The Place of Testing Reform in Educational Reform
A Reply to Cizek

LORRIE A. SHEPARD

Educational Researcher, Vol. 22, No. 4, pp. 10–13.

Cizek offers several criticisms of my article "Psychometricians' Beliefs About Learning" (1991b). Each complaint reflects a serious misreading of the original article.

Cizek's Criticisms

1. Cizek ascribes to me a *model of educational reform*. The gist of this model is that education can best be improved by trying to change psychometricians' beliefs about learning. He fears that this imputed model might do great harm by misdirecting educational reform efforts.

I did not propose such a model and I reject Cizek's Figure 1 as an accurate representation of what I said. My research was focused primarily on describing measurement specialists' beliefs about learning and on tracing the origins of those beliefs in behaviorist psychology for the large group of specialists who held to the "criterion-referenced-testing learning theory." I was interested in the *influence* of beliefs on *testing practice*. Implicit assumptions about learning might explain, for example, why some measurement specialists see teaching to the test as a good thing and others see it as a threat to both validity and learning. I said that implicit beliefs should be examined and made explicit because "an understanding of learning theory is fundamental to evaluating evidence of testing effects and therefore to framing validity investigations" (p. 10). My purpose was not to test strong causal claims but to study more subtle effects of implicit assumptions on thinking, which then influence actions and color interpretations of testing consequences, and so forth.

It is not clear whether the interweaving of thoughts and actions can be neatly separated into cause and effect. Certainly, the importance of these relationships is not captured well by arrows of a certain size in a causal model. Even if the "influences" of psychometricians' beliefs on *testing practice* and indirectly on classroom practice are causal in nature, I did not specify a model. It was not my intention to evaluate the magnitude of these effects in comparison with other important influences. Specifically, interview data from measurement specialists would not have allowed an analysis of curriculum specialists' beliefs or contributions, and they were not mentioned in the article. Most importantly, my purpose in examining the connection between beliefs and testing practice was to gain understanding relevant to current controversies in the field of measurement. I did not suggest, regardless of the causal links, that intervening to change psychometricians' beliefs would be an effective means to reform education.

2. Cizek accuses me of failing to provide evidence demonstrating the connection between beliefs and testing practices. Apparently he ignored the appendix where I presented quotations from testing experts describing their efforts at test-curriculum alignment and perceptions about the instructional influence of tests. I found a compelling link between reported practices and two distinct belief systems derived from categorical sorting of the narrative data. For example, some test coordinators distributed item analyses or objectives lists and encouraged teaching to basic skills tests, while others tried to prevent narrow focus on the test. These two groups gave different rationales for their actions, allowing inferences about their beliefs. The data are there for readers to test the concurrence of beliefs and practices.

The criterion-referenced-testing (CRT) "learning theory" (characterized by propositions 1 and 2, Shepard 1991b, pp. 4–5) was the dominant and most coherently defined belief system (in these highly inferential data), accounting for approximately 50% of the responses. From this perspective, objectives-referenced tests are seen to be the complete instantiation of intended learning goals. Therefore, drill on materials that closely resemble the tests is accepted as a legitimate and effective means to improve achievement. Furthermore, because the CRT position is based on behaviorist theory, which conceives of learning as the sequential mastery of constituent skills, adherents to this view are willing to postpone the development of higher order thinking skills until after basic skills have been mastered. For example, the 22% of respondents represented by Category V all said that teachers had redirected instruction to tested objectives but that *no* important learning goals were slighted because the test covers "essential skills." Measurement specialists who held these beliefs engaged in practices such as the following. They adopted test-retest mastery learning programs and encouraged teachers to teach to essential skills tests (e.g., see quotation III.7 and V.19 in Appendix A). They acknowl-

LORRIE A. SHEPARD is professor, School of Education, Campus Box 249, University of Colorado, Boulder, CO 80309-0249. She specializes in educational measurement and policy research.

edged adding "pieces" to the curriculum to match the test better (IV.[1841] on p. 4; VI.15). Computer-generated lists were used to give teachers feedback on test objectives such as main idea, literal recall, facts or opinion, details, sequence, and so forth (V.14, V.15). Materials and staff development were provided to help teachers target instruction to each test objective (V.16, VI.23). (Note that in many cases a criterion-referenced testing model was used in speaking about a norm-referenced testing program.)

In contrast to the CRT view of testing and learning, the other major perspective was best characterized by proposition three: "Tests should be for monitoring but should not drive instruction" (Shepard, 1991b, p. 5). Whereas CRT-oriented measurement specialists take actions that directly encourage teaching to the test, specialists from the anti-measurement-driven-instruction perspective eschew such actions or engage in practices intended to prevent instruction from being focused narrowly on the test. For example, the 10% of respondents classified in Category II either used teacher observation to ensure coverage of the entire curriculum or purposely rotated tested objectives so that teachers could not anticipate which objectives would be tested in a given year. Thus, it is possible to point to specific practices, emanating from district testing offices, that either encouraged or discouraged teaching to the test. Although I do not claim perfect one-to-one correspondence, there is a pattern of association between statements about practices and statements about beliefs.

3. Because Cizek misspecified a model on my behalf, he was led to make other inaccurate inferences from my analysis. Especially he imagines that because I focused on psychometricians' beliefs, I must think they are *unique* among educators in their adherence to a behaviorist learning model. To the contrary, I agree that this belief system is widely shared by educators and the general public. In the 1991b article, for example, I mentioned that belief in the sequential facts-before-thinking learning model is "consistent with the public's understanding of the immutability of grade level achievement, requiring grade retention as the only remedy to deficient skill acquisition" (p. 6). Also on page 6, I speculated that so many measurement specialists might talk in similar terms because they shared "the same training in the educational psychology of a particular era." By implication, other educators who participated in the same training could be expected to hold similar views.

In a different 1991 publication, I considered at length how outmoded psychological theories—regarding immutable, inherited ability and bit-by-bit sequential learning (the behaviorist model)—underlie educational practices such as tracking, segregated special education placements, grade retention, and 2-year kindergartens (Shepard, 1991a). For example, according to survey data, the great majority of educators practice retention and believe it to be effective; "lack of basic skills" is the most frequently cited reason for retaining students in grade (Byrnes, 1989). In contrast, research evidence from 63 controlled studies shows that retention tends to harm achievement as well as self-esteem (Holmes, 1989). For the practice to persist in the face of overwhelmingly negative evidence suggests the potency of underlying belief systems. Examined at closer range, teachers' beliefs about discrete skills and the linear sequencing of learning are linked to retention practices just as psychometricians' beliefs presage the nature of their tests. Smith and Shepard (1988) found schools with high retention rates to be more bureaucratic and rigidly segregated between grades; teachers delivered a standard curriculum that could be broken down into small learning activities, sequenced, drilled until mastery, and tested in a standardized way. Students all had to be functioning at the same level or were sent elsewhere. In contrast, teachers in low-retaining schools worked more cooperatively between grades and espoused views of child development that "accepted the possibility of spiraling, sudden reorganizations, intuitive leaps to understanding . . . and other unpredictable paths to learning" (Smith & Shepard, 1988, p. 329). Teachers with these beliefs kept children working with their peers on some tasks while using tutoring and other focused interventions to address learning difficulties.

The Smith and Shepard data help make Cizek's point that many teachers and psychometricians share the same training and perspectives. But the argument that the character of tests influences instruction does not depend on the uniqueness of psychometricians' beliefs. In the above study, the principals and teachers in the high-retaining schools found a good fit between their conceptions of achievement and what was measured on the tests, as evidenced by the greater salience of talk about standardized tests in the high-retaining versus the low-retaining schools. Perhaps one could speculate that tests have greater impact when belief systems coincide than when teachers resist attending to the test. In other research, however, teachers lament that they have had to abandon professionally defensible practices in deference to the demands of standardized materials and standardized tests (Hatch & Freeman, 1988).

4. Cizek also accuses me of casting measurement specialists as the "bad guys" because they are "nerdy," "use calculators, enjoy mathematics, require corrective lenses," and make tests that tell us unwelcome news. Cizek made up these insults apparently because he didn't like the substance of my argument. The *ER* article, originally presented as my vice presidential address to Division D of AERA, was intended to improve psychometric theory and practice by promoting debate within the community about the role of testing in curriculum and school reform. It can hardly be classified as wild-eyed test-bashing by an outsider. After twenty-some years as part of the psychometric community, I was addressing my colleagues, former students, and some of my closest friends. If considered and thoughtful criticism from a member of the community is met with portrayals of test-bashing bias, how can the field ever debate critical issues or hope to reform? It was exactly this type of extremism that I sought to avoid by my appeal in the concluding section of the original article.

New Points for Discussion

Although I consider Cizek's essay to be off target as a response to my previous article, he makes two important points that I am willing to discuss directly. First we have the issue of who controls test content; is it psychometricians or psychometricians and curriculum specialists jointly? Then it is a separate question whether test content, from whatever source, should drive instruction. These two issues are entwined in the psychometric hegemony model, which Cizek uses as a foil for his arguments, but they are distinct and should not be treated interchangeably.

1. With his Figure 2, Cizek argues that psychometricians

should act in a *service* role, offering their technical expertise and working collaboratively with curriculum specialists in test development. While I disagree strongly with his assertion that this has always been so, I agree that psychometricians should not have the dominant role in test development. I also agree that explicit consideration of the relative authority of content and technical expertise is an important issue to raise in the context of current efforts to make fundamental changes in the nature of assessment.

To be sure, subject-matter experts have always been called in to develop content frameworks and write items for achievement tests. In my experience, however, psychometricians have been center stage in the development process, defining the rules of test making and compiling the final instrument. Two indicators in support of my conclusion are (a) the frequency of psychometricians (versus subject-matter experts) who were hired as permanent staff by test publishers and state testing offices and (b) the relative frequency of psychometricians as *authors* of standardized tests. The Iowa Test of Basic Skills, for example, is authored by three psychometricians. Furthermore, according to Rudman (1987), the trend is for test publishers, who once had both measurement and curriculum specialists as authors, to phase out the author model in favor of an "in-house" procedure where test development is controlled by "psychometricians, research specialists, and editors" (p. 9) employed full-time by the publishing house.

In these traditional test development arrangements, the transitory participation of subject-matter experts meant that they could contribute pieces to the test but did not necessarily have much say about its overall character. In Rudman's in-house model, for example, "item writers are given test specifications and are hired to write items—often on a pay-per-item basis" (p. 9). Even pertaining to the present-day development of the National Assessment, lack of coherent oversight by content specialists led the National Academy of Education evaluation panel to recommend that the development process be revised to include "knowledgeable persons from the relevant subject areas at all steps of the NAEP development process" (National Academy of Education Panel, 1992, p. 30). A mechanism should be devised "to ensure continuity throughout, beginning with the development of the framework and continuing through the various stages of item specifications, item writing, item scoring, and reporting of the results" (National Academy of Education Panel, 1992, p. 30).

If subject-matter experts had always had equal say, how can we imagine that teachers of English would have given up essay tests in deference to inter-judge reliability coefficients? In my view, the Michigan Assessment example, which Cizek cites, is an instance of a new and significant departure from the old model. The central role played by reading experts in creating both the Illinois and the Michigan assessments and the role of content specialists in several other new state and district assessments are examples of Cizek's model, but they mark a break from the past and are still relatively rare. It is important to observe from these new examples that when subject-matter experts are given greater authority, they make a noticeably different kind of test, or (depending on which way you draw the "causal" arrows) if you want to make a different kind of test, content experts must be given greater voice. Psychometricians typically do not know enough about content to reconceive assessment in terms of integrated performance tasks. In addition, as suggested previously, some psychometricians may be operating from implicit behaviorist assumptions about learning or may be governed too much by technical constraints (like the IRT local independence requirement); therefore, they tend to be less willing than some content specialists to give up on a vision of content in the form of discrete items.

Note that the very notion of "assembling" a test, following an agreed-upon allocation of items, privileges one conception of content over others. For Cizek to look at current standardized tests and say that they reflect an equal partnership with content specialists means that he does not appreciate the way that content is shaded by the "form" of assessment, how selecting right answers is different from generating answers, how ill-structured problems get left out of the content, and so forth.

If subject-matter experts had always had equal say, how can we imagine that teachers of English would have given up essay tests in deference to inter-judge reliability coefficients?

Lest these statements lead to a new round of misunderstanding, let me emphasize that I am not saying that content is all. Nor am I saying that technical issues like task generalizability should be ignored, especially for individual student scores. I disapprove of the following erroneous statements being made in the name of performance assessment—"one task is enough," "you don't need a conceptual framework if you have good tasks," and "teaching to the specific set of tasks on the test is okay if they're authentic tasks." What is important about Cizek's model is that content experts and psychometricians should work together collaboratively. The final product should be negotiated, taking the perspectives of each group into account. If content experts are full participants, rather than being consulted and then sent home, it is more likely that test content will be pedagogically defensible. Psychometricians will be pushed harder to distinguish between technical constraints that are essential to validity and those that are only efficient or habitual.

2. Cizek appears to be against measurement-driven-instruction, concluding that "MDI should not be the reigning model on various—though especially ethical—grounds." He also calls it "naive" to seek educational reform through assessment reform. However, he gives a glowing account of the effectiveness of MDI in his example from Michigan. Does this mean that Cizek agrees with those who think it's okay to drive instruction with the test, as long as it's a good one?

My own view is that MDI—that is, *forcing instructional change* by means of a high-stakes external test—is a mistaken notion regardless of how authentic the examination appears

to be. Any test, even one composed of valued, direct performance tasks, is never a complete instantiation of intended learning goals. Therefore, any test can be corrupted as a valid indicator of what students know, and teaching to any test instead of a curriculum framework can misdirect instructional effort.

Although I reject using assessment as the principal mechanism to reform education, I argue strongly for improving the content integrity of high-stakes tests *to enable reform of instruction* and greater student learning. Not every argument for authentic assessment should be interpreted as an argument for assessment-driven reform. The evidence documenting the negative influence of traditional multiple-choice tests on what teachers teach and how they teach it is irrefutable. Pressure to raise test scores has caused teachers to abandon essay testing and craft their own classroom quizzes in the image of multiple-choice tests; elementary teachers have slighted social studies and science in deference to tested subjects, and in many classrooms even basic skills instruction has been corrupted to mean recognizing (rather than generating) right answers, filling in the blanks on worksheets, reading short passages and answering multiple-choice questions, practicing editing skills rather than writing, and so on (Darling-Hammond & Wise, 1985; Shepard & Dougherty, 1991; Smith, 1991). Changing the character of high-stakes tests should have a positive effect on teaching and learning if it merely stopped the negative effects of misdirected instruction. This does not mean that I think that the same evidence from research on the effects of standardized testing can be used *to prove* that performance assessments will have large positive effects on the order that is sometimes claimed. A great deal will depend on the validity and integrity of new measures in representing the full set of learning goals and on teachers' knowledge about how to attain those goals.

The actual effects and side effects of new forms of assessment will have to be empirically investigated using very nearly the same set of research questions that guided research on the consequences of standardized testing. For example, in addition to the negative effects of MDI on what gets taught, research on existing high-stakes testing has shown negative effects on students and teachers. When there is pressure to raise test scores, hard-to-teach children are more likely to be retained in grade, referred to special education, or drop out of school (Shepard, 1991c), and the professional status and knowledge of teachers are demeaned (Smith, 1991). Although there is some promise for the staff development efforts accompanying new assessments, it remains to be seen whether harmful effects for teachers and students will be re-created by new testing programs.

Summary

Cizek disputes my conclusions that psychometricians' beliefs shape testing practice and that, in turn, high-stakes tests affect instructional practice. By trying to represent my arguments in a causal model, Cizek is led to erroneous inferences: (a) that beliefs and actions can be neatly separated and (b) that psychometricians must be unique among educators in their acceptance of a behaviorist learning theory. Most seriously, Cizek infers that if I saw a causal link, I must be proposing a direct intervention. He thinks I mean to fix education by fixing the thinking of psychometricians and casts me with those who would use authentic assessment to drive educational reform. He calls my line of inquiry (or the line of inquiry I succumb to) a "theoretical, practical, and financial dead end."

I never supposed that investing resources in retraining psychometricians would be the best way to improve education. I did suggest that critically examining underlying assumptions is essential to rethinking both what measurement specialists do when we devise tests and how we should evaluate the effects of our efforts. Belief systems do affect what we accept as evidence of test validity and test consequences. Popham, Cruse, Rankin, Sandifer, and Williams (1985), for example, were willing to take scores on the same taught-to tests as proof that measurement-driven instruction worked; in contrast, Koretz, Linn, Dunbar, and Shepard (1991) gave different tests to assess the effects of high-stakes testing on learning. It doesn't cost much to attempt the kind of critical reflection and debate I proposed to my colleagues in the measurement community. Despite having been misunderstood, I still think it's worth trying for greater understanding.

Note

My thanks to Mary Lee Smith and Sam Wineburg for their comments on a draft of this response.

References

Byrnes, D. A. (1989). Attitudes of students, parents and educators toward repeating a grade. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention.* London: Falmer.

Darling-Hammond, L., & Wise, A. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85,* 315–336.

Hatch, J. A., & Freeman, E. B. (1988). Who's pushing whom? Stress and kindergarten. *Phi Delta Kappan, 69,* 145–147.

Holmes, C. T. (1989). Grade level retention effects: A meta-analysis of research studies. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention.* London: Falmer.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

National Academy of Education Panel on the Evaluation of the NAEP Trial State Asssessment. (1992). *Assessing student achievement in the states: The first report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment.* Stanford, CA: National Academy of Education.

Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66,* 628–634.

Rudman, H. C. (1987). The future of testing is now. *Educational Measurement: Issues and Practices, 6*(3), 5–11.

Shepard, L. A. (1991a). Negative policies for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. H. Hiebert (Ed.), *Literacy for a diverse society: Perspectives, practices, and policies.* New York: Teachers College Press.

Shepard, L. A. (1991b). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2–16.

Shepard, L. A. (1991c). Will national tests improve student learning? *Phi Delta Kappan, 73,* 232–238.

Shepard, L. A., & Dougherty, K. C. (1991, April). Effects of high-stakes testing on instruction. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Smith, M. L. (1991). Put to the test: The effect of external testing on teachers. *Educational Researcher, 20*(5), 8–11.

Smith, M. L., & Shepard, L. A. (1988). Kindergarten readiness and retention: A qualitative study of teachers' beliefs and practices. *American Educational Research Journal, 25,* 307–333.