DOCUMENT RESUME

ED 353 127 SE 052 970

AUTHOR	Koretz, Daniel
TITLE	Evaluating and Validating Indicators of Mathematics and Science Education. A RAND Note.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	RAND/N-2900-NSF
PUB DATE	92
CONTRACT	SPA-8850377
NOTE	54p.; For related document, see SE 053 053.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Construct Validity; Data Analysis; Data Collection; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; Evaluation Needs; *Evaluation Problems; *Mathematics Education; Measurement; National Norms; National Surveys; *Reliability; *Science Education
IDENTIFIERS	*Educational Indicators

ABSTRACT

Indicators of mathematics and science education progress are important to policymakers and the public in order to evaluate the quality of education. Recent reports have evaluated existing indicators and suggested core aspects of mathematics and science education that an indicator system should evaluate. This note discusses two fundamental aspects of the quality of educational indicators: their reliability and validity. The note is divided into five sections. The first section discusses recent concerns for the weaknesses in the academic achievement of science and mathematics students in the United States and the need for availability and quality of indicators to assess that achievement. The second section discusses the concepts of validity and reliability as criteria to gauge the quality of indicators. The third section examines the threats to the validity and reliability of indicators. The fourth section presents approaches for evaluating and validating indicators. The fifth section presents the conclusions of the report. The report recognizes the complex task of building an adequate system of indicators of mathematics and science education. Simplification is cited as essential in order to inform debate and policy, and at the same time conflicting with the goal of adequate description. The report suggests that the issue of simplification can be addressed by choosing measures carefully, evaluating the systems of indicators, and employing multiple measures of important constructs. (Contains over 80 references.) (MDH)

A RAND NOTE

Evaluating and Validating Indicators of Mathematics and Science Education

Daniel Koretz

RAND

The research described in this report was supported by Grant SPA-8850377 from the National Science Foundation.

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of RAND do not necessarily reflect the opinions or policies of the sponsors of RAND research.

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

A RAND NOTE

N-2900-NSF

Evaluating and Validating Indicators of Mathematics and Science Education

Daniel Koretz

Supported by the
National Science Foundation

RAND

PREFACE

In a previous project funded by the National Science Foundation (NSF), RAND analyzed a wide range of options for improving the nation's indicators of mathematics and science education, evaluating them in terms of considerations such as the quality of the information that would result and the burden they would impose (Shavelson, McDonnell, Oakes, and Carey, 1987). In a subsequent NSF- funded project now underway, RAND is implementing one of the options assessed in the earlier project. The option being taken in the new project is called an "extended patchwork." It entails:

- Creating a patchwork indicator system of key aspects of mathematics and science education—that is, synthesizing diverse existing national data to create indicators;
- Evaluating the existing data and the indicators that can be produced from them; and
- Exploring the potential of other types of data to provide information that the patchwork cannot.

In evaluating indicators, RAND is applying a variety of standards. Other recent reports (e.g., Murnane and Raizen, 1988, and Shavelson et al., 1987) have offered a number of useful criteria, most of which focus on what indicator systems should measure and what types of measures are most likely to be useful. In this project, those criteria will be supplemented by analysis of the reliability and validity of indicators—their accuracy and the adequacy with which they support the inferences commonly drawn from them.

This Note, which is one of several documents from the current RAND project, discusses the validity and reliability of indicators. It first sketches a number of common views of validity and reliability that apply to all measurement, not just social indicators. It then discusses threats to validity and reliability that are likely to be particularly important in the case of educational indicators and describes steps that can be taken to address them and maintain the integrity of an indicator system.

SUMMARY

Indicators of mathematics and science education have become increasingly important to policymakers and the public as debate about the quality of education has intensified. Pervasive concerns about the weaknesses of American education generally, and of mathematics and science education in particular, have led to widespread changes in educational policy and practice. These concerns and policy initiatives have focused attention on existing indicators of educational quality and have sparked many efforts to improve them.

What characteristics should a system of indicators of mathematics and science education have to be both informative and trustworthy? A number of recent reports (Murnane and Raizen, 1988; Raizen and Jones, 1985; Shavelson et al., 1987) have evaluated existing indicators and suggested core aspects of mathematics and science education (e.g., aspects of student learning, curriculum, and teacher qualifications) that an improved indicator system should monitor.

This Note discusses two fundamental aspects of the quality of ~~educational~~ indicators: their reliability and validity. The essential concerns of validity and reliability apply to measurement of all sorts, but they take distinctive forms in the case of educational indicators. Because of the attributes they are supposed to measure, the types of data from which they are constructed, and the uses to which they are put, indicators face particular threats to validity and reliability. These particular threats have implications for both building and evaluating a system of indicators of mathematics and science education.

CONCEPTS OF VALIDITY AND RELIABILITY

Many of the measures that must be used as indicators of mathematics and science education are far more ambiguous than many measures encountered in everyday life. The length of a table or a child's fever are easily measured accurately, and there is no uncertainty about the meaning of what is being measured. Moreover, the choice of measures is inconsequential; those who use unusual metrics (such as inches) can easily transform their results into more common measures (centimeters) with trivial loss of accuracy. Unfortunately, these characteristics are not shared by many educational indicators, even though they often appear deceptively straightforward.

One reason for the greater ambiguity of some educational indicators is that the attributes (typically called "constructs") they are intended to measure, such as "science achievement," are frequently complex, poorly defined, and difficult to measure. Often, they can be measured in various ways that can yield markedly different answers; unlike "inches" and "centimeters," these alternative measures are not always simple transformations of each other. Moreover, the error in those measures—for example, the extent to which a score on a given test would vary from one administration to another because of such irrelevant factors as illnesses and mood—can be large and can be estimated in a variety of ways.

Traditionally, two aspects of these questions of measurement have been distinguished. *Validity* refers to the degree to which a given measure supports the inferences that are drawn from it. Contrary to a common misconception, validity is not inherent in a given measure, even though characteristics of the measure can affect validity greatly. Rather, what is valid—or not—is a particular inference that the measure is used to support. Thus, validity depends not only on the care with which a measure is constructed, but also on the nature of the inference and the way that the measure is used. A measure that is highly valid in one context may be entirely misleading in another. *Reliability* refers to the degree to which a measure is free of errors of measurement. Reliability serves to limit validity; if a measure is recording only noise, it cannot be a valid indicator of anything else.

Although reliability and validity are theoretically distinct, they overlap in practice, because a key consideration in evaluating both is the *robustness* of the information they provide. In some cases, inconsistency of measurement (e.g., inconsistency in the results of a single achievement test administered at brief intervals) is always considered unreliability; in others (e.g., differences in the results yielded by substantially different tests of a single domain of achievement), it may or may not be. In both cases, however, the inconsistency of measurement places severe limits on the inferences that the measures can validly support. For example, if two different algebra tests, both of which are considered equally reasonable samples from the domain of interest, provide markedly different estimates of ethnic differences in performance, that inconsistency undermines the validity of inferences about ethnic differences based on only one of the tests.

The reliability and validity of indicators are often paid little heed. For example, there has never been a comprehensive analysis of the validity or reliability of the background information in the National Assessment of Educational Progress (NAEP), even though it is derived from self-reports of students as young as nine. When such analyses of validity and reliability have been done—as in the case of the High School and Beyond study—other analyses of the data are often conducted with no consideration of the findings.

Yet there is ample reason to be concerned about threats to the validity and reliability of indicators. These threats extend beyond measures that are notoriously error-prone, such as student self-reports. They extend, for example, to achievement tests and to such apparently simple and straightforward indicators as measures of course enrollments. Moreover, the threats to validity grow particularly pronounced when indicator systems become important as a gauge of educational performance.

THREATS TO THE VALIDITY AND RELIABILITY OF INDICATORS

The threats to validity and reliability to which educational indicators are particularly prone stem from the characteristics of individual indicators, the characteristics of indicator systems, and the ways in which they are used.

Indicators are used to support broad inferences. For example, the public debate about educational achievement in recent years has focused in large part on very general questions, such as whether the decline in educational achievement has been reversed, and how much worse the mathematics achievement of American students is than that of Japanese students. This tendency is also evident in the reporting of the most recent NAEP mathematics and science assessments, which emphasize (but are not limited to) broad constructs such as "science achievement." These broad inferences pose a threat to validity because they exacerbate the limitations of any one measure. Put differently, the broader the inference, the more numerous and diverse the possible measures across which results should generalize.

Indicator systems rely extensively on simple measures. One reason is cost. Indicator systems require the frequent collection of large amounts of information, making the routine use of expensive measures—such as longitudinal data on student performance or observational data on instructional style and methods—prohibitive. But simple measures are sometimes weak representations of the domains of interest. Data on course-taking by course title, for example, is easy to obtain, but it provides only a very limited window on the curriculum to which students are exposed (Stecher, 1991).

Indicator systems rely on corruptible measures. Many of the measures incorporated into indicator systems are also *corruptible.* That is, it is possible to change the *measure* without causing a comparable change in the *construct* it is supposed to measure. The corruptibility of test scores is increasingly acknowledged; in some instances, if sufficient emphasis is placed on raising scores per se, one cannot safely infer from an increase in scores a corresponding improvement in the dimensions of achievement that the test is intended to measure (e.g., Koretz, Linn, Dunbar, and Shepard, 1991). This problem can affect other types of indicators as well. For example, coursework indicators can be corrupted if courses are re-labelled or modified slightly to give them the appearance of meeting certain requirements. The problem of corruptibility becomes increasingly serious as indicators become more salient and important, because it is only when indicators are important that students and educators have incentives to manipulate their standing on them.

Some indicators are context-dependent. Indicators can have different meanings in different settings, or from one time to another. In some cases, this variation stems from the corruptibility of indicators, but it need not. For example, schools with low-achieving students are likely to implement new requirements for mathematics or science courses differently than schools with very high-scoring students, and those differences may not be apparent from some indicators.

Indicator systems often include a sparse array of measures. Indicator systems often include few if any alternative measures of a given construct. Policymakers will want to draw inferences about the construct, not about the particular measure, but there will be little or no chance to test whether findings generalize across measures. This is particularly problematic when one of the measures is likely to have been corrupted, but it may be a problem in other cases as well, because the measures that are included may simply be an incomplete representation of the constructs of interest.

Indicator systems can be narrower than the effects of the policies they are intended to monitor. Educational policies can have effects that go beyond the domain they are intended to influence. Although indicators are not well suited to determining the actual effects of policies, they are often designed to monitor changes in the domains that policies attempt to influence. Inferences about changes in those domains may be mistaken if the effects of the policies are broader than the indicator system. For example, one recent study found that pressure to raise test scores in a large school district caused an increase in retention in grade. An inference that the rise in test scores signalled a corresponding increase in the achievement of the student population as a whole would be unwarranted; one would need an indicator of retention in grade to reveal that some portion of the increase in scores probably reflects changes in the selection of students for testing because of increased retention.

APPROACHES FOR BUILDING AND EVALUATING INDICATORS

Although the threats to the quality of an indicator system are numerous and severe, a number of concrete steps can be taken to address those threats in building and evaluating an indicator system.

Assess likely threats to validity and reliability. A first step in improving an indicator system is determining the threats it is likely to face. The range of potential threats is large, but four key questions arise:

- Are the indicators based on data of questionable quality, such as student self-reports of family background variables?
- Are the indicators vulnerable to substantial corruption? Test scores are a prime example, but not the only one.
- How broad are the inferences the indicators will be used to support?
- How broad are the likely effects of the policies the indicators are intended to monitor?

Build indicator systems to counter likely threats. Here again, the number of potentially important steps is large, and the most valuable steps will depend in part on the specifics of the indicator system. Still, three steps are likely to be important in many instances:

- Use multiple measures of important constructs. This is particularly important when the individual measures are likely to be weak or narrow (relative to the breadth of inferences they will be used to support), or when they are likely to be corrupted.
- Link the indicator system to other data. For many reasons, including costs, indicator systems are likely to be incomplete. Linking them to other types of data collection can fill in gaps in the information they provide and offer an invaluable opportunity to validate them.
- Include an appropriate wide range of domains.

Evaluate indicators. Evaluations of indicator quality are essential and should be tailored to the particular threats they face. Adequate evaluation requires addressing a variety of questions and may require the use of data external to the indicator system. Among the most important steps are these:

- Assess the match of the data to its proposed use.
- Apply the traditional criteria of validity and reliability.
- Assess the robustness of the information provided, particularly when inferences are broad or when measures are likely to have been corrupted. The limits of robustness determine which inferences are warranted.
- Assess the corruption of indicators.

CONCLUSIONS

Building an adequate system of indicators of mathematics and science education is a complex task that entails many choices and compromises. Education is an enormously complex and varied enterprise, but only a limited number of aspects of schooling can be included in an indicator system, and they must be represented by a severely limited number of indicators. Simplification is essential if the indicator system is to inform debate and policy, but excessive simplification will lead to misconceptions and misdirected policy. The tension between the conflicting goals of sufficient simplification and adequate description sharpens and shapes the threats to validity and reliability faced by educational indicators.

The threats faced by indicator systems, however, can be addressed in constructing indicator systems, in evaluating them, and also in using them. Unreasonable simplification, for example, can be lessened by choosing measures carefully, by relying on multiple measures, and by linking indicator systems to other data. Evaluations that take the particular threats faced by indicator systems can be carried out periodically over the lifetime of the system. Equally important, users of indicator systems must be cognizant of their inherent limitations and should avoid using them to support overly broad, unduly simple, or otherwise unwarranted inferences.

CONTENTS

1. INTRODUCTION

Over the past decade, indicators of mathematics and science education have become more important to policymakers and the public. By the early 1980s, concern about weaknesses in the achievement of American students—manifested in declining test scores during the 1960s and 1970s and in the relatively poor standing of American students in international comparisons—had reached a level not seen for decades, and it continues unabated today. Apprehension about the competitiveness of the American economy and workforce exacerbated these concerns and contributed to a particular focus on several subjects widely thought to be important in that respect, especially mathematics and science. These concerns and the political responses to them made indicators of the condition of education more salient and led to many efforts at the national, state, and local levels to increase both the quantity and relevance of available indicators. Tc take only one example, the Council of Chief State School Officers is currently conducting a multiyear project that will increase the availability of indicators of mathematics and science education that are comparable among states.

Several recent reports have described ways in which indicators of mathematics and science education can be improved. For example, recent reports from the National Academy of Sciences (Raizen and Jones, 1985; Murnane and Raizen, 1988) evaluated the adequacy of existing indicators and made recommendations for improved indicators of learning, student behavior, teaching quality, curriculum, and financial and leadership support. A previous RAND study (Shavelson et al., 1987), supported by the National Science Foundation, criticized current indicators and described a number of different paths—ranging from a "patchwork" synthesis of existing data to the creation of a fully independent, comprehensive data collection system—that could be taken to create a stronger system of indicators.

A variety of criteria can be used to gauge the adequacy of indicator systems. In substantial measure, these earlier reports focused on questions of *what* indicator systems should measure and what types of measures are therefore needed. They also noted concerns, however, about the *quality* of measurement. One apprehension they voiced is that indicators may not accurately reflect core aspects of schooling and therefore may be over- or misinterpreted. A related concern is that some indicators may be unreliable, varying unsystematically from time to time or measure to measure.

This Note addresses the quality of indicators and indicator systems: the related issues of *validity* and *reliability*. It first explains a number of views of validity and reliability, all of which apply to measurement in all fields as well as in the construction and use of educational indicators. It then describes the particular forms those issues take in the case of educational indicators. Some threats to validity, for example, are especially important in the case of indicators, and a variety of factors, such as ways in which indicators are used, can influence validity. It explores the implications for validating educational indicators and for building an adequate indicator system.

2. CONCEPTS OF VALIDITY AND RELIABILITY

Many measures encountered in daily life and in some of the physical sciences are unambiguous and can be taken at face value. Many of the measures used to appraise the condition of education, however—and in the social sciences more generally—are less straightforward. Building an adequate system of educational indicators therefore requires considerable attention to the quality of the measures used. To individuals accustomed to more clear-cut measures, the issues that arise in evaluating educational indicators may seem arcane, but in fact they are fundamental for they determine which inferences an indicator system can adequately support.

This section briefly describes some of the issues that arise in evaluating measures used in educational indicator systems. Readers acquainted with the concepts of the reliability and validity of measurement and with generalizability theory will find this material familiar, although the focus on indicators leads to a particular orientation to it. The following section describes some of the threats to adequate measurement that are particularly likely to arise in the case of educational indicator systems.

Measurement in education is often complex and difficult. Measures in everyday life are usually much simpler. The length of a table is expressed in a clearly understood metric, such as feet or centimeters. Most people have no direct contact with the benchmarks that define everyday measures (such as bars defining standard measures of length or atomic clocks defining units of time), but sufficiently accurate substitutes are readily available. Two different tape measures will provide virtually identical measurements of a table, and when two clocks disagree, it is usually trivial to discover which one is "wrong." Moreover, the accuracy of many common measures can be clearly stated (for example, 4 feet ± 0.1 inch), and the range of error in the measure can often be made small enough to be of little or no practical significance. Finally, alternative metrics and measurement devices often can be substituted for each other with little or no effect on the conclusions one reaches or the decisions one makes.

In education, however, measurement is frequently difficult and ambiguous. One reason is that many of the attributes one wishes to measure are very complex and cannot even be defined precisely. There are many competing definitions, for example, of "proficiency in mathematics" or "knowledge of biology." Thus there is often no unambiguous, single benchmark analogous to bars of standard length or atomic clocks. Moreover, the individual's performance on a given measure can be scaled and reported in many different ways. Unlike different scales of length, these alternative scales are often not simple linear transformations of one another, and the methods chosen can influence the conclusions that are reached. In addition, the accuracy of the measures—for example, the extent to which a student's score on a given test would change if the student took an equivalent test on a different day because of such irrelevant factors as mood, hours of sleep the night before, and so on—is often neither trivial nor obvious. Finally, many measures in education, unlike everyday measures of length or time, can easily be *undermined*; their meaning depends on their context and the way they are used and is not always what it seems at first glance.

In education, therefore, there is often a critical difference between a *measure* and the *attribute* it is intended to gauge, which is often called a "construct." A score on an achievement test in mathematics is not the same thing as "achievement in mathematics;" rather, it is only one of many possible measures that could be used to approximate the unattainable "true" measure of the construct of interest. In the case of achievement and many other aspects of education, these various measures are all simplifications of the complex and often poorly defined construct.

These complexities have profound implications for building a system of educational indicators. Different approaches to measuring core aspects of schooling can suggest different conclusions and point to different policy responses. Building a system of indicators therefore entails many decisions about which measures to use and how to scale and report them. Often, neither the choice of measures nor the decision about scaling and reporting is entirely clear-cut, and there is, accordingly, ample reason for argument about the extent to which the selected measures and metrics support the inferences that are drawn.

GAUGING THE QUALITY OF INDICATORS

Because many of the available measures are only incomplete proxies for an unattainable ideal, the decision among them depends not only on the choice of constructs to be gauged, but also on the quality or adequacy of the measures. Conventionally, issues of quality are subsumed by two terms: *validity* and *reliability*. A measure is valid to the extent that the inferences drawn from it are appropriate and justifiable—for example, the degree to which it is reasonable to infer "mathematics achievement" from performance on a given test. Reliability refers to the "degree to which test scores are free from errors of measurement" (AERA, APA, and NCME, 1985, p. 19), that is, from certain types of inconsistencies from one measure (or time of measurement) to another. Reliability can be seen as a necessary but insufficient condition for validity, or, in a loose sense, as an upper bound on validity. To the extent that a measure is unreliable, it reflects only measurement error rather than any construct of interest.[1]

Issues of validity and reliability appear to be considered only intermittently in constructing educational indicators. For example, the National Assessment of Educational Progress (NAEP) incorporates a wide array of non-cognitive variables—measures of student background, school characteristics, attributes of instruction, and so on—that are commonly used as indicators of the condition of education. Yet very little effort has been made to assess the validity and reliability of the NAEP's non-cognitive items. No one knows, for example, the extent to which students' reports of most family characteristics, reading materials in the home, or homework assignments are meaningful, even though other studies of student self-reports and one limited study of such variables in the NAEP offer ample grounds for concern. The one study of such issues in the NAEP which considered only Asian and Hispanic students, found that students' reports of parental education agreed with parents' reports less than half the time in most of the third-grade groups, never more than 61 percent of the time among seventh-grade groups, and reached a maximum of 65 percent among eleventh-grade students (Baratz-Snowden, Pollack, and Rock, 1988). The NAEP is not exceptional in this respect; indicator systems routinely make use of data that is subject to serious questions about validity and reliability without adequate evaluation.

In part, the lack of attention to the validity and reliability of indicators might stem from the seeming simplicity of many indicators. For example, an indicator such as "the number of mathematics courses taken" seems straightforward. Many people would infer that its meaning is therefore unambiguous and that analysis of validity and reliability—cloaked as it is in arcane terminology—need not be a concern. In fact, however, not even an indicator as apparently simple as this is really straightforward. Even courses with the same name often include different content or are taught differently, and people may be inconsistent in classifying courses for reporting purposes. Care is therefore needed in selecting and validating even apparently simple indicators.

Indeed, the apparent simplicity of many indicators is itself one reason why validity is an issue. Indicators represent a *necessary simplification* of educational conditions and practices, and that very simplification is one reason why the meaning of indicators is not always what it seems to be. In addition, *using a measure as an indicator* can undermine its validity, for the ways in which individuals and educational systems respond to indicators can

[1] In educational measurement, the concepts of validity and reliability have been developed primarily in conjunction with the testing of cognitive attributes, such as achievement and intelligence. Much of this discussion is accordingly couched in terms of tests. The issues involved, however, apply to many types of measures, not just tests.

change their meaning. Subsequent sections of this Note provide a number of concrete examples illustrating that these problems can be severe.

The rest of this section briefly describes several views of reliability and validity and some types of evidence used to gauge them. A theme that runs through discussions of both reliability and validity is the *robustness* or *consistency* of measures or indicators—across alternative measures of the same construct, over time, and so on. The question of the robustness of measures is in some respects the key question in the evaluation of indicator systems.

RELIABILITY

Reliability refers to "the degree to which test scores are free from errors of measurement" (AERA, APA, and NCME, 1985, p. 19). In popular use, "error" often connotes "an avoidable or correctable mistake" (Feldt and Brennan, 1988). In error of measurement, however, "error" refers to *imprecision,* which typically has many sources and cannot be entirely avoided.

Traditional Views of Reliability

As Feldt and Brennan (1988) note, reliability is traditionally measured in one of two ways. One approach focuses on the *standard error of measurement*, which is an estimate of the degree to which a person's score would vary from one instance of measurement to another—for example, if different but equivalent forms of a single test were administered. The second approach focuses on *reliability coefficients,* which indicate the consistency of alternative measures in a specific group. In this view, a score (x) is seen as the sum of two components: a "true" score, τ, and measurement error, ε. Theoretically, the reliability coefficient is defined as the squared correlation between the unmeasured true scores, τ, and scores on a given test, x, or $\rho^2_{X\tau}$. In practice, because true scores are not known, it is typically calculated as $r_{xx'}$, the correlation between two alternative measures, x and x', of the same construct.

An important criticism of the reliability-coefficient approach is that the value of the correlation between alternative measures depends on the characteristics of the particular group in which it is calculated. In contrast, estimates of the standard error of measurement tend to be more stable from one group to another, and that has led many workers in educational measurement to favor that approach (Feldt and Brennan, 1988). This particular criticism, however, is generally not germane to indicators, which are typically based either on the population of interest (e.g., all eleventh-graders in a state) or on a large and representative sample. The reliability-coefficient approach in fact can be particularly useful in evaluating indicators, in that it can be a part of a broader evaluation of the *robustness* of information across measures.

A critical point about the different measures, x and x', used to estimate a reliability coefficient is that they are, to the extent practical, *equivalent*. What is meant by "equivalent" varies, and the estimation of reliability depends on just how similar two measures are. But when two measures differ markedly—for example, students' versus parents' reports of background variables, or standardized multiple-choice tests versus teachers' grades as a measure of achievement—the correlations between them cease to be purely a matter of measurement error and become a matter of validity as well. For this reason, the reliability coefficient is narrower than a test of a measure's overall robustness across diverse measures, which, as we will see, is a key aspect of validity as well.

Generalizability Theory

An alternative approach to assessing reliability that is infrequently employed but that has major implications for evaluating indicators is *generalizability theory*, or G theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Shavelson and Webb, 1981; see also Feldt and Brennan, 1988). G theory departs from traditional views of reliability in that measurement error is not treated as a single, random quantity. Rather, inconsistencies in measurement are seen as being both random and systematic and as stemming from a variety of sources, such as the choice of a measurement instrument, the nature of the sample, and the conditions of test administration or observation. Analysis of variance is used to estimate the components of error variance attributable to each of these factors (called *facets* in G theory). G theory thus extends to the analysis of error variance a common method for assessing the role of substantively important factors.

Underlying G theory is an important change in the way in which ideal measures are conceived. Classical measurement theory stresses "true" scores for which actual measures are fallible proxies. In contrast, G theory focuses not on a single true score, but rather on the wide array of alternative, fallible measures—the *universe* of possible scores—that would be the ideal basis for a given decision. Since the decisionmaker has access to only a small portion of this universe, the question of reliability becomes *accuracy of generalization, or generalizability* from the set of measures chosen to the universe of alternatives (Cronbach et al., 1972; emphasis in the original). G theory entails analyzing facets that determine the error in the measures available to the decision-maker. If, for example, one finds that a large amount of "error variance" actually reflects systematic differences among alternative instruments, then the results from one of these instruments, taken alone, do not generalize well to the universe of alternatives and are a weak basis for decisionmaking.

While the contrast between G theory and the traditional view of measurement error may seem arcane, its implications for constructing indicator systems are extremely important. From the perspective of G theory, it cannot be enough to improve any single measure in an effort to make it a better proxy for the unobtainable "true" measure. Rather, G theory underscores the importance of arraying *a variety* of measures and of analyzing the consistencies and inconsistencies among them. This question of the consistency among alternative measures is also central to recent notions of validity.

VALIDITY

Validity refers to the "appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores [or other measures]" (AERA, APA, & NCME, 1985). Validity is often, but incorrectly, construed as an attribute of a particular measure—that is, the extent to which a measure assesses what it purports to assess. While this can be a convenient shorthand, it can also be fundamentally misleading, because "what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation" (Messick, 1988).

This is a critical point, for it implies that the quality of the measure used or of the data more generally is not sufficient to guarantee validity. These factors can limit validity, but validity also depends on the use to which a measure is put and the specific inferences that are drawn from it. A measure that is highly valid as a basis for one inference in one context may be invalid as a basis for the same inference in another context, or for a different inference in the same context. As illustrated below, a common threat to the validity of indicators arises when indicators that are sufficient for one use are assumed—incorrectly—to be sufficient for another.

Three Types of Evidence of Validity

Many kinds of evidence can be used to test the validity of a measure. The relative importance of different types of evidence depends in part on the use to which a measure will be put. Nonetheless, it is generally preferable to examine several different types of evidence, for a single type can be misleading.

The type of validity evidence that is probably most commonly used in the case of indicators—when any evidence is presented at all—is *content-related* evidence. Content-related evidence "demonstrates the degree to which the sample of items, tasks, or questions on a test [or other measure] are [sic] representative of some defined universe or domain of content" (AERA, APA, & NCME, 1985, p. 10). For example, the cognitive items on the NAEP reflect the judgment of panels of experts about the knowledge and skills that are the most important exemplars of, for instance, eighth-grade mathematics. Content-related evidence is often a tacit concern in constructing indicators even when validity is not explicitly addressed, but such informal consideration of content is not a substitute for a *systematic* analysis of content-related evidence, in which the domain of interest is carefully described and the content of the measure is compared to it. Systematic analysis of content-related evidence of validity is often lacking in the construction of educational indicators.

A common threat to validity that would be addressed by content- related evidence is the failure of an indicator to measure enough of the construct of interest. This problem has been termed "criterion deficiency" or, more accurately, "construct underrepresentation," which Cook and Campbell (1979) characterized as "operations failing to incorporate all the dimensions of the construct" of interest. Student achievement provides obvious examples of this problem, but it arises with many other types of indicators as well. To take one example, consider tests of the mathematical knowledge and skills of eighth-grade students. A test of basic mathematical skills comprising multiple-choice computation items and some very simple word problems might be a valid measure of eighth-grade students' mastery of those basic skills. But what if the focal concern is the students' "mathematical achievement," which is taken to mean, not only those basic skills, but also a variety of other skills and knowledge? For example, the curriculum standards of the National Council of Teachers of Mathematics for grades 5 through 8 call for instruction in algebra, statistics, probability, and geometry, among other areas (National Council of Teachers of Mathematics, 1988). Because the test of basic skills represents only one component of this broader construct of "mathematical achievement," it alone cannot be counted on to support inferences about it. Similarly, preservice coursework is one element of the qualification of mathematics and science teachers, but it is not alone adequate to support inferences about teachers' qualifications, more broadly construed.

A measure can also be undermined for the opposite reason, that is, for incorporating information that is irrelevant to the construct of interest. For example, suppose that the same test of mathematical basic skills was used to support a very different inference: not to gauge students' mastery, but rather to appraise the effectiveness of school programs. That is, it would be used to reach conclusions, not merely about differences among schools in achievement, but rather to support conclusions about the educational *causes* of the differences in mean scores. Much of the variation in scores on achievement tests, however, results from non-educational factors, such as parental education and school-level poverty rates. Unless one can rule out such background factors as causes of differences among schools, it would not be valid to infer differences in school quality from differences in test scores, even if the test is deemed adequate as a measure of achievement.[2]

Criterion-related evidence indicates the degree to which scores on a measure are related to some less arguable outcome criterion. Criterion- related evidence is most germane when there is another measure available that is closer to the construct of interest but less practical to employ. For example, suppose a measure of a mathematics curriculum is considered high-quality but is impractically expensive because it relies on actual classroom observation. A cheaper alternative measure—perhaps, a questionnaire administered to teachers about opportunities to learn—is developed as a surrogate. Criterion-related evidence supporting the validity of inferences based on the cheaper measure might take the form of a high correlation between scores on it and scores on the more costly measure. Criterion-related evidence also is particularly important when a test is used to predict an event in the future. For example, if a test (such as the Scholastic Aptitude Test the SAT) is used to support predictions about future grades in college, correlations between SAT scores and freshman grades constitute criterion-related evidence pertaining to that inference.

Construct-related evidence is the most diffuse category and is defined in diverse ways. The most recent edition of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) stresses the extent to which the measure in question exhibits the empirical relationships that it should, if it is indeed measuring the construct of interest. Such evidence can take a variety of forms. One form is positive relationships between the measure in question and other measures that measure the same or similar constructs—e.g., evidence that one test of mathematical achievement correlates highly with other tests of that domain. This is called *convergent evidence*. Conversely *discriminant evidence* comprises evidence that the measure is not too highly related to measures of different constructs—for example, evidence that a mathematics test correlates substantially less well with reading tests than with other mathematics tests.

[2] In fact, there are several reasons why mean scores may not be a valid basis even for judging the relative achievement level of schools, even if the *causes* of differences are not at issue. Differences among schools in the use of tests—such as the amount of explicit test preparation—may undermine the validity of inferences about relative levels of skills. Differences among schools in background characteristics may also be a threat to validity if the test used is biased against groups disproportionately represented in some schools. In addition, the rank ordering of schools based on mean scores can vary with the scaling of scores unless the distributions of scores have a particular characteristic (i.e., they must be stochastically ordered; see Spencer, 1983). The example above, however, is independent of any of these concerns and is intended to illustrate the *additional* threat to validity that arises when the inference is changed from one about differences in achievement to one about the *causes* of those differences.

The Relevance of Indicator Use to Validity

Finally, consider a case in which the inference to be supported seems, on the surface, to be the same in two contexts, but the use to which the measure is put is quite different. Suppose that a test of mathematics achievement is administered first under "low-stakes" conditions—that is, where scores have few serious consequences for students or teachers—and a rise in scores is observed over several years. To the extent that one accepts the particular test as a sufficient representation of "mathematics achievement," one might accept this trend as valid evidence of improved performance. In contrast, suppose that in another jurisdiction, the stakes attached to the same test were raised greatly—for example, by using scores as a minimum criterion for promotion between grades and as a basis for evaluating teachers. Under such circumstances, a rise in average scores may not be valid evidence of improved mastery of the domain in question. In other words, scores on the test can become inflated as a measure of mastery of the broader domain (e.g., Koretz, 1988; Koretz, Linn, Dunbar, and Shepard, 1990; Linn, Graue, and Sanders, 1990; Shepard, 1988).

Thus, changing the *use* to which a measure is put, or otherwise changing the context in which it is used, can add additional threats to the validity of an inference. In other words, many educational indicators, unlike more clear-cut measures such as measures of length, are *corruptible*—a point that is elaborated in the following section. In this instance, one reason for the corruption is that teaching can become unduly focused on the specific content of the test at the expense of other material, thus undermining the relationship between the specific test and the broader construct which it is intended to measure (e.g., Salmon-Cox, 1984; Shepard, 1990, 1991).

THE CONVERGENCE OF VALIDITY AND RELIABILITY: ROBUSTNESS AS A REQUIREMENT OF VALIDITY

Although the consistency of a measure—for example, the stability of its results over time—is often considered a key component of reliability, it is also a critical aspect of validity. What policymakers want in using indicators is information about broad constructs, for example, the level of students' proficiency in mathematics or the extent to which students study algebra. If two alternative measures of these constructs offer different answers, the validity of inferences about the constructs themselves is called into question. This notion has been articulated in the technical literature for decades but has had disappointingly small influence on the recent educational indicator movement.

The Multitrait-Multimethod Matrix Approach to Validation

The robustness of measures is central to the "multitrait-multimethod matrix" approach to validation (Campbell and Fiske, 1959), which combines convergent and discriminant evidence. This method calls for "triangulation"—that is, obtaining information about a construct of interest from two or more sources, such as student and parent questionnaires, or achievement tests and teachers' grading. Ideally:

- Relationships among variables in data from each source should show relationships that support the validity of the measures;
- These relationships should be similar from one method or data source to another (e.g., the same in student questionnaires and parent questionnaires); and
- Estimates for each variable should be consistent across sources. (For example, estimates of family income from parents and students should be highly correlated.)

Thus, one key aspect of this approach is ascertaining the *robustness* of information across methods and sources. This is similar to the generalizability-theory approach to reliability. In both cases, a given measure is viewed as only one of a set of defensible alternatives, and the question is the extent to which the results on the first measure *generalize* to the alternative measures, or to scores on the same measure administered under different circumstances. While the specific multitrait- multimethod matrix approach is open to a variety of criticisms (Messick, 1988), this emphasis on the robustness of information across methods or sources of data is particularly useful in evaluating indicators.

The Sampling Model of Validity

The critical importance of the robustness of information has been extended into a formal theory of validity, the sampling model of validity (Kane, 1982). Kane noted that generalization over alternative measures is essential to valid measurement. He noted, for example:

> A statement that the length of a metal bar is 1.5 meters treats length as a property of the bar and implies that length does not depend, for example, on the location, orientation, or temperature of the bar, or on the identity of the observer (Kane, 1982, p. 127).

More generally, he argued, the validity of a measurement procedure depends on the extent to which "it provides accurate estimates of the expected value over the universe of observations defining the attribute." Validation, in this view, requires that one assess the degree to which measurements are *invariant* across reasonable alternative measures.[3]

In this view, a key difference between validity and reliability is the range of alternative measures under consideration. When one asks how much a student's score is likely to fluctuate on a given standardized test—an archetypal question of reliability—one is asking about the generalizability of results across a narrow range of "allowable observations." In contrast, in assessing validity of inferences, one must consider generalizability to the broad universe of possible measures that "define" (or are implied by) the construct one is theoretically measuring. In this view, in estimating the reliability of a standardized test of mathematics achievement, it is reasonable to restrict the range of measures to some that are very similar to the test in question. In contrast, in evaluating the validity of inferences about "mathematics achievement" based on that test, it is necessary to consider the extent to which results on the test are similar to the results that would be obtained with substantially different measures of the construct of "mathematics achievement."

One strong argument in favor of the sampling view of validity is that it formally acknowledges the importance of multiple measures of a given construct. Any single measure is likely to be flawed and therefore misleading to some degree. As Messick (1988, p. 34) notes,

> Tests are not only imprecise or fallible by virtue of random errors of measurement but [are] also inevitably imperfect as exemplars of the construct they are purported to assess. Tests are imperfect measures of constructs because they either leave out something that should be included . . . or else include something that should be left out, or both.

These problems are characteristic, not just of tests, but of measurement in the social sciences more generally, and methodologists have argued for decades that the appropriate response is to rely on multiple measures. For example, Webb, Campbell, Schwartz and Sechrest (1966, p. 3) argued that:

[3] Generalizability and validity are not necessarily the same. Messick (1988) noted that several types of generalizability—across populations, settings, times, and tasks—have been labelled validity, often inaccurately. He noted that validity need not imply generalizability over all these dimensions; the appropriate degree of generalizability depends on the construct and inference. Generalizability across tasks representative of the domain in question, however, is similar to the question of the sample tasks' representativeness of the domain, which is a matter of content-related evidence of validity. Generalizability across alternative measures is merely an extension of this traditional component of validation.

> The operational implication of the inevitable theoretical complexity of every measure . . . calls for . . . multiple measures which are hypothesized to share in the theoretically relevant components but have different patterns of irrelevant components.

In other words, one should "triangulate," using methods that are unlikely to share the same inadequacies or distortions. The sampling model of validity is an elaboration of view, because it stresses the comparison of results among different categories of measures.

Finally, it is essential for present purposes to consider what Kane called "the tradeoff between validity and import." Kane maintained that "if import is ignored, it is easy to generate measurements with a high degree of validity by defining the universe of generalization narrowly enough so that the inferences to the universe scores involve generalization over few facets" (Kane, 1982, pp. 151-152). That is, one can narrow the universe to rule out alternative measures that are likely to produce discordant results, but the result may be that valid inferences are restricted to relatively unimportant, narrow constructs. In the previous example, one could ignore generalization from multiple-choice, standardized tests to other measures of mathematics achievement, but the price is that valid inferences would be restricted to aspects of mathematics achievement measured by multiple-choice tests, not to mathematics achievement more generally. As is shown in the following section, the focus on broad constructs of clear import is one of the primary reasons why the validity of many indicators is problematic.

3. THREATS TO THE VALIDITY AND RELIABILITY OF INDICATORS

The concerns of reliability and validity noted above apply to measurement in general, but they take particular forms in the case of educational indicator systems. This section discusses some characteristics of educational indicators and examines some of the particular forms that questions of validity and reliability take as a result.

The particular issues of validity and reliability faced by indicators reflect both attributes of the indicators themselves and the characteristic ways in which they are used. One important aspect of their use is the types of inferences that they are employed to support; these often differ from the inferences that the same measures are used to support in other contexts. Equally important are the ways in which indicators are used to monitor the performance of educational systems and, in the extreme case, to hold educators or students accountable for their performance. When indicators are used in this fashion, the responses of individuals in the educational system can fundamentally alter the meaning of the indicators and may undermine the validity of the inferences they are intended to support.

CHARACTERISTICS OF INDICATORS

The adequacy of an indicator system is affected by the characteristics of the individual indicators comprising the system, as well as by attributes of the indicator system as a whole. Characteristics of individual indicators—and of the uses to which they are put—are considered here, and aspects of indicator systems are discussed subsequently.

Indicators Are Used to Support Broad Inferences

One of the most important characteristics of educational indicators is that they are typically used to support very broad inferences. In deciding how educational policy should be changed, policymakers need a portrayal of the big picture: the condition of educational achievement, the adequacy of course offerings, the dropout rate, and so on. In terms of the tradeoff Kane (1982) noted between import and validity, indicator systems lean heavily in the direction of importance to the policy debate, and that entails breadth of inference. The broader an inference becomes, however, the more problematic validity is likely to be, because it becomes increasingly likely that any one measure will be too narrow to support it adequately. In other terms, as Kane (1982) noted, the broader the inference, the more numerous and diverse will be the measures across which results should generalize.

Indicators of student achievement provide a clear example of this, although examples could be chosen from many other aspects of education as well. The intense debate of the past decade about the level of achievement in the United States has been largely couched in very general terms: Has achievement been deteriorating or improving? How much lower is the mathematics achievement of American students than that of Japanese students? Which states demonstrate the highest levels of achievement? This tendency has been apparent both in public debate and in policy-oriented analytical work (see, for example, Harnischfeger and Wiley, 1975; Koretz, 1986, 1987; National Commission on Excellence in Education, 1983). In some instances, this tendency has been carried to the extreme of using available measures of "achievement" with little or no consideration of their specific content. For example, throughout the 1980s, the U.S. Department of Education periodically released "wall charts" of state comparisons in which the broad construct of "performance outcomes" was represented by two numbers: the average college-admissions test score (total battery scores on either the SAT or the American College Testing Program [ACT] tests, depending on the state) and the percent of high school graduates taking that test (Office of Planning, Budget, and Evaluation, U.S. Department of Education, 1985). No consideration was given to the specific attributes measured by those tests—for example, to the fact that the ACT, unlike the SAT, includes subtests covering natural and social sciences, or to the fact that the SAT is not closely linked to secondary school curricula and is not validated as a measure of achievement.

This emphasis on general constructs is reflected in the current form of the NAEP. A particularly striking example is the reporting of "science" as a single domain of achievement (Mullis and Jenkins, 1988). In the 1986 administration, the NAEP science test comprised five content areas (each represented by a subscale): life science, chemistry, physics, earth and space science, and the nature of science. These content areas were scaled separately, and some analyses were reported for individual subscales. The primary emphasis in reporting the NAEP, however, was on an overall scale of "science proficiency," which is a weighted average of scores on the five subscales (see Mullis and Jenkins, 1988, pp. 138-141), where the weights accorded each subscale reflect expert judgment about each area's importance. Very few actual test items were reported, and those that were presented were selected, not to illustrate the specific content areas, but rather to exemplify various points on the overall scale of science proficiency. The concurrent NAEP mathematics assessment (Dossey, Mullis, Lindquist, and Chambers, 1988) was handled in the same fashion.

These characteristics of NAEP reporting illustrate the emphasis on very broad questions that often characterizes the use of achievement test data as indicators. In contrast, many other uses of test data would entail narrower inferences and, in many cases, a much tighter match between the instrument used and the construct represented. For example, a district superintendent wanting to appraise the success of a new curriculum in chemistry would need a measure focused specifically on the content implicit in the new curriculum.

Indicators Are Often Built from Simple and Inexpensive Measures

For several reasons, measures used to construct indicators are often relatively simple. In part, this is deliberate. An indicator system must simplify the condition of education in order to inform rather than overwhelm debate about policy. (The trade-off between simplification and comprehensiveness is illustrated concretely in a companion Note on indicators of mathematics and science achievement; see Koretz, 1991a.) In addition, policy often addresses simple aspects of the educational system (e.g., how many courses students must take in mathematics), and indicators will often be designed to parallel them. The desire for *comparability* of information across jurisdictions may also increase the tendency toward simplification, for disparate districts and schools may more easily be compared in terms of simple variables that are easily pulled out of context, such as the proportion of students taking algebra.

In part, however, the simplicity of indicators is a matter of cost. Indicator systems are intended to obtain a substantial amount of information, typically on a wide variety of topics, from many sources on a routine basis. Routine use of expensive measures—such as observational measures of instructional content and approach, longitudinal studies of students, and representative questionnaire studies of parents—would make most educational indicator systems prohibitively expensive. To constrain costs, indicator systems therefore typically rely on relatively inexpensive measures that are not too burdensome either to those administering them or to those responding. This can threaten both the validity and the reliability of the measures; that is, it can introduce both biases and measurement error into the results.

Inaccuracy can be a consequence of the reliance on simple and inexpensive measures. For example, some indicator systems have relied on student reports for information on background factors. As noted above, very limited analyses of student reports in the NAEP revealed serious deficiencies in the responses. A similar but larger-scale analysis of the quality of responses in the High School and Beyond study provided similarly sobering results. Correlation coefficients between sophomores' and parents' reports of background variables ranged from very low to quite high—for example: .21 for the presence of a specific place to study in the home; .35 for the presence of an encyclopedia in the home (an item used in the NAEP as well); .44 for mother's occupation; .50 for family income; .56 for whether the family owns or rents its residence; .81 for mother's education; and .87 for father's education (Fetters, Stowe, and Owings, 1984).

Simple measures may also be insufficiently detailed or comprehensive, even when they are not clearly inaccurate. Curriculum indicators provide a clear example. Many observers have noted that there can be important disparities between *surface components* of coursework or curriculum and the content of actual instruction. Murnane and Raizen (1988) distinguished between the "intended" curriculum, which includes the likes of curriculum guidelines, texts, and tests, and the "implemented" curriculum, which is the actual content and style of instruction. For present purposes, however, it would be more useful to see this distinction as a continuum rather than as a dichotomy. At one end of the continuum, some components of curriculum contain little detail and can be considered to be only distantly related to instruction. The most extreme example is the requirement that students merely take a specified number of courses in a given subject area. At the other end of the continuum is the actual, often unmeasured content of instruction. At intermediate positions on this continuum are a variety of other components of curriculum, such as fine-grained instructional objectives, textbook requirements, and curriculum-linked competency tests. Curriculum indicators have typically focused on the distant, general end of this continuum, such as measures of course enrollments in broad subject areas or by course title. On a national basis, the information available about more detailed aspects of curriculum, such as details of course content, instructional style, and so on, is sparse, limiting what one can say about group differences in exposure or about reform-related changes in curriculum (Stecher, 1991).

An additional consequence of using simple measures is that the meaning of indicators can vary substantially from one setting to another. Thus, for example, many jurisdictions maintain data on the proportion of students who take Algebra I, but courses with that title can vary markedly in content even with one jurisdiction (e.g., Massachusetts Department of Education, 1986). In some instances, it may be possible to reduce sharply the variation in the meaning of indicators; an instance is recent efforts to standardize the collection and interpretation of data pertaining to dropouts. In many instances, however, inability of simple measures to capture important variations among settings is likely to be an inherent weakness of indicator systems. As explained in the following section, one approach to this problem would be to supplement the collection of indicator data with periodic, more intensive data collection.

Indicators Are Often Constructed from Corruptible Measures

In daily life, we are accustomed to using measures that are not affected by the attention we pay them. For example, readings from fever thermometers are usually valid and reliable regardless of the use to which the resulting information will be put. Telling a child that the presence or absence of fever will determine whether she can stay home from school is unlikely to affect the reading, provided the thermometer stays where it belongs for the appropriate time. There is little reason to doubt that a change in the reading reflects a true change in the underlying condition in question—that is, the fever. In such circumstances, few people would bother trying to manipulate the measure directly rather than manipulate the underlying condition, and fewer yet would succeed.

Unfortunately, such is not the case with many educational indicators. We know that many can be—and are—directly manipulated. This manipulation changes the relationship between the measures and the constructs they are intended to represent. That is, it threatens the validity of the inferences one wishes to draw from the measures. Moreover, the more important an indicator is, the greater the incentive to manipulate it. The current period of reform, with its strong emphasis on indicators as ways of judging performance and holding people accountable, provides many incentives to manipulate indicators.

Currently, the most widely recognized examples of this problem pertain to the use of achievement tests as indicators. Some observers in the educational measurement field have warned that increased attention to test scores as a criterion for judging schools could lead to an inflation of test scores that would undermine inferences about educational performance (e.g., Koretz, 1987, 1988; Linn, 1983, 1987; Madaus, 1988; Shepard, 1988). This inflation of scores can stem from many factors—such as a shifting of instructional time from untested to tested parts of the curriculum, other forms of inappropriate teaching to the test, outright cheating, and the use of dated or otherwise inadequate norms—all of which undermine the generalizability of scores on the particular test to other measures of the domain of achievement in question. Until recently, these warnings seemed to be largely unheeded by educational policymakers and the press, but over the past several years, there has been a dramatic increase in the attention paid to this problem. An event that did much to bring about this change was the publication of *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average* (Cannell, 1987). This booklet, written by a West Virginia physician, maintained that most states and an implausibly large number of school districts are claiming to be above average in terms of standardized test scores. The booklet ignited a firestorm of controversy; despite arguments about the report's flaws, it did much to make policymakers and the public aware of the potential of high-stakes test scores to provide a misleading view of achievement. Cannell's study has since been more carefully replicated (Linn, Graue, and Sanders, 1991), and other research has confirmed the inflation of test scores (Koretz, Linn, Dunbar, and Shepard, 1991).

The area of curriculum also provides clear, if less well known, examples of the corruptibility of indicators. For example, many recent reforms have increased the minimum requirements for academic coursework—that is, core courses in the traditional academic disciplines, such as mathematics, science, and English (1985). In many jurisdictions, simple indicators that can serve to monitor compliance with these new requirements are already in place. In response to the requirements, some jurisdictions have found ways to change the ways that courses are categorized with respect to these indicators. For example, some have found ways to count vocational courses as academic. One way has been to relabel courses; efforts have been made to better document the academic content of vocational courses to justify their being counted as academic. Another approach has been to repackage course content—increasing the academic content of vocational courses enough to justify their being counted as academic (National Assessment of Vocational Education, 1988).

In each of these instances, participants in the educational system *change their behavior in response to an indicator in a way that changes that indicator's meaning.* Just as teaching to the test can change the meaning of test scores—that is, change the inferences for which test scores offer valid support—the re-labelling of courses alters the meaning of course enrollment statistics. In both cases, unless one has information on these changes in behavior, the inferences one is likely to draw from the indicators can be fundamentally misleading. One may read in the indicators evidence that achievement is improving or that coursework is becoming more rigorous, when the truth is not as encouraging.

What makes the corruptibility of many indicators particularly vexing is that the odds of corruption increase with the salience and importance of the indicator. In particular, using a measure for *accountability* will often undermine its value as an *indicator*. If no one is attending to a particular indicator, participants in the system have no reason to modify their behavior to maximize their standing on that variable. But when an indicator becomes more important as an index of the condition or performance of the educational system, the incentives to act in ways that threaten the meaning of the indicator will increase.

The Meaning of Indicators Can Drift Over Time

Corruption in the face of reform is not the only reason why the meaning of indicators may be unstable over time. Gradual changes in practice can occur without the stimulus of a specific reform and can change the meaning of an indicator. A well-known instance is "grade inflation." A variety of data sources confirm the popular view that high school grading standards became easier during part of the 1960s and 1970s; the proportion of students receiving high grades increased even as average test scores fell (Advisory Panel on the Scholastic Aptitude Test Score Decline, 1977; Koretz, 1987; National Center for Education Statistics, 1982). A less well known example is unintended changes in the meaning of scales used to report achievement-test results. For example, despite careful equating of the test from one year to the next, the scale used to report the SAT gradually became easier during part of the 1960s and 1970s, partially attenuating the well-publicized drop in SAT scores during that period (Advisory Panel on the Scholastic Aptitude Test Score Decline, 1977).

Indicators Are Aggregate Measures

Indicator systems typically make use of aggregate measures, such as average test scores (for a school, district, or state), the percentage of students taking a given course, and the percentage of teachers reaching some level of qualification. This has substantial implications for evaluating their reliability and validity.

On the positive side, aggregate measures tend to be more reliable, in the narrow sense of the word. That is, aggregate scores on a given indicator will generally be more consistent than scores of individuals over time and across measures, such as similar achievement tests. Indicator systems are sometimes deliberately constructed to take advantage of the greater reliability of aggregate measures. For example, if the purpose of a test is to provide only reliable estimates of aggregate performance, individual students can be given tests that are too short to provide reliable estimates of their own performance. This can be advantageous for several reasons: it lessens respondent burden, lowers costs, and permits the assessment to cover a broader range of material *in the aggregate* than it would otherwise. This approacn is used in the NAEP and in a variety of state assessment programs.

Although the use of aggregate measures reduces some sources of measurement error, it does not affect many of the important sources of potential inconsistency across measures. Moreover, some observers appear willing to draw broader inferences from aggregate measures than for individual measures, a tendency that intensifies threats to the robustness—and therefore the validity—of the conclusions reached. For example, numerous states and localities use the results of achievement testing in a few grades and subject areas as a basis for broad inferences about aggregate levels of achievement of schools or districts, implicitly assuming that testing in those few instances provides a reliable and valid guide to overall levels of achievement. In some instances, however, aggregate levels of achievement have been shown to vary substantially from grade to grade, subject to subject, and year to year, particularly when the effects of other factors (such as demographic factors or prior achievement levels) are taken into account (e.g., Frechtling, 1982; Guskey and Kifer, 1990; Helmstater and Walton, 1985; Kippel, 1981; Mandeville and Anderson, 1987; Matthews, Soder, Ramey, and Sanders, 1981; Rowan and Denk, 1983). In other words, schools, and to some degree districts, are ranked differently—sometimes dramatically so—depending on the grades and subjects tested and the year in which tests are administered. Moreover, rankings are sometimes dramatically different depending on the method used to construct rankings from aggregate scores and background data (Frechtling, 1982).

In addition, aggregation can cause potentially misleading consistency. For example, tc the extent that the ranking of schools in terms of mean achievement is consistent from one test to another, that consistency is often attributable in large part to factors not directly related to any specific domain of instruction or achievement, such as the ethnic or social-class composition of the schools. That is the reason that school rankings on tests become much less consistent when the effects of such background factors are controlled. Because the apparent consistency of rankings stems in substantial part from such factors, however, rankings of schools are sometimes quite similar from one test to another even when the tests used reflect different domains. For example, in the 1986 NAEP, school means on the mathematics composite scored about .90 with school means on the science composite.[1] Unless the effects of background factors are taken into account, observers may falsely infer from similarity in aggregate scores on two tests that the tests measure similar constructs, or may misconstrue the similarity as evidence of validity for either one.

CHARACTERISTICS OF INDICATOR SYSTEMS

Indicators are used not only individually, but also as parts of a *system* of measures intended to monitor core aspects of schooling. The characteristics of the indicator system as a whole must be considered as well in evaluating the validity and reliability of the information obtained from it.

Indicator Systems May Be Insufficiently Broad

A variety of institutional and contextual constraints can cause a policy to have broader or more diverse effects than intended. This expansion of effects can pose a serious threat to

[1] This correlation is attenuated by the small number of tested students in some schools. Schools with very few students tested in mathematics were eliminated as described in the accompanying report on achievement indicators (Koretz, 1991a, Appendix C), but because of the design of the NAEP, correlations across subject areas involved very few students tested in science in many schools.

the validity of indicators that are focused on that policy unless the indicator system is broad enough to capture the range of effects.

An educational system has only a certain amount of "give"—after a certain point, pressure on one aspect of the system will likely cause some other aspect of the system to change to relieve the pressure, just as pressure on one part of a balloon will cause a bulge elsewhere. Where the system bulges depends on the specific constraints in that particular educational system. The impact of these constraints can be quite general, and they are a primary reason why indicator systems can be insufficiently comprehensive to provide a valid basis for inference.

In some cases, these constraints can lead to effects that go beyond the specific intended focus of the policy in question but still remain within the part of the educational system that is the policy's target. For example, new requirements for coursework in mathematics and science can have diverse effects on the mathematics and science curriculum beyond the intended increase in enrollments in certain courses. For example, an increase in the number of science courses required for graduation in one midwestern state may have led to a reduction in the number of advanced science courses offered in some districts because of a shortage of qualified teachers. The teachers who had formerly taught them had to reallocate their time to teach the additional introductory courses that were the predictable consequence of the new requirement (Schrock, 1988). Other possible unintended consequences of increased coursework requirements might be an increase in teaching by out-of-field teachers, a watering-down of course content, and an increase in class size.

In such instances, the adequacy of an indicator system would hinge on whether the system measures the unintended as well as the intended effects *within* the domain in question—in this instance, the mathematics and science curriculum. This is analogous to a traditional question of validity, the adequacy with which a measure samples from the domain of interest.

In other instances, however, the bulges in the educational system will occur *outside* the domain of interest. For example, students may respond to new mathematics requirements by taking fewer history courses. This can make the validation of indicators more problematic than the validation of many other measures, for the question is not merely whether the indicator system samples adequately from the target domain, but also *whether the indicator system samples adequately from the range of domains that are likely to be affected by the same policies or linked in other ways.*

Achievement testing provides a particularly clear example of the more problematic case in which the meaning of an indicator hinges on changes *outside* the domain of interest.

At the beginning of this decade, for example, Pittsburgh began an intensive program of measurement-driven instruction (Monitoring Achievement in Pittsburgh, or MAP) that was coupled in some cases with the adoption of new texts. MAP was phased in one subject at a time. One of the contextual constraints that influenced the effects of this program was the nearly fixed amount of instructional time available to teachers. Thus, one study found that in the first years of the program, when the testing program included only reading and mathematics, some teachers in self-contained classrooms responded to the MAP program by taking instructional time away from other, untested subject areas, such as science and social studies (Salmon-Cox, 1984). At that time, then, a comprehensive assessment of the effects of MAP on achievement would have required tests in subject areas, such as science, that were not then included in MAP. Similarly, in one mid-Atlantic state, some schools responded to the imposition of a minimum competency test in citizenship by reducing the amount of time some students spent in science classes in order to free up time for additional citizenship instruction (Wilson and Corbett, 1991). To put the first of these examples in the terms used here, it could be quite misleading to construe the early increases in MAP mathematics scores as simply an increase in mathematics achievement and to infer from that an improvement in mathematics instruction, because the rise in scores represented in part a *transfer* of achievement from science to mathematics.

One can find even more extreme examples, in which the effects of a policy are even farther removed from the target concern. For example, indicator-based accountability systems might alter the tracking of students so that some of those whose performance would harm a system's standing on the indicators are removed from consideration. For example, Gottfredson (1988) examined aggregate trends in one Southeastern county after the implementation of a test-based reform measure. He found that scores on both a criterion-referenced test and a norm-referenced test increased markedly, but those increases were accompanied by a sizable rise in retention in grade.

Many characteristics of schools and school systems can cause the effects of policies to spread in ways that can threaten the validity of indicators. Two factors already noted are the supply of qualified teachers and the total amount of available instructional time. The internal organization of schools can also be germane. For example, the response of teachers to the Pittsburgh MAP system depended on whether they were in self-contained classrooms. Teachers in self-contained classrooms tended to take time away from subject areas that were not included in MAP, such as science and social studies. In contrast, teachers in departmentalized structures generally taught only one subject, so that option was not available to them; instead, they took time from material within their subject areas that was not covered by the exams (Salmon-Cox, 1984). Thus, in the departmentalized structure, separate examinations that more fully covered the district's curricular goals in the tested subjects might have served as a valid indicator of the program's effects on achievement. In contrast, in self-contained classrooms, an indicator system would have had to cover the other subject areas from which time was taken in order to provide a comparable level of validity.

Indicator Systems Often Include Few Alternative Measures

Indicator systems often include only a few—or even only a single—measure of a given construct. Some jurisdictions have been working to lessen this problem, but the cost and other burdens of indicator systems would seem to guarantee that it will not be solved fully in the near term.

Achievement again provides a clear example. Many jurisdictions are expanding the set of "performance outcome" indicators they collect in order to provide a more comprehensive view of outcomes. For example, many jurisdictions consider, not only achievement tests, but also progress through the grades and high school graduation rates. In addition, some jurisdictions use a number of tests as indicators of achievement. Many relevant inferences, however, are not about "performance," but rather about narrower constructs, such as "mathematics achievement of junior high school students." At this level of inference, indicator systems tend to include very few if any alternative measures. On a national level, for example, there is only one ongoing source of nationally representative data about junior-high mathematics achievement: the NAEP. The situation is a bit better in some states that administer both "functional" (or criterion-referenced) and norm-referenced, standardized achievement tests in mathematics. Still, in most jurisdictions, the opportunity to validate one measure of a construct by comparison to others is very limited, if present at all.

One reason why the scarceness of alternative measures threatens indicator systems was noted in the preceding section: any given test is only one of many possible incomplete measures of the construct of interest. The concern of policy is rarely scores on that particular measure, but rather performance on the construct it is supposed to represent. Alternative measures will sometimes provide similar information—achievement tests are often validated in part by showing that they yield information similar to that provided by alternatives—but they do not always. Different achievement tests, for example, can rank schools or districts differently (e.g., Koretz, 1986). Some of these discrepancies may be explicable; significant discrepancies can be caused by differences in the degree of overlap between tests and curricula (e.g., Bianchini, 1978; Cooley and Leinhardt, 1980; Miller and Linn, 1988). But many of the discrepancies among alternative achievement measures remain unexplained, which underscores the risk of relying on any one of them.

A second reason that reliance on a single measure is risky is that occasionally the information from a single measure is simply anomalous. Perhaps the best know recent example of this is the "reading anomaly" in the 1986 NAEP. As the primary report of that assessment noted:

> The results of the 1986 reading assessment seemed to be out of line with previous NAEP reading assessment results. In particular, they indicated precipitous declines in average reading proficiency at ages 9 and 17. The nature of these drops across only a *two*-year period, taken in the context of only modest changes in reading proficiency across a succession of four-year periods since 1971, was simply not believable (Applebee, Langer, and Mullis, 1988, pp. 56-57; emphasis in the original).

A wide-ranging analysis of this anomaly by the Educational Testing Service, the prime contractor for the NAEP, was largely inconclusive, although it ruled out certain explanations (Beaton, Ferris, Johnson, Johnson, Mislevy, and Zwick, 1988). With one dissension, an independent panel set up to investigate the anomaly concluded that it probably reflected procedural differences rather than a real decline in reading ability or changes in population characteristics (Haertel et al., 1988).

Finally, the corruptibility of many indicators further heightens the risk of relying on indicators in the absence of alternative measures. Where there is a risk that the use of an indicator has corrupted it, validating inferences based on the indicator requires, not just alternative measures, but alternative measures that are relatively unaffected by the factors that corrupted the original measure. In many cases, that means that the alternative should be a measure to which relatively little attention has been paid.

Indicator Systems Often Lack Measures Relevant to Alternative Explanations

As noted in the previous section, validating an inference often requires, not only evidence supporting the inference, but also evidence that undermines or rules out alternative explanations of the observed relationship. When indicators are used purely for descriptive purposes—which many would argue is their appropriate use—this is usually not a major concern. Often, however, audiences use indicator data to support inferences that go beyond simple description, in particular, to draw inferences about school quality or program effectiveness from differences in test scores. Indeed, the desire for such information is a major motivation for many recent expansions of indicator systems. In such cases, the availability of measures pertaining to alternative explanations can be critical.

The NAEP provides a good example, because its recent expansion to provide comparisons among states was motivated in substantial part by a desire for information about the relative quality of school systems. As currently conducted, however, the NAEP cannot support most inferences about the relative effectiveness of educational systems (Koretz, 1991b), in large part because of insufficient information about other potential causes of differences in scores. For example, NAEP is cross-sectional and thus lacks information about earlier performance levels (and, therefore, about improvement); it lacks adequate information about some important background variables; and it has no information whatever about some others. In addition, it includes insufficiently detailed information about many educational variables, including information on students' educational histories and important details about their current programs. The inability of NAEP to support causal inferences about educational effectiveness is illustrative, because many of its characteristics are shared by other indicator systems.

It is ironic that the current widespread interest in expanding indicator systems rests in part on the desire to use them to evaluate programs. The contemporary interest in educational indicators is a resurgence of a longer term interest in social indicators, but it appears that some of the lessons of the earlier social indicator movement have been forgotten. Shavelson (1987), for example, argues that by the mid-1970s, the social indicator movement had largely reached a consensus that indicator systems (in general, not just in education) cannot substitute for careful program evaluation because the systems lack sufficient detail and control.

4. APPROACHES FOR EVALUATING AND VALIDATING INDICATORS

The potentially serious threats to the validity and reliability of education indicators are too numerous and diverse to be addressed by a single, cookbook method. Nonetheless, the preceding discussion suggests a number of general approaches that would be beneficial in designing and evaluating them.

ASCERTAINING LIKELY THREATS TO VALIDITY AND RELIABILITY

A first step for evaluating or improving an indicator system is to determine what the likely threats to validity and reliability are. Among the important questions to address are the following:

Are indicators based on data of questionable accuracy or reliability? Although student self-reports were used above as an example of data that are questionable in this regard, many other types of data may be suspect as well. For example, dropout statistics are a notorious problem—inconsistent in definition from one jurisdiction to another and often based on badly flawed data. Data quality is especially likely to be a problem if indicator systems are built with "data of convenience," such as extant administrative-record data or easily obtained but questionable reports from staff and students. Determining which data are liable to be flawed is essential if some of the additional steps suggested below—such as collecting supplementary data for evaluative purposes—are to be targeted efficiently.

Are the indicators subject to corruption? Whether—or to what degree—indicators are vulnerable to corruption hinges both on the nature of the measures and the uses to which they are put. The more the indicators are used to hold people accountable, the more likely they are to be corrupted. Test scores appear relatively easy to corrupt, particularly if the tests are not secure and are census-based. Some participation indicators, such as retention in grade and dropout rates, can be corrupted, although the imposition of common definitions, data collection standards, and reporting standards can greatly reduce the risk. Course-taking measures are also vulnerable when standards for content and approach below the level of course title are weak.

How broad are the inferences the indicators will be used to support? The broader the inferences, the more likely it is that a single measure, or a few very similar measures, will be insufficient. That is, the breadth of inference is a key to the problem of *robustness*: broader inferences make it more likely that the results of alternative measures would differ substantially.

How broad are the likely effects of policies that the indicator system is intended to monitor? For example, will mandated increases in basic coursework in one subject lead to a watering down of course content? Or is it likely to lead to decreases in advanced coursework in that subject or decreases in basic coursework in another subject? The answers to these questions depend in part on the characteristics of the particular settings in which the policies are implemented, such as school size, current enrollment patterns, availability of staff, and the characteristics of the student population.

BUILDING INDICATOR SYSTEMS

It is not possible to build indicator systems that are immune to the threats described above, but it is feasible to build them in ways that both lessen the severity of the threats and make it practical to evaluate their severity. Among the most important steps are the following:

Use multiple measures of important constructs, particularly when there are reasons to expect individual measures to be incomplete or corruptible. When feasible, the multiple measures should be vulnerable to *different* weaknesses, so that they provide complementary information and provide a check against misinterpretations that might stem from using any one of them alone. For example, if the incompleteness ("construct underrepresentation") of a measure is a potential problem, as is often the case with achievement tests, it may prove important to find alternative measures that tap different portions of the broad domain of interest. If source of data is a potential problem, for instance, using principals' estimates of occupational profiles of school communities—alternative measures should rely on a different source. Where routine use of multiple measures for all units of analysis is impractically expensive, some of the measures can often be administered on a sample basis or at less frequent intervals.

Include an appropriately wide range of domains. The indicator system as a whole should provide, not only sufficiently comprehensive coverage of the domains of primary interest, but also information about other aspects of education that are likely to be affected by spillover from policies in those primary domains. For example, one study cited earlier that the use of achievement tests as a criterion for promotion or graduation may have substantial effects on retention on grade. In such an instance, trends in test scores cannot be interpreted correctly without reference to trends on retention.

Link the indicator system to other data. The incompleteness of any single indicator system can be offset to some degree if it is designed to mesh with other types of data collection efforts. For example, at the federal level, periodic, large-scale longitudinal surveys such as the High School and Beyond survey and the National Educational Longitudinal Study include measures of many of the same constructs as are tapped by the more frequent NAEP. Tailoring the measures used in each to take advantage of this can widen the range of available information and increase the data available to address the quality of the indicators used in each.

EVALUATING INDICATORS

The issues that arise in evaluating indicators are fundamentally the same as those that arise in appraising other types of data, but the characteristics and uses of indicators suggest that particular approaches are worth emphasizing. Many of the approaches that should be followed require only the data from the indicator system itself, but others require external data cf other types.

Apply traditional criteria of validity and reliability. This obvious step bears reiteration because it is so often disregarded in building indicator systems, perhaps because of the apparent simplicity of the indicators themselves. If there is reason to suspect significant measurement error—as often there is—traditional measures should be used to appraise their reliability.[1] The validity of measures should be explored by a variety of approaches, such as examining relationships among measures for both convergent and discriminant evidence of construct validity and comparing data across sources.

Assess the robustness of information. An essential step in evaluating the validity and reliability of indicators is assessing the robustness of the results they yield, over time, measures, and sources of data.

Assessing the robustness of results addresses both reliability and validity. The basic approach can be the same in both cases; what differs is the range of alternatives over which robustness is tested. If the measures compared are nearly equivalent—for example, two alternative forms of a given test separated by only a brief period—then the consistency between them could be seen as a test of reliability. But if the comparison is between considerably different measures that nonetheless are implied by the single construct in question—say, between a multiple-choice test and a free-response test of the same domain—the degree of consistency is a test of the validity of the measures, as conceptualized in the sampling model of validity. In between these two extremes is a continuum of comparisons among measures varying in their similarity. Many of the comparisons among alternative

[1] As noted earlier, because indicators are generally based either on the population or on representative samples drawn from it, the reliability-coefficient approach to assessing reliability (which is one test of robustness) is less subject to criticism than it is in many other contexts (in which the value of the coefficient can vary with the characteristics of the group studied).

measures in a patchwork indicator system, such as comparisons among the results of similar achievement tests, lie between these two poles.

Accordingly, when the results of indicators are found *not* to be robust, it may not always be particularly useful to try to pigeonhole that finding as pertaining either to reliability or to validity. Rather, it may be more useful to ask: *What specific inferences does this particular inconsistency undermine?* In what ways are conclusions likely to be misleading? What inferences are warranted nonetheless?

It can be revealing to assess the robustness of findings even without prior consideration of any specific threats to validity. Regardless of the types of threats noted above, any number of characteristics of data sources can cause inconsistencies. Examples include differences in the wording of questions, in the content or format of test items, in the scaling of results, in the conditions under which tests are administered, in the patterns of non-response by sampled individuals, and so on. It is also important, however, to tailor tests of robustness to the particular threats to validity that appear most serious for a given indicator system. Comparisons should be focused on attributes of the data that are likely to account for significant discrepancies in results. In the terminology of generalizability theory, these attributes identify *facets* across which results may not generalize.

Source of data is one attribute that can cause significant discrepancies in results. For example, as noted above, students' reports of background variables are often strikingly inconsistent with those of their parents, and an exclusive reliance on student reports for such information—as in the case of the NAEP—is therefore potentially a serious threat to validity. Testing the consistency of information across sources of information is an instance of the multitrait-multimethod approach noted earlier; one can have confidence in the results to the extent that they are similar from one method (in this case, one source of information) to another. The method of data collection—for example, observation versus interviews versus questionnaires—can also be important. In the case of achievement indicators, test format can be an important facet.

Although some of the facets that are most important, such as source of data, are important in social science generally, others are specific to educational indicators. For example, measures at different points on the continuum from distal aspects of the intended curriculum to the implemented curriculum in the classroom may yield very different results, and that variation is key to the validity of the inferences each warrants. Ideally, a test of the robustness of curriculum indicators should incorporate measures at the implemented end of the continuum, so that the extent of agreement between distant and proximal measures can be ascertained.

Assess the corruption of indicators. If an indicator system is used in ways that threaten to corrupt some of the indicators in the system, the only way to estimate that corruption—and retain confidence in the indicators—is to test the generalizability of results to measures that are less likely to be corrupted. This is a just a special case of assessing robustness, in which the *proposed uses* to which indicators are put are the facet across which variation in results is assessed.

To obtain alternatives that are less likely to be corrupted, it may be necessary to create *measures that have little salience* to educators and students. Measures that are not used for formal accountability but that are nonetheless salient can be perceived as a source of pressure—and can be corrupted as a result—when other indicators in the same system are used for accountability or monitoring. Achievement tests provide a good example of this; in the current climate, merely publicizing results of tests has sometimes been sufficient to make them "high-stakes" in the eyes of educators, even when no explicit sanctions or rewards are attached to them.

One approach to this problem is to include measures that are collected on a sample basis that does not permit evaluations of the individuals whose behavior might corrupt the indicators in question. For example, sample-based achievement testing in which the data collected are insufficient to evaluate individual teachers or schools might prevent some corruption, although if districts or states can still be evaluated, the potential for corruption from local and state actions remains. Another approach to gauging corruption is to use *benchmarking studies* of the sort described below.

Assess the match of the data to its proposed use. In recent years, indicators have often been used—and misused—opportunistically; an indicator or system that has been created (and perhaps validated) for one purpose is used for quite another, because it is at hand. Two examples were given in earlier sections. Aggregate scores on the SAT, which was validated only as a predictor of college performance and is administered to a fluctuating, non-representative, self-selected sample of students, were nonetheless employed as a purported indicator of states' levels of educational achievement. Another, more general example is the frequent misuse of data designed to provide *descriptions* of the condition of education as the basis for *causal inferences* about the effectiveness of educational programs. An essential step in evaluating indicator systems is to look at the match between the characteristics of indicators or systems and the various uses to which it is put; validation for one purpose need not imply adequacy for another.

Conduct periodic benchmarking studies. In some instances, the steps proposed above can be carried out with data from an indicator system itself. Often, however, additional data will be needed. For example, data on actual classroom instruction are burdensome and expensive to collect and are unlikely to be included in many indicator systems, which makes it uncertain which inferences can be supported by the more distal curriculum measures that are likely to be included. Similarly, in some environments, it may be difficult or impossible to protect even sample-based achievement test data from corruption.

In some instances, the needed external data will be available from other large-scale data collection efforts. For example, trends on the NAEP, which at least until the present was not "taught to" in a way that would corrupt scores, provides evidence that scores on some state- and local-level achievement indicators may have been corrupted, because many of the latter showed sizable increases during the 1980s, while the NAEP showed only very meager gains. The large-scale longitudinal studies fielded by the U.S. Department of Education provide an opportunity to compare certain types of data across sources—for example, student and parent reports of background factors.

Other instances, however, may require specialized *benchmarking* or validation studies that are designed specifically to obtain data that are impractical to collect in larger-scale or less focused data collection efforts. These studies need to be tailored to specific gaps in the validity evidence that can be provided by the indicator system itself. They can often be relatively small, however, and they can be conducted less frequently than data collection for the indicator system.

Benchmarking studies can contribute to validation in several essential ways. By virtue of their lesser frequency and smaller scale, they can provide an opportunity to provide criterion-related evidence that would be too burdensome or expensive to incorporate into an indicator system. For example, benchmarking studies can be used to provide the data on the actual implemented curriculum that is needed to help validate the more distal measures used as indicators. Benchmarking studies can also be used in many ways to test the robustness of the information provided by indicators. For example, they can be used to provide data from additional sources (e.g., parents versus students) or measures (e.g., additional achievement tests with different content or format). Similarly, benchmarking studies can provide the data needed for assessing the corruption of more routinely collected data.

5. CONCLUSIONS

Building a useful and adequate system of indicators involves many choices and compromises. Both schooling and its outcomes are tremendously complex and variable. To be comprehensible and useful for informing debate and policy, an indicator system must reduce this complexity markedly. At the same time, an overly simple system of indicators is likely to mislead, and the boundary between necessary and excessive simplification is anything but clear.

One step in this necessary simplification is defining the core aspects of schooling that an indicator system should monitor (e.g., Shavelson et al., 1987), such as student outcomes, curriculum, and the quality of the teaching workforce. This is only a partial solution, however. These core aspects of schooling are numerous; moreover, they are themselves multi- faceted and encompass constructs that are complex, poorly defined, and very difficult to measure. Comprehensive measurement of these core aspects is therefore often impractical, particularly in the light of the practical and financial limitations imposed by the frequent and large-scale data collection entailed in an indicator system. Even within each of the core aspects of schooling, simplification is inevitable, and the trade-off between necessary and excessive simplification arises.

The unavoidable incompleteness of many indicators is exacerbated by the uses to which they are put. First, because of their role in public debate and policymaking, indicators are often used as a basis for very broad inferences about the condition of education. The broader the inference, however, the more likely a given set of indicators will be insufficient. Second, indicators are increasingly used to hold policymakers, educators, and students accountable. One consequence of this use is that some indicators—test scores are a good example—may become corrupted. This corruption heightens the problem of incomplete measurement, because it makes the available partial measures unrepresentative of the broader constructs they are intended to represent.

This unavoidable simplification has diverse ramifications for building and evaluating indicator systems. One concern is the *adequacy of description* afforded by a set of indicators. For example, to what degree are simple means an adequate basis for characterizing group differences in achievement? Analyses of the National Assessment of Educational Progress presented in another Note from this project (Koretz, 1991a) suggest that this question has no single answer; sometimes additional information (such as information on differences in variances) adds little to the characterization of group differences, while in other instances, it adds important information. In other words, inferences about mean differences—even if reliable and valid—sometimes should be supplemented by other information to obtain a reasonably comprehensive view of group differences.

The tension between simplification and comprehensiveness also colors the issues of validity and reliability discussed in this Note. To begin with, it is one reason why the issues of validity and reliability are more complex in the case of educational indicators than in the case of many measures in everyday life and in the physical sciences. The fundamental criteria of validity and reliability that apply to indicators are essentially the same as those that apply to all measurement. For example, a thermometer that gives too high a readin, is analogous to an achievement test the scores on which have been inflated; in both cases, the systematic bias in the results undermine the validity of the inferences the measures are intended to support. The reading from a properly calibrated fever thermometer, however, is not an incomplete measure of the construct of interest. In contrast, scores on a single achievement test typically are an incomplete measure, and many of the issues of reliability and validity discussed here—such as the robustness of information across alternative measures and the corruption of measures used for accountability—stem directly from that incompleteness.

The issues inherent in simplification cannot be avoided, but they can be addressed in building indicator systems, in evaluating them, and in using them. In constructing an indicator system, unreasonable simplification can be avoided—or at least lessened—by choosing measures carefully and by employing multiple measures of important constructs. In addition, if the indicators are embedded in a broader system of data collection, less frequent or smaller special studies can be used to provide information (for example, data on instructional approaches) that is likely to be lacking in the large-scale routine data collection upon which indicator systems are based. Evaluations of indicator systems should take into account the particular threats to validity and reliability faced by indicators, including their incompleteness and vulnerability to corruption. Here again, the use of both multiple measures within the indicator system and other measures from supplementary studies is likely to be central. Finally, users should be cognizant of the limitations of indicator data and should avoid using them to support overly broad or unduly simple inferences.

REFERENCES

Advisory Panel on the Scholastic Aptitude Test Score Decline, *On Further Examination*, College Entrance Examination Board, New York, 1977.

Airasian, P. W., and G. F. Madaus, "Linking Testing and Instruction: Policy Issues," *Journal of Educational Measurement,* Vol. 20 (2),1983, pp. 103–118.

Alexander, K. L., and M. A. Cook, "Curricula and Coursework: A Surprise Ending to a Familiar Story," *American Sociological Review,* Vol. 47, 1982, pp. 626–640.

Alexander, K. L., and A. M. Pallas, "Curriculum Reform and School Performance: An Evaluation of the "New Basics,'" *American Journal of Education,* Vol. 92, 1984, pp. 391–420.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, American Psychological Association, Washington, D.C., 1985.

Applebee, A. N., J. A. Langer, and I.V.S. Mullis, *Who Reads Best? Factors Related to Reading Achievement in Grades 3, 7, and 11*, Educational Testing Service, Princeton, 1988.

Baratz-Snowden, J., J. Pollack, and D. Rock, *Quality of Responses of Selected Items on NAEP Special Study Student Survey*, Educational Testing Service, Princeton, 1988, unpublished.

Beaton, A. E., J. J. Ferris, E. G. Johnson, J. R. Johnson, R. J. Mislevy, and R. Zwick, *The NAEP 1985-86 Reading Anomaly: A Technical Report*, Educational Testing Service, Princeton, February 1988.

Bianchini, J. C., "Achievement Tests and Differentiated Norms," in M. J. Wargo and D. R. Green, *Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation*, CTB/McGraw Hill, Monterey, California, 1978.

Campbell, D. T., and D. W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, Vol. 56, 1959, pp. 81–105.

Cannell, J. J., *Nationally Normed Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average*, Friends for Education, Daniels, West Virginia, 1987.

Cook, T. D., and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston, 1979.

Cooley, W. W., and G. Leinhardt, "The Instructional Dimensions Study," *Educational Evaluation and Policy Analysis,* Vol. 2 (1), 1980, pp. 7–25.

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam, *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley, New York, 1972.

Dossey, J. A., I.V.S. Mullis, M. M. Lindquist, and D. L. Chambers, *The Mathematics Report Card: Are We Measuring Up? Trends and Achievement Based on the 1986 National Assessment*, Educational Testing Service, Princeton, 1988.

Education Week, "Changing Course: A 50-State Survey of Reform Measures," *Education Week*, Vol. 4 (20), February 6, 1985, pp. 11–30.

Feldt, L. S., and R. L. Brennan, "Reliability," in R. L. Linn (ed.), *Educational Measurement*, third edition, McMillan, New York, 1988, pp. 105–146.

Fetters, W. B., P. S. Stowe, and J. A. Owings, *High School and Beyond: Quality of Responses of High School Students to Questionnaire Items*, National Center for Education Statistics, Washington, D.C., 1984.

Frechtling, J. A., *Alternative Methods for Determining Effectiveness: Convergence and Divergence*. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.

Frederiksen, N., "The Real Test Bias," *The American Psychologist*, Vol. 39 (3), 1984, pp. 193–202.

Gottfredson, G. D., *You Get What You Measure, You Get What You Don't: Higher Standards, Higher Test Scores, More Retention in Grade*, Johns Hopkins University, Center for Research on Effective Middle Schools, Report #29, Baltimore, 1988.

Guskey, T. R., and E. W. Kifer, "Ranking School Districts on the Basis of Statewide Test Results: Is It Meaningful or Misleading?" *Educational Measurement: Issues and Practice*, Vol. 9 (1), 1990, pp. 11–16.

Haertel, E., H. Walberg, J. Baldwin, J. S. Chall, L. V. Hedges, T. Pandey, W. H. Schmidt, and D. E. Wiley, *Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons: Report of the Subpanel on Anomaly and Trends*, 1988, unpublished.

Harnischfeger, A., and D. E. Wiley, *Achievement Test Score Decline: Do We Need to Worry?* CEMREL, Chicago, 1975.

Helmstadter, Gerald C., and Mary M. Walton, *The Generalizability of Residual Indexes of Effective Schooling*. Paper presented at the annual meeting of American Education Research Association, Chicago, April 1985.

Heyns, B., and T. L. Hilton, "The Cognitive Tests for High School and Beyond: An Assessment," *Sociology of Education*, Vol. 55, 1982, pp. 89–102.

Hoover, H. D., personal communication to the author, 1987.

Horn, E. A., and H. J. Walberg, "Achievement and Interest as Functions of Quantity and Level of Instruction," *Journal of Educational Research*, Vol. 77(4), 1984, pp. 227–232.

Jones, L. V., "The Influence on Mathematics Test Scores, by Ethnicity and Sex, of Prior Achievement and High School Mathematics Courses," *Journal for Research in Mathematics Education*, Vol. 18(3), 1987, pp. 180–186.

Jones, L. V., E. C. Davenport, A. Bryson, T. Bekhuis, and R. Zwick, "Mathematics and Science Test Scores as Related to Courses Taken in High School and Other Factors," *Journal of Educational Measurement,* Vol. 23(3), 1986, pp. 197–208.

Kane, M. T., "A Sampling Model for Validity," *Applied Psychological Measurement,* Vol 6, Spring, 1982, pp. 125–160.

Kippel, Gary, "Identifying Exceptional Schools," *New Directions in Program Evaluation*, No. 11, September 1981, pp. 83–100.

Koretz, D., *Trends in Educational Achievement*, Congressional Budget Office, Washington, D.C., 1986.

Koretz, D., *Educational Achievement: Explanations and Implications of Recent Trends*, Congressional Budget Office, Washington, D.C., 1987.

Koretz, D., "Arriving in Lake Wobegon: Are Standardized Tests Exaggerating Achievement and Distorting Instruction?" *The American Educator,* Vol. 12, Summer, 1988, pp. 8–15, 46–52.

Koretz, D., *Indicators of Mathematics and Science Achievement,* RAND, N-3439-NSF, 1991a.

Koretz, D., "State Comparisons Using NAEP: Large Costs, Disappointing Benefits," *Educational Researcher,* 20(3), April 1991b, pp. 19–21.

Koretz, D., R. L. Linn, S. B. Dunbar, and L. A. Shepard, "The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests," in R. L. Linn (Chair), *Effects of High Stakes Educational Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991.

Laing, J., H. B. Engen, and J. Maxey, *Relationships Between ACT Test Scores and High School Courses,* Research Report #87-3, ACT, Iowa City, Iowa, 1987.

Linn, R. L., "Curricular Validity: Convincing the Courts that It Was Taught Without Precluding the Possibility of Measuring It," in G. F. Madaus (ed.), *The Courts, Validity, and Minimum Competency Testing,* Kluwer-Nijhoff, Boston, 1983, pp. 115–132.

Linn, R. L., "Accountability: The Comparison of Systems and the Quality of Test Results," *Educational Policy,* Vol. 1, 1987, pp. 181–198.

Linn, R. L, M. E. Graue, and N. M. Sanders, "Comparing State and District Test Results to National Norms: The Validity of Claims that Everyone is Above Average," *Educational Measurement: Issues and Practice,* Vol. 9 (3), 1990, pp. 5–14.

Madaus, G. F., "The Influence of Testing on the Curriculum," in *Critical Issues in Curriculum: 97th Yearbook of the National Society for the Study of Education*, University of Chicago Press, Chicago, 1988, pp. 83–121.

Madaus, G. F., T. Kellaghan, E. A. Rakow, and D. J. King, The Sensitivity of Measures of School Effectiveness," *Harvard Educational Review,* Vol. 49 (2), 1979, pp. 207–230.

Mandeville, Garrett K., and Lorin W. Anderson, "The Stability of School Effectiveness Indices Across Grade Levels and Subject Areas," *Journal of Educational Measurement*, Vol. 24 (3), Fall 1987, pp. 203–214.

Massachusetts Department of Education (Bureau of Research and Assessment), *Course Taking Among Massachusetts High School Students*, publication #14476-56-1000-6-86-C.R., Boston, 1985.

Massachusetts Department of Education (Bureau of Research and Assessment), *The High School Experience in Massachusetts*, publication #14438-74-500-5-86, Boston, 1986.

Matthews, T. A., Soder, J. B., M. C. Ramey, and G. H. Sanders, *Use of Districtwide Test Scores to Compare the Academic Effectiveness of Schools*. Paper presented at the annual meeting of the American Educational Research Association, April 1981.

Maxey, J., S. Cargile, and J. Laing, "Three Measures of Academic Achievement and Their Association with Performance on the ACT Assessment," *NCEOA Journal,* Spring 1987, pp. 6–10.

Messick, S., "Validity," in R. L. Linn (Ed.), *Educational Measurement*, third edition, McMillan. New York, 1988, pp. 13–103.

Meyer R. H., *Applied Versus Traditional Mathematics: New Econometric Models of the Contribution of High School Courses to Mathematics Proficiency*, working paper, National Assessment of Vocational Education, Washington, D.C., 1988.

Miller, M. D., and R. L. Linn, "Invariance of Item Characteristic Functions with Variations in Instructional Coverage," *Journal of Educational Measurement,* Vol. 25 (3), 1988, pp. 205–220.

Mullis, I., Comments of the Chair in *The NAEP Report Cards: Theory, Methods, and Implications for Policy and Practice*, Symposium delivered at the annual meeting of the American Educational Research Association, New Orleans, April 1988.

Mullis, I.V.S., and L. B. Jenkins, *The Science Report Card: Elements of Risk and Recovery. Trends and Achievement Based on the 1986 National Assessment*, Educational Testing Service, Princeton, 1988.

Murnane, R., and S. Raizen, *Improving the Quality of Indicators of Science and Mathematics Education in Grades K–12,* National Academy Press, Washington, D.C., 1988.

National Assessment of Vocational Education, *First Interim Report from the National Assessment of Vocational Education,* U.S. Department of Education, Washington, D.C., 1988.

National Center for Education Statistics, *The Condition of Education, 1982 Edition*, U. S. Department of Education, NCES Report #NCES-82-400, Washington, D.C., 1982.

National Commission on Excellence in Education, *A Nation at Risk*, U.S. Government Printing Office, Washington, D.C., 1983.

National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics*, National Council of Teachers of Mathematics, Reston, Virginia, 1989.

Office of Planning, Budget, and Evaluation, *State Education Statistics: State Performance Outcomes, Resource Inputs, and Population Characteristics, 1982 and 1984*, U.S. Department of Education, Washington, D.C., 1985.

Raizen, S., and L. V. Jones, *Indicators of Precollege Education in Science and Mathematics*, National Academy Press, Washington, D.C., 1985.

Resnick, L., *Education and Learning to Think,* National Academy Press, Washington, D.C., 1987.

Rock, D. A., R. B. Eckstrom, M. E. Goertz, and J. Pollack, *Study of Excellence in High School Education: Longitudinal Study, 1980–82 Final Report,* Center for Statistics, U.S. Department of Education, Washington, D.C., Contractor Report CS 86-231, 1986.

Rowan, Brian, and Charles E. Denk, *Modelling the Academic Performance of Schools Using Longitudinal Data: An Analysis of School Effectiveness Measures and School and Principal Effects on School-level Achievement*, Far West Laboratory, San Francisco, 1983.

Salmon-Cox, L., *MAP Reading End-of-Year Report*, Learning Research and Development Center, Pittsburgh, 1984, unpublished.

Schmidt, W. H., "High School Course-Taking: Its Relationship to Achievement," *Journal of Curriculum Studies*, Vol. 15 (3), 1983, pp. 311–332.

Schrock, J., Personal communication to author, May 1988.

Shavelson, R. J., *Historical and Political Considerations in Developing a National Indicator System*. Paper presented at the annual meeting of the American Educational Research Association, April 1987.

Shavelson, R. J., L. McDonnell, J. Oakes, and N. Carey, *Indicator Systems for Monitoring Mathematics and Science Education*, RAND, R-3570-NSF, 1987.

Shavelson, R. J., and N. M. Webb, "Generalizability Theory: 1973-1980," *British Journal of Mathematical and Statistical Psychology,* Vol. 34, 1981, pp. 133–166. (Reprinted by RAND under the same title, P-6580.)

Shepard, L. A., *Should Instruction Be Measurement-Driven?* Presented at the annual meeting of the American Educational Research Association, New Orleans, April 1988.

Shepard, L. A., "Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?" *Educational Measurement: Issues and Practice,* Vol. 9 (3), 1990, pp. 15–22.

Shepard, L. A., "The Effects of High-Stakes Testing on Instruction," in R. L. Linn (Chair), *Effects of High Stakes Educational Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991.

Spencer, B., "On Interpreting Test Scores as Social Indicators: Statistical Considerations," *Journal of Educational Measurement,* Vol. 20 (4), 1983, pp. 317–333.

Stecher, B., *Developing Indicators of Secondary-Level Curriculum in Mathematics and Science*, RAND, N-3406-NSF, 1991.

Webb, E. J., D. T. Campbell, R. D Schwartz, and L. Sechrest, *Unobtrusive Measures*, Rand-McNally, Skokie, Illinois, 1966.

Welch, W. W., R. E. Anderson, and L. J. Harris, "The Effects of Schooling on Mathematics Achievement," *American Educational Research Journal,* Vol. 19 (1), 1982, pp. 145–153.

Wilson, B., and R. D. Corbett, "Two State Minimum Competency Testing Programs and Their Effects on Curriculum and Instruction," in R. Stake (ed.), *Advances in Program Evaluation: Effects of Mandated Testing on Teaching*, Vol. 1, Part B, JAI Press, Greenwich, Connecticut, 1991, pp. 7–40.

RAND/N-2900-NSF