pupils in special programs?**

Shepard: Yes. In the case of testing to identify children in mildly handicapped categories the costs of assessment and staffing procedures use up half of the extra per-pupil resources available without any evidence that pro forma administration of tests adds to the scientific integrity of placement decisions. In research that Mary Lee Smith and I have done, and in other studies, there is a very high correspondence between initial teacher referrals and final placement decisions, with all of the testing in between serving to justify placement. At least half of the children labeled by schools as learning disabled (LD), by far the largest category of handicap, are misidentified. Rather than fitting the original clinical definition of LD, they are more aptly described as slow learners, linguistically different children, misbehaving boys, children who are absent or whose families move too frequently, or as average learners in above-average contexts. And again, in the case of special education placement for these children, there is no evidence that pull-out programs they receive are certain to be a beneficial treatment.

ER: **Your view is that you're concerned equally about both the quality of testing and the quality of the educational intervention; it's the two together, not just one or the other.**

Shepard: That's right. If you had an unambiguously wonderful treatment, people would be clamoring to get into it. They'd be clamoring for retention; they'd be clamoring for special education placement. It would be reasonable to use a fallible measuring device, on the grounds that some information is better than none, and err in the direction of giving special treatment.

But time and again we have seen the parallels among special educational treatments that are not benign: tracking, special education placement for mildly affected learners, extra-year programs before first grade, and grade retention. So it's really the harm of the treatment that is more worrisome than the fallibility of the measure.

ER: **What are your views about the merits of measurement-driven instruction?**

Shepard: Measurement-driven instruction comes from the behavioristic test–teach–test learning model. It assumes that all of the constituent elements of important insights and understandings can be broken down and taught one by one. As I indicated earlier, this learning theory is seriously flawed and has a

(**Shepard** continues on p. 27)

in the field, So the chances for getting more competition here—do they look good to you?

Popham: We need to find some assessment agencies that are willing to invest the resources necessary to develop first-rate assessment alternatives for prospective teachers. Educational Testing Service (ETS) is currently creating a battery of tests for beginning teachers that they promise will be a substantial improvement over the tests currently available in the NTE program. The new ETS tests are scheduled to be available in a few years.

It is an eminently reasonable expectation to want a history teacher to know history, and a mathematics teacher to know mathematics. We also want our teachers to be fundamentally literate. Finally, if possible, we want beginning teachers to know something about instructional principles. Thus, I think it is highly appropriate to have incoming teachers demonstrate their skills via some kind of testing program. At the moment, however, we don't have an adequate number of high-quality tests from which to choose.

ER: **We've covered a fair number of test-related topics today. Given your responses, I suspect you'll agree that the assessment world is not exactly static.**

Popham: Years ago, as an undergraduate philosophy major, I learned that it was the view of Heraclitus that everything was always in a state of flux. It's certainly true that the educational assessment world is currently in an almost frenzied state of flux. And flux, as we know, is an F-word with four letters.

(**Shepard** continued from p. 23)

deadening effect on instruction, especially because it postpones attention to thinking and problem solving.

Very recently we are seeing a new version of measurement-driven instruction from advocates of authentic and performance assessments. Although I generally concur that more admirable assessments will have a more salutary effect on instruction and learning, I have two reservations about using assessments (however impressive) to leverage educational reform. (a) Under great pressure, the weaknesses of any assessment will be exaggerated. Therefore, you are always in danger of encouraging teaching to the assessed version of the learning goals rather than the original goals. (b) Forcing modes of instruction via external high-stakes assessments detracts from the professional role of teachers. It trades making the worst 10% of teachers better by fiat against empowering the other 90%. Both of these concerns can be alleviated, of course, if the assessments are sufficiently broad so that tasks are not prespecified and taught to, and there are multiple paths to successful performance. But in litigious environments these features are often negotiated out of testing programs because there is safety in specificity. These problems have yet to be worked out and should be resolved before powerful assessment programs are installed.