Using Multiple Measures to Address Perverse Incentives and Score Inflation

Daniel Koretz, *Harvard Graduate School of Education*

The principle that important decisions should not be based on a single measure is axiomatic, if widely ignored in practice. The traditional rationale is the risk of incorrect decisions from incomplete and error-prone data. The current high-stakes uses of test scores increase the need for multiple measures for two distinct reasons: the risk of score inflation and the potential for perverse incentives for educators and students. Addressing these two issues may require focusing accountability on measures of schooling as well as a much wider range of measures of student outcomes. The difficulties of pursuing this approach are described, and some possible directions for research and development are noted.

Keywords: accountability, incentives, multiple measures, score inflation

It has long been axiomatic in the field of educational and psychological measurement that important decisions should not be made on the basis of a single measure. For example, the current edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states:

> Standard 13.7. In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision. (p. 146)

This principle is of course widely ignored in practice. Roughly half the states now have policies making high-school graduation contingent on performance on a single test or battery. Several large districts make promotion between grades contingent on exceeding a cut score on one test battery, and some states (including Florida, Georgia, and Texas) have recently taken steps to implement such promotional-gates testing (Harrison, 2003; Keller, 2001). Both of these policies appear to violate the spirit of Standard 13.7. Indeed, the proportion of large-scale testing programs that seem to violate this principle has been growing rapidly in recent years.

Nonetheless, the principle that important decisions should be based on multiple measures is gradually becoming more prominent in the policy debate about large-scale assessment and accountability. This principle has been noted in federal statute for some years and is referenced in Section 1111 of the recently passed No Child Left Behind (NCLB) bill, which stipulates that the newly mandated annual testing in grades 3 through 8 should use assessments that "involve multiple up-to-date measures of student academic achievement." There has been considerable discussion in the policy and research communities about how multiple measures might best be employed—for example, what measures might be used and how they might be combined. As yet, we have no clear models to follow. In addition, there has been some debate about the actual meaning of the requirement for multiple measures and what types of testing systems meet it—for example, whether offering students multiple chances to take an exit examination meets the requirements of Standard 13.7.

The traditional discussion of the importance of multiple measures focused on the risk of incorrect decisions based on incomplete and error-prone cognitive measures. Recently, as test scores have been used increasingly to estimate the quality of schools, the notion of multiple measures has been expanded to include a modest number of measures other than tests of student achievement, such as school-level dropout and attendance rates.

Because of the currently pervasive use of high-stakes testing to create incentives for teachers and students, however, we may need to broaden markedly our view of multiple measures to address two issues of fundamental importance. The first and more manageable of these issues is that the high-stakes use of scores raises additional threats to the validity of score-based inferences, and multiple measures may be needed to address these threats. The second and seemingly far more challenging issue is that reliance on one measure or a narrow set of achievement measures appears to create a complex and not entirely desirable mix of incentives for teachers. Devising a system of multiple

Daniel Koretz is Professor, Harvard Graduate School of Education, 415 Gutman Library, 6 Appian Way, Cambridge, MA 02138; e-mail: daniel_koretz@harvard.edu. His areas of specialization are measurement, education policy, and social context of education.

measures for purposes of accountability that will produce a more consistently desirable mix of incentives may require major and difficult changes from current practice. For example, to improve incentives for teachers, it might be necessary to consider measures of schooling as well as multiple measures of student performance, and the former raise formidable problems of measurement.

This article describes the risks that arise in the current context from insufficient use of multiple measures. It describes a number of possible approaches for the use of multiple measures that might address these risks. Some, like the use of audit tests, have been attempted and appear practical. Others, such as the use of direct measures of educational practice, are likely to be far more difficult and may even prove impractical. The measurement and incentive issues raised by high-stakes uses of tests are critical, however, and this article therefore argues less for a specific approach than for an agenda of research and development on the use of multiple measures to address the incentive effects of high-stakes measurement systems.

Traditional Concerns About Multiple Measures

The long-standing dictum against making important decisions based on a single measure appears to rest on a concern about incorrect decisions based on a necessarily incomplete and error-prone measure. For example, the *Standards* illustrate Standard 13.7 with the example of screening students for special placements and warn that the screening may require "more comprehensive evaluation" than the original test (American Educational Research Association et al., 1999, p. 146).

The basis for this principle, however, has received surprisingly little discussion, and there does not appear to be a consensus within the measurement community about its meaning or justification. The one basis for this principle that is universally acknowledged in the profession is simple measurement error. This requires no discussion here, except perhaps to note that decision inconsistency remains substantial even at levels of reliability conventionally considered high (e.g., Wainer & Thissen, 1996). Moreover, decision inconsistency is exacerbated when cut scores are set at the moderate levels characteristic of many current testing programs rather than the low levels characteristic of earlier minimum-competency testing programs. Table 1, taken from Klein and Orlando (2000), shows the probability that two instances of measurement will yield inconsistent pass-fail decisions as a function of the reliability of the test and the level of the cut score. Unless the cut score is set extremely low or high, even highly reliable tests will result in considerable decision inconsistency. Even a test with a reliability of .9 will result in inconsistent decisions for 12% or more of students if the cut score is placed so as to pass anywhere between 30% and 70% of test takers.

However, simple measurement error is not the only threat to the robustness of an inference or decision based on a single test. Differences among tests present another reason to use multiple measures. Because tests of large domains are necessarily incomplete, alternative measures of the same domains will often classify individuals quite differently (e.g., Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Inconsistencies among procedures for setting standards or cut scores provide another justification for multiple measures. The setting of performance standards is inherently judgmental and to some degree arbitrary, and as a consequence, different methods of setting standards can result in substantial differences in the percentages of students deemed to be passing (e.g., Heubert & Hauser, 1999; see also Jaeger, 1989, pp. 497–500). The necessary incompleteness of achievement tests as a genre of evaluation presents yet another reason to avoid decisions based on a single measure. Many aspects of student achievement are difficult to test, and some students will perform better or more poorly on on-demand tests than on other important measures of performance, such as writing an analytical paper or applying knowledge in real-world contexts. And of course many people hold that outcomes other than achievement as such—such as an interest in learning that students will take with them after they leave school—are important goals of schooling as well.

When one goes beyond simple measurement error, however, one finds that consensus about the importance of multiple measures for drawing inferences about individuals breaks down. For example, Koretz (2001) noted several of the reasons mentioned above and argued that providing multiple opportunities to pass a single test does not meet the requirement in Standard 13.7 for multiple measures because it addresses only one of the limitations of reliance on a single measure. In contrast, Schafer, writing about the *GI Forum v. Texas*

Table 1. Percentage of Students Whose Pass/Fail Status Would Be Changed with a Second Testing at Various Combinations of Passing Rate and Score Reliability

	Score Reliability									
Percent Passing	**.00**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**
90	19	17	17	16	14	13	12	11	9	6
80	32	30	28	27	25	23	20	17	14	10
70	42	39	37	35	31	29	25	22	17	12
60	48	45	42	39	36	32	29	25	20	14
50	50	47	44	40	37	33	30	26	21	14
40	48	45	42	39	36	32	29	25	20	14
30	42	40	38	35	32	29	26	22	18	13
20	32	30	28	27	25	23	20	18	14	10
10	19	18	16	16	15	13	12	11	09	6

Reprinted with permission from Klein and Orlando (2000).

Education Agency lawsuit against the exit examination system then in effect in Texas, wrote that "The AERA, APA, and NCME (1999) standard that a single test score should not be used for a high-stakes decision is not violated if multiple opportunities exist for the student to pass" (Schafer, 2000, p. 414). The aim here is not to resolve this argument, but rather simply to note that disagreement about this issue persists. Even when the issue is only the accuracy of decisions about individuals, the profession does not show a consensus about the reasons for or the importance of multiple measures.

As large-scale assessments increasingly came to be used to support inferences about the quality of schools and to provide incentives for school improvement, the discussion of multiple measures broadened modestly, and some states (e.g., Kentucky and Maryland) began measuring noncognitive student outcomes such as dropout, attendance, and retention rates. It is important to note, however, that within the policy community, these outcomes generally receive far less attention than test scores and are often portrayed as a way of bolstering the utility of scores (e.g., by making it more difficult for schools to boost scores inappropriately by allowing students to drop out). Moreover, these measures receive less weight in accountability systems than do scores. This new emphasis therefore represents only a modest departure from the traditional focus on multiple measures as a means of improving the measurement of the performance of individuals.

Types of Outcome Measures

To discuss a broadening of the notion of multiple measures, a terminology is needed to describe the various types of measures that may be useful in an accountability system. The terms *external* and *internal* tests are used here to refer to tests that are either imposed by agencies outside of the school (external tests) or devised or selected by school personnel (internal tests). It is important to note that external tests need not be selected and mandated by states. For example, Nebraska's unusually decentralized state testing program entails performance standards that are set by the state but that are measured using assessments chosen by districts. Districts may select from "norm reference [*sic*] tests, criterion reference [*sic*] assessments, or locally developed classroom assessments" (Nebraska Department of Education, 2002, p. 4). To the extent that districts choose either of the first two options, the tests would be considered external tests for the purposes of this discussion. In contrast, the partially decentralized assessment system now under development in Maine will include both state and local components, and the expectation is that the latter will include internal measures (Maine Comprehensive Assessment Advisory Committee, 2000).

Distant outcomes are aspects of achievement that are far removed in time from many of the immediate concerns of most teachers. Distant measures include tests designed to assess these outcomes. Distant measures are typically external tests, although external tests need not be distant measures. One example of a distant measure would be what Stecher and Barron (1999) labeled "milepost" testing, that is, testing that is carried out only in a handful of grades. For teachers in other grades, these outcomes are quite remote and compete for attention with more immediate goals.

Intermediate outcomes are less far removed in time from the immediate concerns of teachers but nonetheless are often tested with external assessments. Examples of intermediate measures would be state end-of-course tests or the Advanced Placement tests. These are intermediate measures for the teachers teaching the pertinent classes, although they are distant measures (or even simply irrelevant) for other teachers.

Proximate outcomes are the short- and moderate-term products of schooling that occupy much of the daily attention of many teachers. These are diverse and include not only increases in knowledge and skills but also changes in motivation, other attitudes, and behavior. While in theory, measures of proximate outcomes might be external, in practice, most will be internal. Some proximate outcomes may play a central role in the evaluation of schools and teachers by students, parents, and educators. Some of these proximate outcomes will be reflected in performance on intermediate and distant measures of achievement; others will not be, or will be only faintly echoed. To the extent that these proximate outcomes are important aspects of educational quality and are not well reflected in distant and intermediate outcomes, a failure to incorporate them into an accountability system may distort incentives and undermine the quality of schooling.

All of these classes of outcomes may be important for an effective outcomes-based accountability system, and if so, all are pertinent to a discussion of multiple measures.

Current Uses of Scores and Their Implications for Multiple Measures

The uses of large-scale assessments of achievement—and the range of inferences they are used to support—have evolved markedly over the past several decades. One change has been the growing use of students' scores to evaluate schools. Many traditional standardized achievement test batteries were designed to provide diagnostic information, not to evaluate entire school programs. For example, the manuals for educators accompanying the Iowa Tests of Basic Skills have long warned explicitly against using students' scores to judge entire school programs (see, for example, Hoover et al., 1994, p. 13). The use of test scores to evaluate schools is now so pervasive, however, that it receives little comment.

The fact that many key inferences based on scores now pertain to schools rather than students is itself a sufficient reason to look beyond measures of student achievement. It has long been known that noneducational factors such as student background predict much of the variance in performance among schools, and typical test score databases are ill-suited to disentangling their effects from the effects of educational quality. These databases are usually cross-sectional, contain very little information about differences in intake characteristics, and include little information about other factors (such as student transience rates) that can affect scores. In recent years, some have advocated using value-added modeling of longitudinal test score data to gauge school quality, in part with the view that doing so would free estimates from the conflated influences of student background characteristics. Recent work, however, suggests that current value-added models may not adequately address this confounding. When only characteristics of individual students are pertinent and students are close to randomly distributed among schools, some current models may be reasonably robust to the omission of those variables. However,

when students are clustered in schools in terms of variables correlated with achievement, and when teachers mix little across these clusters, current models may leave the impact of teaching confounded with the effects of background characteristics (McCaffrey et al., in press). Therefore, it is likely that accurate measurement of school quality will require the use of additional measures beyond student test scores.

More important for the present discussion, however, are several other major changes in the use of tests: the great increase in the stakes attached to scores, the focus on changes in scores as a measure of quality, and the now almost ubiquitous intent to use tests to generate incentives. Nearly 20 years ago, when Popham, Cruse, Rankin, Sandifer, and Williams (1985) published their seminal paper on measurement-driven instruction, the use of testing programs primarily to create incentives was controversial; now it is so pervasive that it generates little debate. Kentucky's accountability system, with rewards and sanctions for schools based on changes in test scores, was pathbreaking when it was first implemented in the early 1990s, but such programs had become commonplace only a decade later. The rapid evolution of this use of scores in state policy was codified last year by the new federal statute, No Child Left Behind, which mandates that the assessment systems used to meet requirements of the act:

> include sanctions and rewards, such as bonuses and recognition, the State will use to hold local educational agencies and public elementary schools and secondary schools accountable for student achievement and for ensuring that they make adequate yearly progress. [Sec. 1111 (b)(2)(A)(iii)]

These uses of tests raise issues of both inferential and consequential validity. One of the most important score-based inferences in today's large-scale testing systems—indeed, arguably the most important inferences—pertain to increases in scores, and thus the validity of inferences about these increases is now of paramount importance. Given that the primary motivation for these uses of tests is to provide incentives to educators to improve schooling, a primary question about consequential validity is the actual incentive effects of these systems for educators. Both of these questions pose important and in some cases difficult questions about the use of multiple measures.

Current Uses of Tests and the Validity of Gains

Empirical research on the validity of score gains on high-stakes tests is limited, but the studies conducted to date show that under these conditions, scores may become highly inflated by inappropriate test preparation and other factors (e.g., Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar, & Shepard, 1991). These studies typically compare trends in scores on a high-stakes test and a lower-stakes test that serves as an audit measure; the logic is that validity of the inference about gains requires some, but not complete, generalizability of gains over measures of the same domain (e.g., Koretz, 2002b; Koretz & Barron, 1998). In some cases, studies have found gains on high-stakes tests several times as large as those on an audit test. For example, Koretz and Barron (1998) and Klein et al. (2000) found that gains on the National Assessment of Educational Progress (NAEP) were far smaller than the gains observed on Kentucky's KIRIS assessment and the Texas TAAS test, respectively. In some cases, studies have found sharp gains on a high-stakes test accompanied by no gain whatever on an audit test. For example, Koretz and Barron compared trends on the Kentucky KIRIS assessment to trends on the ACT, considering only the students (nearly half of the cohort at that time) who took both tests. They found a gain of nearly 0.7 *SD* in mathematics on the KIRIS assessment, accompanied by no gain in mathematics on the ACT. One study (Koretz et al., 1991) employed an experimental design in which randomly selected classrooms were administered one of several tests in addition to the one used by the host district. The district had changed tests 4 years earlier, and the switch to the new test had been accompanied by a drop of about half an academic year in mathematics. Four years later, scores on the new test had reached the level previously reached on the old test. When Koretz et al. (1991) readministered the test that the district had used until 4 years previously, they found that scores on that test had dropped by roughly half an academic year.

Although the few available studies of the validity of gains under high-stakes conditions show sizable inflation of scores, they are far too few to indicate how common this problem is or how inflation (or, conversely, the validity of gains) is distributed across types of schools, types of students, or types of assessment and accountability programs. Nonetheless, they warrant two conclusions relevant to the present discussion. First, the distortions caused by this inflation need not be limited to incorrect inferences about overall progress. The responses to test-based accountability appear to be highly variable, and therefore there is no reason to expect score inflation to be uniform across schools or students. Accordingly, it could produce relative errors in classification. Second, score inflation can be very large, and therefore it has the potential to create distortions far larger than those created by the measurement error in a reasonably reliable test.

Current Uses of Tests and Incentives

Research on educators' responses to high-stakes testing is also limited, although it is far more copious than research on the validity of gains. This research suggests that the incentives created by current systems are not entirely what is desired, and they provide some clues about the mechanisms underlying score inflation. Stecher (2002) provided a concise review of many of these studies. Stecher (2002) cited a number of studies that showed positive effects on behavior, such as encouraging desired changes in instructional practice and encouraging teachers to allocate resources to new elements of a curriculum. However, Stecher's review noted numerous undesirable effects as well. For example, he cited numerous studies showing that teachers take time away from valued parts of the curriculum that are not emphasized on the test, as well as studies that showed emphasis on undesirable forms of test preparation. In some instances, administrators shifted teachers among grades to put less able teachers in untested grades. In addition, consequences may increase the frequency of inappropriate testing practices. Koretz, Barron, Mitchell, and Stecher (1996) and Koretz, Mitchell, Barron, and Keith (1996) conducted state-representative surveys gauging the responses of educators to the testing programs then in effect in Kentucky and Maryland, which were chosen in part because the stakes attached to scores were at that time considerably higher in Kentucky than in

Maryland. Teachers in Kentucky were substantially more likely to report having witnessed inappropriate testing practices in their own schools.

The possible link between some of these behavioral responses to testing and score inflation seems clear, but research to date has not progressed far in tying educators' specific behaviors to the validity of gains. Koretz, McCaffrey, and Hamilton (2001) suggested a taxonomy of seven categories of response that vary in their effects on the validity of score gains. *Teaching more, working harder,* and *working more effectively* all could produce unambiguously meaningful gains in scores. At the other extreme, *cheating* can only produce spurious gains. Between these extremes lie classes of behavioral response that can generate either meaningful or inflated gains. *Reallocation* refers to a transfer of instructional resources (such as time) from material de-emphasized on the test to material emphasized on the test. This can occur between subject areas (e.g., reducing instruction in science or music to permit more time in mathematics) and within subjects (e.g., taking time away from aspects of algebra not emphasized by the test to concentrate more on emphasized material). Reallocation of resources causes a reallocation of achievement; whether that reallocation inflates scores depends on the importance of both the emphasized material and the de-emphasized material for the inferences that users base on scores. *Alignment,* which some observers believe protects against all sins, is simply a special case of reallocation, and it too can inflate scores if it causes teachers or students to reduce emphasis on important material. Finally, Koretz et al. (2001) used *coaching* to refer to focusing instruction on narrow, specific aspects of the test. For example, if a teacher notices that the high-stakes test always uses regular quadrilaterals and triangles in area and perimeter problems, he or she may decide not to use irregular polygons or figures with more than four sides in instruction. Coaching may also focus on aspects of the test that are entirely nonsubstantive, such as aspects of item format.

Although it should be possible to link the behaviors of educators to score inflation, it would be a mistake to consider the validity of gains the only reason to be concerned about possibly undesirable incentives. For example, consider a hypothetical teacher who opts for one type of response that Koretz et al. (2001) maintain should produce meaningful score gains: working harder. She increases the pace of work in class and assigns very large amounts of homework, much of it difficult and all of it graded. To parallel these changes, she decides to make her own internal tests more difficult. In some instances (depending, for example, on the initial level of demand in the class), this might be all for the good, but in other classes, it might be excessive and might lead many of the students to react to the subject matter with anxiety and repugnance. Many would question whether the latter outcome represents improved education, even if it leads to a meaningful increase in test scores.

Possible Uses of Multiple Measures to Address Incentives and Score Inflation

Although extant research is sufficient to document problems of score inflation and unintended incentives from test-based accountability, it provides very little guidance about how one might design an accountability system to lessen these problems. More extensive and creative uses of multiple measures may prove to be a key to addressing these problems. This section suggests several possible uses of multiple measures to address both problems.

Using Distant and Intermediate Outcome Measures

One use of multiple measures that would help address problems of both incentives and the validity of gains would be to pair distant high-stakes tests with distant or intermediate audit measures. Audit measures are assessments that are less prone to score inflation from inappropriate test preparation. These may be independent tests (e.g., NAEP), but it may also be feasible to use material embedded in operational, high-stakes assessments for this purpose.

This use of multiple measures is simple in concept but may prove complex in practice. The audit measure must be constructed to support inferences similar to those based on the high-stakes test, but at the same time, it must be dissimilar enough that the effects of inappropriate preparation for the high-stakes test will not generalize strongly to it. These two criteria are of course in tension, and how they can best be balanced is not yet clear. For example, one might begin construction of an audit measure by looking for content deemed important but given little emphasis on the high-stakes tests, and conversely by looking for material that has received inadvertent overemphasis on the high-stakes test. (See Koretz et al. [2001] for more discussion of inadvertent "overweighting" of content.) However, it may also be necessary to design audit tests to take into account more subtle aspects of the style of items used in the high-stakes assessment (e.g., Koretz, 2002a).

To the extent that audit measures address the possibility of score inflation, they should also improve the incentives facing teachers. That is, if audit measures are successful in capturing the effects of inappropriate narrowing of instruction and other forms of inappropriate test preparation, the accountability system should give teachers less of an incentive to engage in these undesirable behaviors.

The possible use of audit testing on a large scale has begun to receive some attention in the policy community. NCLB took a tentative step in this direction by requiring that all states participate in the state-level NAEP, but it did not in the end specify that NAEP was to be used to evaluate gains on states' high-stakes tests. Discussion about the use of NAEP in NCLB, however, did show a lack of consensus in the profession about how NAEP should be used as an audit measure, if the decision is eventually made to use it in that way. At one extreme, most extant studies of score inflation (e.g., Klein et al., 2000; Koretz et al., 1991; Koretz & Barron, 1998) have focused on the magnitude of disparities in trends between a high-stakes tests and an audit test, interpreting a large disparity as evidence of score inflation. In contrast, in a recent statement, the National Assessment Governing Board (NAGB) did not address the size of a disparity in trends, instead arguing that "any amount of growth on the National Assessment should be sufficient to 'confirm' growth on state tests" (National Assessment Governing Board, 2002, p. 9).

The root of this disagreement can be clarified by focusing on the example of possible differences in item formats between a high-stakes test and an audit test. The NAGB statement argued that differences between the results of the tests could stem from any number of factors, including differences in item format as well as differences in difficulty, reporting metrics, motivation,

and standard-setting approaches. The NAGB statement does not focus specifically on inferences users base on scores or on the relevance of format (or the other listed factors) to these inferences. However, it is a consensus in the profession that validity is an attribute of an inference, and therefore Koretz et al. (2001) argued that the disparity in trends between high-stakes and audit tests must be evaluated in the context of the specific inferences that users base on scores on the high-stakes tests. While differences in format may help explain a difference in performance between a high-stakes test and an audit test, they may not undermine the usefulness of the audit test for evaluating the gain in scores on the high-takes test. For example, Shepard (1988) showed that format differences can have a major impact on performance. She noted that on one state test during the 1970s, the p values for subtraction items presented in a vertical format were far higher than those for items presented in a horizontal format. Suppose one found a similar difference in performance between a format used on a high-stakes test and another format used in NAEP or another audit test. Whether this disparity would undermine the validity of inferences based on the high-stakes test would depend on the nature of the inferences. If the inferences assume (even tacitly) that performance increases are not limited to the particular format used on the high-stakes test, then the disparity in performance across formats does undermine the validity of the inference. The same logic would apply to all of the other characteristics of tests noted by NAGB.

Using Proximate Measures of Achievement

To the extent that external distant and intermediate measures of achievement are as a set insufficient to measure educational quality, improvements in their use will not be sufficient to avoid undesirable incentives. In theory, using proximate measures might be a method of addressing these limitations and their effects, but in practice, it may prove extremely difficult to incorporate these measures into centralized accountability systems.

To what degree are external, distant and intermediate achievement tests insufficient to evaluate educational quality? In the present environment, arguments about the limitations of external tests for this purpose often seem to be the province of opponents of standardized testing, but it is worth recalling that these limits have long been emphasized by some of the most prominent developers and proponents of standardized testing. For example, in his chapter in the initial edition of *Educational Measurement,* E. F. Lindquist, the creator of the Iowa Tests of Basic Skills (ITBS) and the co-founder of ACT, argued that in the ideal case, an assessment of "educational development" would use direct measurement of criterion behaviors and would also extend beyond what he called "intellectual development" and "rational behavior" to include measurement of attributes such as artistic abilities and managerial or executive abilities (Lindquist, 1951). He argued that available tests fell far short of this ideal, in part because some criterion behaviors and some desired outcomes of education are either difficult or impossible to measure. Others have emphasized the inability of external, distant or intermediate tests to supplant more proximate measures. For example, a recent ITBS manual warns administrators that "Though standardized achievement scores cannot and should not replace teacher observations and classroom assessment information, they can provide unique supplementary information" (Hoover et al., 1994, p. 11) and, as noted earlier, warns administrators not to use scores to evaluate entire school programs. Policymakers today clearly pay little heed to such warnings.

The advice offered by Hoover et al. (1994) is that their own distant, external measure of achievement be coupled with internal, proximate measures in educators' internal evaluations. It is less clear, however, how one might incorporate proximate measures of achievement into a centralized, state-level accountability system. Course grades are the most common proximate measure, but the inconsistency of grading standards and the susceptibility of grades to inflation would seem to limit their usefulness in centralized accountability systems unless steps were taken to benchmark them in some manner.

Portfolio assessments have been tried as a method of centralizing and imposing some degree of standardization on measures of the quality of ongoing classroom work. There is some evidence that portfolio assessment systems, while burdensome to teachers, can provide incentives to improve practice (e.g., Koretz, Stecher, Klein, & McCaffrey, 1994; Stecher, 1998). However, portfolio assessments are only partially standardized; scoring generally is standardized, but the specific tasks selected and the conditions under which they are completed are not. Perhaps for those reasons, the evidence about the quality of measurement provided by the few large-scale portfolio assessments in the United States is discouraging (Koretz, 1998; Koretz et al., 1994).

To my knowledge, we have as yet no documented example of a reasonably successful integration of internal, proximate achievement measures into a large-scale accountability system in the United States. The ongoing efforts in a few locales (e.g., the Maine effort referenced above) may provide additional information once fully implemented. Instead of attempting to standardize these measures directly, it may prove more useful to deal with them by means of some form of moderation or by incorporating them into direct measures of educational practice, as described below. Additional research is needed to address this uncertainty.

Using Measures of Other Proximate Outcomes

Many parents, when choosing a school for their children, consider characteristics of schools that may be construed as proximate outcomes. For example, do the students seem engaged? Are they motivated? Are they willing to tackle difficult and ill-structured problems? Do they offer comments and answers without excessive worry about being wrong? Factors outside of school clearly influence these characteristics, but good teachers also strive to encourage them. It is precisely qualities such as these that some critics of test-based accountability argue are being undermined in good schools because of excessive attention to test scores. We have no method for standardizing measures of these characteristics at present, and we may never have them. Ignoring considerations such as these, however, distorts the incentives provided to teachers. This may be another argument in favor of direct measurement of educational practice.

Using Direct Measures of Educational Practice

For a number of years, many policymakers have insisted that educational accountability must be focused on outcomes. Yet if we cannot directly measure

some of the most important outputs of schooling, direct measurement of educational practice may be the only method available for balancing the incentives created by an accountability system.

The arguments against direct measures of practice are numerous and serious. In many instances, there is no consensus about the types of practice that foster achievement and other desired outcomes. (For that matter, there is often not consensus about the desired forms of achievement.) In many instances, a variety of different practices may work, and different teachers—even when faced with similar classrooms—may find different approaches effective. Some measures are "fakable," response bias can distort results of surveys, and teachers have been known to have special lessons in reserve for the day when an observer appears unexpectedly. Measuring practice is expensive, and it is often to some degree subjective as well. Evaluations of teachers that are dependent on expert judgment require that individuals with expertise in teaching be removed from the classroom to conduct evaluations.

The arguments in favor of direct measures of practice, however, are also strong. Perhaps the strongest argument is that we do not have a sufficient range of other measures, such as measures of proximate outcomes, that are sufficient to balance the incentives given to teachers by current accountability systems. Measures of practice are used in other areas of public policy, such as health care. The pressures of test-based accountability suggests a need for further research exploring the feasibility of using direct measures of practice to balance the incentives facing teachers.

Using Measures of Noncognitive Outcomes

Using noncognitive student outcomes, such as dropout and attendance rates, is seemingly uncontroversial and has been tried in several states. In many cases, however, these measures either show little variability or are substantially out of the control of schools, and their utility as measures of school quality may therefore be limited. These measures may prove to be more useful as a means of controlling inappropriate responses to test-based accountability (e.g., attempts to boost scores by retaining students in grade or by allowing low-performing students to drop out).

Discussion

This article has argued that the current high-stakes uses of tests create two additional reasons to rely on multiple measures: the risks of inflated score gains and perverse incentives for educators. It has also argued that in order to create the desired incentives for teachers, we may need to go beyond multiple distant outcome measures and may need to use proximal outcome measures and perhaps direct measures of schooling as well. However, the practicality of integrating these additional measures into a centralized accountability system is by no means clear. As yet, we have no good models of how best to do what this article suggests, and the research base is woefully inadequate. The proposed steps may also be politically difficult, for reasons of financial costs and vested interests of stakeholders.

The least problematic step of those suggested here—and the one that requires the smallest departure from both current practice and the long-standing focus of the measurement profession—would be to design ways to incorporate multiple intermediate and distant achievement measures, including audit tests, into accountability systems. Audit measures created specifically for this purpose are equally feasible at all grade levels, but other uses of distant and intermediate achievement measures may be more feasible at the high-school level because of the range of tests and other outcomes that can be monitored. However, this step will require more than the design or selection of additional measures. It will also require that the measurement profession accept a common framework for evaluating data from these measures (such as gaps in the generalizability of gains) and develop additional methods for analyzing these data (Koretz et al., 2001). Even in the absence of such a common framework, however, the mere presence of audit measures may help balance the incentives faced by teachers.

Addressing proximate outcomes is likely to prove far more challenging. There are at least three paths that might be followed, all of which appear difficult. The first would be to standardize proximal measures sufficiently that they can be incorporated directly into a centralized accountability system. If experience with large-scale portfolio assessments is any indication, this approach is likely to be very hard, if not impractical. A second approach would be to use some form of moderation, perhaps both statistical and social moderation, to benchmark local proximal measures. Whether this can be done effectively and at reasonable cost remains to be seen. A third path would be to limit centralized accountability to intermediate and distant measures but to encourage localities and schools to emphasize proximate measures as well. In this case, the greatest difficulty may be structuring the accountability program to give localities reason to take the local measures seriously.

Adding direct measures of practice to centralized accountability systems would likely be similarly challenging as well as politically difficult. A few jurisdictions are experimenting with this now, for example, using peer review of teachers or surveys of students, parents, or staff. To my knowledge, however, no states have undertaken this, and we have little research examining the effectiveness of alternative approaches.

Incorporating direct measures of educational practice and proximate outcome measures into centralized accountability systems would create problems of measurement that go far beyond reliability in the classical sense of random error, such as variations in standards among groups, individual subjectivity, and bias. Steps might be taken to lessen these problems, but it seems virtually certain that these problems would remain substantially more serious and less tractable in the case of these measures than in the case of many distant and intermediate measures of student achievement.

It is worth noting that the problem of subjectivity in personnel management is not restricted to education and has received considerable attention in economics. For example, Neal (2002) noted that "straightforward incentive systems based on objective standards are often problematic because objective performance standards are often easy to game" (p. 36). He further argued that professionals in other fields are rarely given incentive pay based on "simple formulae tied to an objective performance standard" (p. 37). Of course, the cornerstone of many current education reforms, including NCLB and many state programs, is precisely to base accountability for educators on "simple formulae tied to an objective standard." Rather, Neal maintained, these professionals in other fields are given incentives that are based in part on more subjective

criteria that reflect the multidimensional nature of their roles and that are not easily manipulated by them.

The practicality of moving public education to the type of accountability system that Neal maintains is common in many other professions, and that this article argues for, is debatable. Neal is skeptical because managers in public employment have no financial stake in making sure that subjective evaluations are accurate. Other reformers might respond that structures can be built to give managers a stake in the accuracy of these outcomes. Our concern, however, is not the organizational issues raised by these notions, but rather the issues of measurement they elicit.

From the perspective of measurement, the discussion above suggests a fundamental tension between two goals of measurement in current education policy: maximizing the quality of the measurement of student performance, and creating the best possible mix of incentives for teachers. The question is how the conflicting demands of these two goals can best be reconciled.

This article does not propose any retreat from the goal of maximizing the rigor of intermediate and distant assessments of student achievement. For a decade or more, some reformers have argued that incentives could be made right by using distant and intermediate "tests worth teaching to." These tests often sacrificed psychometric rigor in favor of authenticity of tasks. This strategy did not work. Certainly, steps can often be taken to increase the authenticity of tasks used in distant and intermediate measures, and evidence has also shown that "tests worth teaching to" can encourage desired changes in practice (e.g., Koretz et al., 1994; Stecher, 2002). But research has also shown that even assessments designed to be "worth teaching to" can generate distorted incentives and inflated test scores (e.g., Koretz & Barron, 1998). Moreover, the psychometric costs of going too far in this direction can be high. Currently, the phrase "tests worth teaching to" is sometimes used to refer to tests aligned with standards, rather than to tests emphasizing complex tasks, but as noted above, alignment, while desirable, is insufficient to ensure appropriate incentives.

One possible approach to this fundamental tension—to balance the need for appropriate incentives with the need for high-quality measurement—may be to use multiple, rigorous measures of intermediate and distant outcomes carefully (e.g., by using audit tests for validation) while pairing these with other types of measures. At the same time, intensive research and development would be needed to improve the quality of the lower-quality measures in the system. That some of these other measures, such as direct measures of practice, are likely to be error-prone and somewhat subjective is unfortunate, but the negative effects of unbalanced incentives may offset these drawbacks.

Another possible approach for integrating intermediate and distant achievement measures with less tractable proximate measures and direct measures of schooling would be to use test scores to trigger audits. That is, rather than serving as the end point of evaluation in a centralized accountability system, test scores would signal the need for more labor-intensive and perhaps more subjective forms of evaluation. For example, a school with persistently low scores would be investigated to determine the likely causes of the low scores. Similarly, very rapid rises in test scores could trigger audits to look for inappropriate forms of test preparation, such as inappropriate reallocation or excessive coaching. This approach would seem to have two advantages. First, it would help target remedies to the particular difficulties faced by individual schools. These particular difficulties are often not apparent in a system that relies solely on distant outcome measures but could be revealed by audits. A school with a solid staff but a highly transient student population, for example, would warrant an entirely different response than would a school with a stable student population and a weak staff, even though the two schools might have very similar trends in test scores. Second, this approach would limit the use of the more problematic measures, thus saving resources and allowing these measures to be utilized more carefully.

In sum, the current uses of large-scale assessments in centralized accountability systems pose two risks, already documented by research, that have implications for the use of multiple measures: the possibility of score inflation and the possibility of perverse incentives for educators. It seems likely that the first of these problems can be addressed (although perhaps not completely) by expanding the traditional notion of multiple measures to incorporate multiple distant and intermediate achievement measures. Doing so poses substantial challenges (e.g., designing audit measures to be sufficiently similar to but sufficiently different from high-stakes tests) but appears tractable and would not require a major change in the field's thinking about multiple measures. In contrast, dealing with the problem of perverse incentives may require a more substantial and much more difficult expansion of the notion of multiple measures, perhaps to include both proximal outcomes and direct measures of schooling. The challenges posed by such changes would be far more substantial, and addressing them would require an uneasy compromise between the apparently partially conflicting goals of measurement quality and improved incentives. Clarifying these challenges and evaluating possible responses will require an ambitious agenda of research and development.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Harrison, S. (2003, January 20). Florida set to tighten reigns on promoting third-graders. *The Miami Herald.* (Retrieved February 4, 2003, from http://www.miami.com/mld/miamiherald/living/education/4986683)

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Cantor, N. K., Bray, G. B., Lewis, J. C., & Qualls-Payne, A. L. (1994). *Iowa Tests of Basic Skills interpretative guide for school administrators, levels 5–14.* Chicago: Riverside.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/MacMillan.

Keller, B. (2001, March 28). Ga. OKs social-promotion ban as Texas revisits its own. *Education Week.* (Retrieved February 4, 2003, from http://www.edweek.org/ew/ewstory.cfm?slug=28social.h20)

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND (http://www.rand.org/publications/IP/IP202/).

Klein, S. P., & Orlando, M. (2000). *CUNY's testing program: Characteristics, results, and implications for policy and research.* MR-1249-CAE. Santa Monica, CA: RAND.

Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. In D. Koretz, A. Wolf, & P. Broadfoot (Eds.), *Records of achievement.* Special issue of *Assessment In Education, 5*(3), 309–334.

Koretz, D. (2001, April 13). Examples of standards for accountability programs. In E. Baker (Chair), *Holding accountability systems accountable: Research-based standards.* Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Koretz, D. (2002a, September 11). Believe me, it is not cheating, but some strange method. Invited presentation at the annual meeting of the Center for Research on Evaluation, Standards, and Student Testing, Los Angeles.

Koretz, D. (2002b). Limitations in the use of achievement tests as measures of educators' productivity. In E. Hanushek, J. Heckman, & D. Neal (Eds.), *Designing incentives to promote human capital.* Special issue of The *Journal of Human Resources, 38*(4), 752–777.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU. Santa Monica, CA: RAND.

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *The perceived effects of the Kentucky Instructional Results Information System (KIRIS)* (MR-792-PCT/FF). Santa Monica, CA: RAND.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing: Preliminary evidence about generalization across tests, In R. L. Linn (Chair), *The effects of high stakes testing.* Symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Tech. Rep. No. 551). Los Angeles: Center for the Study of Evaluation, University of California.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *The perceived effects of the Maryland School Performance Assessment Program* (CSE Tech. Rep. No. 409). Los Angeles: Center for the Study of Evaluation, University of California.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lindquist, E. (1951). Preliminary considerations in objective test construction. In E. F. Linquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Maine Comprehensive Assessment Advisory Committee. (2000, June). *Measured measures: Technical considerations for developing a local assessment system.* Augusta, CA: Author.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., Hamilton, L., & Kirby, S. (in press). *Models for value-added modeling of teacher effects.* Santa Monica, CA: RAND.

National Assessment Governing Board. (2002, March 1). *Using the National Assessment of Educational Progress to confirm state test results.* Washington, DC: Author.

Neal, D. (2002). How would vouchers change the market for education? *Journal of Economic Perspectives, 16*(4), 25–44.

Nebraska Department of Education. (2002, June). *STARS: School-based Teacher-led Assessment and Reporting System: A summary.* Author.

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*(9): 628–634.

Schafer, W. D. (2000). *GI Forum v. Texas Education Agency:* Observations for states. *Applied Measurement in Education, 13*(4), 411–418.

Shepard, L. A. (1988, April). The harm of measurement-driven instruction. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Stecher, B. M. (1998). The local benefits and burdens of large-scale portfolio assessment. In D. Koretz, A. Wolf, & P. Broadfoot (Eds.), *Records of Achievement.* Special issue of *Assessment in Education, 5*(3), 335–351.

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S., Hamilton, B. M., Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education.* Santa Monica, CA: RAND.

Stecher, B. M., & Barron, S. I. (1999). *Quadrennial milepost accountability testing in Kentucky* (CSE Tech. Rep. No. 505). Los Angeles: Center for the Study of Evaluation, University of California.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22–29.