APPENDIX 3: KIRIS Summary

INTRODUCTION

Kentucky’s educational reform program, named KERA for the Kentucky Education Reform Act from which it emanated, is one of the most prominent and influential standards-based reform efforts in the nation. KERA has entailed a thorough overhaul of the state’s K-12 educational system. It has altered, for example, the financing of schools, the organization of primary grades (instituting a statewide ungraded primary program), professional development, the governance of schools, and the state’s assessment program, and it instituted a statewide program of school-level accountability for performance.

Despite its scope, KERA is known outside the state primarily because of its assessment and accountability system—the Kentucky Instructional Results Information System (KIRIS). This system used a complex, partially performance-based assessment, in conjunction with noncognitive indicators such as dropout rates that are given less weight, to judge the performance of schools. Scores were reported annually, but schools were evaluated every two years on the basis of changes in performance from one two-year average to the next. These periods were called bienniums or accountability cycles. Each school was assigned a performance target for each cycle based solely on its performance level during the first two years of that cycle. Schools that fell below their targets could be sanctioned, and schools that exceeded them by a sufficient amount received cash rewards. This system was a source of intense controversy within the state almost from its inception, but it has been viewed as a promising model by many reformers and observers of education throughout the nation.

KERA quickly produced large gains in scores on the KIRIS assessment. These gains were seen by many as signs that the reforms were improving educational performance; for example, the Kentucky Department of Education (KDE) entitled its press release of 1994 scores “Celebrate the Progress!” (Kentucky Department of Education, 1995a). Assessment-based accountability, however, can lead to inflated gains. That is, scores can increase more than actual student learning warrants, creating an illusion of progress. This inflation can arise from numerous factors. For example, teachers may de-emphasize important material that is not likely to be tested in order to free up time for material emphasized on the test (e.g., Koretz, Barron, Mitchell, and Stecher, 1996). That is, they may focus on the content of the test itself rather than on the domain it is supposed to represent, thus raising scores more than they improve mastery of the domain. In extreme cases, teachers may focus on the content of specific test items. They may also narrow instruction relative to the intended domain by tailoring instruction narrowly to the rubrics used to score student work on the assessment (e.g., Stecher and Mitchell, 1995). And they may engage in inappropriate activities during testing sessions.

In the case of systems using the more traditional, multiple-choice tests, the inflation of gains can be large, and the entire gain may even be illusory in some instances.
(Koretz, Linn, Dunbar, and Shepard, 1991). Indeed, some of the impetus for the “second wave” of education reform, in which educators are held accountable for scores on various performance assessments, was a growing recognition of the inflation of gains that can arise when traditional tests are used for accountability. There is no evidence, however, to indicate that forms of testing other than multiple choice are less susceptible to this problem. Therefore, to evaluate the impact of the KERA program on student learning, it is essential to validate the observed gains in KIRIS scores—that is, to evaluate the extent to which these gains support the inference that student learning has indeed increased. This monograph reports the results of a study that used a variety of approaches to assess the validity of KIRIS gains from the first administration in 1992 through the 1995 and 1996 administrations.

**SCOPE OF THE STUDY**

While this study represents an unusually intensive effort to validate gains, it has major limitations. Methods for validating gains are relatively primitive, and the data useful for this purpose in Kentucky—as in most states—are severely limited. Most traditional methods of validation do not take the possibility of inflated gains into account. Indeed, most of these methods are cross-sectional and therefore cannot address issues of change over time. Conventional methods for validating tests include examining the adequacy of an assessment’s content, the behavior of individual items (e.g., tests of homogeneity and item bias), and convergent/discriminant evidence (the extent to which an assessment’s correlations with other measures conform to expectations). All of these methods are typically applied cross-sectionally, and while they are essential for validating gains, they are insufficient for that purpose. For example, it will be shown below that scores on an accountability-oriented assessment and another assessment may maintain their correlation (a cross-sectional indication of validity) even while trends in mean scores on the two assessments show a huge and rapid divergence that calls the validity of gains into doubt.

Relatively little attention has been given to developing methods directly germane to the validation of gains. This report discusses cross-sectional evidence of validity only incidentally. Other efforts have been under way for years to obtain such evidence. For example, considerable evidence of this sort can be found in the KIRIS technical reports released periodically by KDE and in the report of an expert panel commissioned by the Office of Education Accountability (OEA) of the Kentucky General Assembly (Hambleton, et al., 1995). Although this evidence is mixed, it is not discussed here. Rather, we evaluate evidence directly pertinent to the validity of changes in scores.

Two broad types of evidence are discussed here: internal and external. Internal evidence derives from the KIRIS assessment itself. For example, one can look for clues about the validity of gains in performance differences on new test items versus items carried over from previous years. External evidence is based on comparisons of scores on KIRIS and other tests.
Only three sets of test scores suitable for comparison with KIRIS scores are available for much or all of Kentucky: scores from the National Assessment of Educational Progress (NAEP), from the American College Testing (ACT) college-admissions tests, and from two commercial achievement test batteries—the CTBS4 (Comprehensive Tests of Basic Skills) and CAT5 (California Achievement Tests)—currently used by many Kentucky districts. Of these three, the CTBS4 and CAT5 scores are not considered here for a variety of methodological reasons. (A comparison of trends on these tests and KIRIS can be found in Nitko, Stone, and Wang, 1997.) ACT scores are used, but ACT data are for a large but self-selected sample of high school students, and the overlap in content between the ACT and KIRIS tests is only moderate. NAEP scores provide the strongest evidence.

The KIRIS assessment frameworks in reading and mathematics were expressly modeled after NAEP’s, and NAEP occasionally tests representative samples at the state level. These samples, however, are obtained only infrequently. It was possible to compare NAEP and KIRIS for grade 4 reading from 1992 through only 1994 and for grade 4 and grade 8 mathematics from 1992 through 1996. It is important to note that our analyses considered only the assessments from 1992 through 1995 and 1996. There were no aggregate gains on KIRIS in 1996, but our analysis of NAEP data necessarily included 1996 because NAEP was not administered in 1995.

A period of four years may seem long enough for evaluating an assessment program such as KIRIS, but it may not be. Experience has shown that scores often rise rapidly during the first few years of a testing program as familiarity with the test grows. A substantial share of these gains may not generalize to other tests and therefore may be seen as inflation of gains. Thus, one might find substantial inflation of gains over the first two to four years of an assessment and more meaningful gains thereafter. Moreover, KIRIS was evolving during the period covered by the analysis, and its effects in later years might have differed as a result. Nonetheless, the validity of KIRIS gains during the first four years is not a trivial concern—these gains have been widely presented as evidence of the program’s success and have been used to allocate tens of millions of dollars in rewards to schools.

**PLAN OF THE REPORT**

Chapter 2 provides a very brief description of the KIRIS assessment and accountability system. The system is complex and has been documented in many other publications, so no effort is made to provide a detailed description here. It is necessary, however, to explain the basic characteristics of the system in order to make the evidence presented later in the report meaningful.

Chapter 3 provides a framework for validating gains on the KIRIS assessment—in particular, for evaluating external comparisons with other test data.
Chapter 4 describes trends in scores on KIRIS over the first four years of the program. Again, more detail is available in other reports. The purpose here is to describe the overall, statewide trends in performance, the validity of which is the focus of this report. To provide consistency over time and across the various types of analyses, the results reported here have all been calculated by RAND from Kentucky’s assessment data and do not necessarily agree completely with those published by KDE. In most cases, any discrepancies are minor. A few, more substantial differences are noted in the text.

Chapter 5 summarizes evidence from past studies pertaining to the validity of score gains on KIRIS. Two documents are summarized: (1) the report of an expert panel (often called “the OEA Panel” or “the Hambleton Panel”) convened by an agency of the Kentucky General Assembly to evaluate KIRIS after its first three years (Hambleton, et al., 1995); (2) a RAND series of surveys of Kentucky teachers and principals (Koretz, Barron, Mitchell, and Stecher, 1996).

Chapters 6 through 9 describe the results of our analyses. Chapter 6 discusses trends in the retention of students in grade. Chapters 7 and 8 discuss external evidence of the validity of gains, specifically, trends for NAEP and then for the ACT. In both cases, the findings of the OEA Panel (Hambleton, et al., 1995) are presented and then extended with additional analyses of newer data. Chapter 9 then presents a variety of internal evidence created from the KIRIS performance data themselves. It also details the several unconventional methods we used in our analyses.

Chapter 10 describes a number of the characteristics of KIRIS, such as changes in the design of the assessment over time, that might bear on the interpretation of the data presented in the earlier chapters.

Chapter 11 summarizes the empirical evidence, discusses its bearing on the validity of KIRIS score gains, and offers implications for both policy and further research.

The Appendix provides additional methodological detail pertaining to several aspects of the study: the creation of the KIRIS trend database we used; the methods used to standardize the KIRIS, NAEP, and ACT data; and the steps taken to create a merged file of ACT and KIRIS data.