The Kentucky Instructional Results Information System (KIRIS) is a multidimensional measurement and assessment system that supports the statewide educational accountability system in Kentucky. It was initiated by the Kentucky Department of Education in 1991 in response to a comprehensive statewide educational reform law. KIRIS collects data on cognitive outcomes in grades 4, 8, and 12 and combines them into one accountability index for each school. Schools that achieve adequate gains on the index receive financial rewards; consistent failure triggers state intervention. The cognitive measures are primarily performance based; they include on-demand constructed-response questions and performance events, as well as portfolios.

Educators in the state report that KIRIS has had strong effects on curriculum and instruction. External evaluators invited by the legislature to review KIRIS raised serious concerns about the quality of the measures and their validity for the state’s purposes. The Kentucky Department of Education has made many changes to the system in response to these concerns, and additional changes were being considered at the time of this writing. KIRIS is the first example of a strong statewide accountability system built on performance measures that has been implemented, and it has interesting lessons for all educators.

DESCRIPTION AND PURPOSE

The Kentucky Educational Reform Act of 1990 (KERA) represented a dramatic reform of the state’s educational system, with a strong em-
phasis on accountability. KERA embodied a particular approach to education in that it

- Set goals for the educational system.
- Created a mechanism for assessing progress toward those goals.
- Established rewards and sanctions for schools based on performance.

Figure B.1 lists the six major goals that were set for learners in the areas of basic communication and mathematics skills; application of concepts and principles to real-life situations; self-sufficiency; school attendance; school dropout/retention rates; elimination of barriers to learning; and transitioning from high school to work, further

1. Students are able to use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives.

2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, practical living studies, and vocational studies to what they will encounter throughout their lives.

3. Students shall develop their abilities to become self-sufficient individuals.

4. Students shall develop their abilities to become responsible members of a family, work group, or community, including demonstrating effectiveness in community service.

5. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life.

6. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned, and build on past learning experiences to acquire new information through media sources.

Figure B.1—Kentucky’s Six Learner Goals
education, or the military (Kentucky Department of Education, 1993). Schools are the basic unit used to measure performance in Kentucky. The state expects schools to steadily improve their performance relative to these six goals.

The Kentucky Department of Education was charged with creating a system to measure and report school performance against these goals; KIRIS was the result. KIRIS scores are made up of two components, noncognitive and cognitive measures. The noncognitive measures account for about 16 percent of the total score for a school and include attendance, retention, dropout, and transition rates. The cognitive measures, which are collected only in grades 4, 8, and 12, cover the core academic subjects (mathematics, reading, science, social studies, writing, humanities and the arts) and practical living and vocational studies. Standards for performance have been set for the cognitive measures, and student work is classified into one of four performance levels: Novice, Apprentice, Proficient, or Distinguished.

The cognitive measures, which are primarily performance based, originally included open-ended items, performance events, and portfolios (Kentucky Department of Education, 1995a). The open-ended items are in both the short-answer and the essay format. Performance events, which last about one class period, include some group work followed by individual work leading to an individual written product. They are administered on a matrix-sampled basis, with each student working on just one or two events. Portfolios are collected in writing and mathematics; each one contains five to seven “best pieces” of student work that cover different content areas and different core concepts. There are no content requirements for the portfolios, but they are supposed to demonstrate breadth as well as higher-order skills in each domain.

Measures from all domains (cognitive and noncognitive) are combined into a single accountability index for each school. The relative weights assigned to the content areas for the next cycle of accountability are summarized in Table B.1. A baseline index was computed using 1991–1992 performance, and an improvement target was

---

1From 1991–1992 to 1993–1994, the number of multiple-choice items was cut in half and the number of open-response items was doubled.
Table B.1

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics</td>
<td>14</td>
</tr>
<tr>
<td>Reading</td>
<td>14</td>
</tr>
<tr>
<td>Science</td>
<td>14</td>
</tr>
<tr>
<td>Social studies</td>
<td>14</td>
</tr>
<tr>
<td>Writing</td>
<td>14</td>
</tr>
<tr>
<td>Arts and humanities</td>
<td>7</td>
</tr>
<tr>
<td>Practical living/vocational studies</td>
<td>7</td>
</tr>
<tr>
<td>Noncognitive index</td>
<td>16</td>
</tr>
</tbody>
</table>

established using this score. \(^2\) (Greater gains on the baseline index are expected for low-scoring schools than for high-scoring schools.) Subsequent biennial averages are used as baselines for future improvement targets. Kentucky's long-term goal over a twenty-year period is that all schools will score above the 100 level (which is equivalent to having all students at the Proficient level).

Kentucky has a strong commitment to inclusion, and very few students are excluded from participation in the assessment. Special-educa-tion students complete an alternative portfolio based on their individual educational plans. Scores from these students are included in the computation of the school's accountability index.

**RELATIONSHIP TO OTHER PROGRAMS**

KERA created the framework for a new educational system incorporating the six goals shown in Figure B.1. KIRIS is the measurement and accountability system created to support KERA. KIRIS is seen as one part of the “complex network intended to help schools focus their energies on dramatic improvement in student learning” (Kentucky Department of Education, 1995b, p. 2). The state's goal is to create an integrated program of assessment, accountability, curriculum reform, and staff support. Because there are high stakes at-

\(^2\)Recently, changes were made to improve the technical quality of KIRIS scores. For example, performance events are no longer included in the accountability index.
attached to performance, education officials expect to observe “teaching to the test,” so they have tried to design an assessment system based on events worth “teaching to.”

In accordance with KERA, KIRIS was built to assess school performance against the six broad learner goals shown in Figure B.1. KERA also required the department of education to create a performance-based assessment program to measure success. Goals 1, 2, 5, and 6 address the application of cognitive skills, and the contractor responsible for developing KIRIS worked with educators in Kentucky to develop assessments that measured these cognitive outcomes. The learner goals themselves are too broad to serve as test specifications, so in 1991 the Kentucky Board of Education adopted a more detailed set of valued outcomes that described in greater detail the skills learners should possess in the fields of mathematics, science, art, humanities, social studies, practical living, and vocational studies.

For the next two years, these outcomes were used as the basis for developing assessment tasks. However, these outcomes proved to be confusing to many important audiences, including parents, and were replaced by a set of fifty-seven “academic expectations” describing what Kentucky students should know and be able to do when they graduate from high school. Subsequent KIRIS assessment development has focused on these academic expectations (Kentucky Department of Education, 1995b).

KIRIS was built to assess school performance in response to broad new demands placed on education. The associated outcomes or expectations were derived by panels of educators to reflect this new direction, not existing programs. In particular, the vocational outcomes are quite general and do not necessarily match with the objectives of specific vocational programs. Only three of the fifty-seven academic outcomes relate to vocational studies:

- Students use strategies for choosing and preparing for a career.
- Students demonstrate skills and work habits that lead to success in future schooling and work.
- Students demonstrate skills such as interviewing, writing resumes, and completing applications that are needed to be
accepted into college or other postsecondary training, or to get a job.

KIRIS is not focused specifically on assessing learning in vocational classes. In both 1992–1993 and 1993–1994, only three performance events and eleven open-response items per grade level were used to assess practical living and vocational studies combined, and this content area counted for only 7 percent of the overall accountability index (Kentucky Department of Education, 1995b). Most students completed only one performance event and one open-response item in this domain. This does not provide enough information to be useful for evaluating vocational programs, either at the individual or program level. Over time, one might expect to see greater coordination between specific instructional activities and the statewide assessment. Furthermore, the career skills measured by KIRIS might be useful indicators of one aspect of vocational education. However, as presently conceived, KIRIS itself will not be sufficient for evaluating specific vocational programs. Rather, vocational educators may be able to learn about performance-based accountability systems from the KIRIS model.

IMPLEMENTATION AND ADMINISTRATION

Kentucky has supported the implementation of KIRIS with extensive teacher training and technical assistance (Kentucky Department of Education, 1993). It established eight regional service centers to train district staff as associates, whose role is to help their districts further professional development. Districts and schools report that the centers are a valuable resource. The department of education funded a program to train KERA assessment fellows, whose role is to be available throughout the state to help schools and districts prepare for and interpret KIRIS; over 300 educators have participated in this program. Over 100 teachers have been trained as Distinguished educators, their function being to help schools succeed (particularly those whose scores are low). The Kentucky educational television network has broadcast fourteen professional development sessions. In addition, colleges and universities in Kentucky have offered courses and contracted with individual districts to train teachers in the new assessment methods and other aspects of KERA school reform.
The contractor responsible for KIRIS has trained 700 people as mathematics portfolio cluster leaders to help teachers in their area understand the portfolio guidelines and implement appropriate classroom procedures. Over 1,000 teachers have participated in guided scoring practice workshops for the writing portfolios. Teachers also have been involved in summer scoring of portfolios, which they report is beneficial for their professional development. Overall, the state has engaged in a broad and thorough effort to provide information and training to prepare teachers for the new assessment and accountability system.

Kentucky contracted with Advanced Systems in Measurement and Education (ASME) to develop and administer KIRIS. ASME worked closely with teams of Kentucky educators to formulate plans for the assessment, develop test items and open-response tests, administer the performance events, score the assessments, and set standards for student performance. ASME, in turn, contracted with WestEd for collection and analysis of the noncognitive data on attendance, retention, dropout, and transition rates.

It is difficult to estimate the total cost of KIRIS. The contractor receives about $6 million per year for developing the assessments, administering them, scoring the results, and reporting to schools and the state. This funding also covers some staff development activities. The Kentucky Department of Education also spends about $2 million a year on professional development of this type for teachers. In addition, some districts contract separately with ASME for additional scoring for continuous assessment, and the annual budget for rewards to schools is estimated to be about $18 million (Kentucky Institute for Education Research, 1995).

In addition, the KIRIS assessment requires school time. Each student completes four periods of on-demand assessment (periods were ninety minutes long in grades 8 and 12, and sixty minutes long in grade 4). If students need additional time, they are given a half-period more to complete the activities. Each student also devotes one period to a performance event, administered at the school by ASME staff. Writing and mathematics portfolios are collected throughout the year, but we were unable to find an estimate of the additional time spent preparing the portfolios (above and beyond the time required to do the assignments).
Teachers also devote some school time to preparing for KIRIS, but whether this is a cost or a benefit depends on the nature of the activities. KIRIS is designed to promote changes in curriculum and instruction, and, in theory, the time schools devote to preparing for KIRIS can be considered instructional time. Surveys administered by RAND suggest that teachers put considerable time into test preparation (Koretz, Mitchell, Barron, and Stecher, 1996). However, there is little evidence indicating whether this is appropriate preparation (activities that promote improvement in the broad domain of skills measured by KIRIS) or inappropriate preparation (time spent narrowly preparing students for specific KIRIS tasks or activities that might not generalize beyond the particular content of the test).

TECHNICAL QUALITY

In 1994, a panel of distinguished measurement specialists was appointed to investigate the technical quality of KIRIS. Their specific charge was to determine whether the accountability index was sufficiently robust to support how it was being used. The panel concluded that KIRIS is “significantly flawed and needs to be substantially revised” (Hambleton et al., 1995, Exec. Summary, p. 1), and it made fourteen recommendations for improving the system. The panel members were particularly concerned that the public was being misinformed “about the extent to which student achievement has improved” and about the “accomplishment of individual students” (ibid, p. 5).

The panel based its conclusion on evidence relating to six aspects of KIRIS, all of which are important considerations in the use of alternative assessments in vocational education. Each is discussed briefly in the following paragraphs, much of the discussion adapted directly from Hambleton et al. (1995).³

³In 1995-1996, the Kentucky Department of Education and its assessment contractor implemented all of the evaluation panel’s recommendations with the exception of eliminating the writing portfolios. Analysis of the 1995-1996 writing portfolios showed continued improvement in scoring reliability and accuracy.
Assessment Development Process

The greatest weakness that the panel found in the development and documentation process was that the specifications (frameworks) do not communicate clearly what students are expected to know and be able to do, and therefore do not provide adequate signals to teachers and test developers. Since the test emphasizes cross-cutting themes rather than traditional discipline-based knowledge, an understanding of the exact nature of the expectations is important. In Kentucky, the test frameworks vary in detail and specificity across subjects, and frequently they do not contain any information about variations in expected student performance across grade levels. It is important to note that the greatest weaknesses in this area were found in the first year, and the process has been improving since then.4 The panel was also critical of the process that was used to develop assessments, recommending that the state clearly follow four steps: specify goals explicitly, construct exercises that measure progress toward these goals, evaluate the exercises by having judges examine pilot-test results from students, and select and assemble test forms using acceptable items.

Accountability Index

A second problem was that the scores reported for schools did not have adequate reliability for accountability purposes: the scores reported for students were less reliable than the usual standard for such tests. The panel concluded that a substantial number of schools probably were assigned to the wrong reward category (Eligible for Reward, Successful, Improving, Decline, In Crisis), and that such errors of assignment were particularly likely for small schools. Furthermore, there was inadequate information to determine the likely level of error due to differences in task sampling from year to year, so the problems the panel was able to identify probably underestimated the true error of classification. Another problem was that student score reports did not convey information about the margin of error of reported scores, which should be included,

---

4Unfortunately, scores from the first year helped to establish each school’s baseline performance level, so the initial weak test development process affected later rewards and sanctions.
according to accepted test standards (American Psychological Association, 1985). The panel noted that reliability of both student scores and school scores (i.e., information used for assessment purposes and for accountability purposes) could be improved by using both multiple-choice and open-response tasks to obtain scores, an option that was rejected by Kentucky in its commitment to emphasize performance assessment.

**Portfolio Scoring Procedures**

The panel examined separately the scores generated by the portfolio component of KIRIS, reporting negative findings about the reliability and validity of these scores as well. It is important to remember that the Kentucky portfolios served dual purposes: to provide measures of student achievement for the accountability system and to encourage changes in curriculum and instruction. On the first point, the panel found that scores were insufficiently reliable to support their use for accountability. Specifically, although raters were moderately consistent in ranking student work, they disagreed about the percentage of portfolios reaching each of the KIRIS performance levels. More damning was the fact that ratings by students’ own teachers were higher than ratings by independent judges.

There was little evidence available about the validity of scores, but the panel was particularly concerned about the lack of standardization in the way portfolio entries are produced and the amount of assistance students receive. (This is a problem that undermines the validity of portfolio scores in other states as well.) Another problem of interpretation is that portfolios constructed of “best pieces” may not reflect sustainable levels of performance under normal conditions. The panel was more optimistic about the potentially beneficial effects of the portfolios on curriculum and instruction. Little information had been gathered about instructional impact at the time of the review, but, based on evidence from other portfolio assessment systems, the panel encouraged Kentucky to maintain the system on a low-stakes basis while gathering evidence about its long-term effects on classrooms.
Making Scores Comparable Across Years (Equating)

Next, the panel tackled the difficult question of the comparability of scores over time. KIRIS allocates rewards and sanctions on the basis of comparisons between performance in baseline years and subsequent years. Therefore, it is essential that the scores be comparable from one administration to the next even though the tasks, events, and items may vary. Although much of the panel’s analysis was highly technical, involving the appropriate statistical equating designs, its conclusions were clear: the equating process was insufficient. KIRIS used too many judgmental procedures without adequate standardization, particularly in the translation from raw scores to performance levels. This introduced errors into the year-to-year comparisons. Other problems that undermined the equating of scores from year to year included changes in procedures and the exclusion of multiple-choice items (which have higher reliability) from the accountability index. Overall, the panel found that the equating did not support year-to-year comparisons, and it recommended a number of changes to strengthen the process.

Setting Performance Standards

Classification of students into proficiency levels is at the core of KIRIS, and the accuracy of these classifications affects the accuracy of each school’s accountability index. Students are classified as Novice, Apprentice, Proficient, or Distinguished on each assessment, based on their scores. The assignment of scores to proficiency levels is done through judgmental processes in which panels review student responses and classify them according to descriptions of performance at the four levels. The investigatory panel found that these processes were not adequately described and appeared to lack appropriate standardization. It particularly criticized the standard-setting process, which at times assigned students to a proficiency level on the basis of as few as three test items.

Impact on Student Learning

Finally, the panel looked at the evidence of educational improvement in Kentucky; in other words, Has KIRIS had the desired effects
on student performance? The Kentucky Department of Education trumpeted the improvement in student scores from 1991–1992 to 1993–1994, and the general public was led to believe that substantial progress had been made. The panel tried to determine to what extent these score changes reflect real differences in student learning. It concluded that the reported gains “substantially overstate improvements in student achievement” (Hambleton et al., 1995, sec. 8, p. 2). Panel members based this judgment on external evidence on student performance, such as NAEP, which does not show any improvement over the same time period (although there is a limit to how many such comparisons can be made at the same grade level and for the same subject). Though the panel members could not explain the differences, they suggested that inflated gains were attributable to two factors: the high stakes attached to KIRIS led to inappropriate teaching to the test, and the desire to show big increases in scores led to overly poor performance during the baseline year.

CONSEQUENCES AND USE OF ASSESSMENT RESULTS

The accountability index was used for the first time in 1994 to reward and sanction schools. Each school received a detailed report of its students’ performance and its overall accountability index. Additional money was awarded to schools that met the threshold for rewards. The reports have been used in a variety of ways that are “consistent with the intent of KIRIS” (Kentucky Department of Education, 1995b, p. 222), including to monitor the progress of programs over time and to target instructional program improvement efforts.

KERA and KIRIS have had broad effects on curriculum assessment and professional development. There is clear evidence that some teachers are changing instructional practices in response to KIRIS assessment content and processes. For example, the use of writing portfolios has led to an increased emphasis on student writing. However, there is evidence that teachers are lagging in reforming many practices, including some assessment-related ones. They “are struggling with the use of learning centers and theme-centered units; are failing to use recommended practices in science, social studies and the arts; are not planning their instructional program around
Kentucky’s Learning Goals and Academic Expectations; are having difficulty implementing a variety of continuous, authentic assessments; are neglecting to plan with special area teachers; and [are] failing to involve parents in the primary program” (Kentucky Institute for Education Research, 1994, pp. xvii–xviii).

**APPLICABILITY TO VOCATIONAL EDUCATION**

Much can be learned from KIRIS that has value for vocational education. On the positive side, some of the changes that proved most difficult for Kentucky educators should be relatively easy for vocational educators already accustomed to using performance as a basis for assessment. Similarly, the development of clear descriptions of desired outcomes and student proficiencies that has proved so difficult in Kentucky is very akin to the task analyses common in vocational education and so should create fewer problems. When vocational educators try to design assessments to measure unfamiliar skills and performances (e.g., generic skills, such as teamwork or understanding of systems), they will face similar problems of definition and communication, but their experience with task delineation and performance specification should stand them in good stead.

On the negative side, strong accountability requirements seem to make most aspects of assessment more difficult. Greater resources will be needed for everything from development to training to implementation if such an assessment is used to structure an accountability system.

Not one of the assessment elements of KIRIS is new; other testing programs use portfolios, performance events, and open-ended responses, and other states produce school “report cards” with indicators of both cognitive and noncognitive outcomes. What is unique about KIRIS is the use of these measures in a strong accountability context. The presence of high stakes exacerbates the political problems, raises the necessary technical standards, and heightens the anxiety level of educators, all of which would make it difficult to implement KIRIS-like assessments in similar contexts. The use of a single summary index of performance without the high stakes might be beneficial for some purposes, however.
Of particular concern is the need for high-quality measurement, a goal that still eludes KIRIS after four years (according to the technical experts). Such quality standards increase the time and resources needed for all aspects of the assessment, including developing student outcome goals, producing assessment specifications, developing tasks, scoring student responses, setting standards, equating forms, and reporting. These types of technical issues must be confronted by vocational educators if they want to use performance assessment for certifying competency, awarding certificates of mastery, or other important ends. In fact, the technical demands will be greater if the assessments are going to be used to make decisions about individuals. The KIRIS experience suggests that such an approach will require advanced technical expertise as well as considerable time and resources. Despite the criticisms of KIRIS, political support for educational improvement continues in Kentucky, as do efforts to improve the accountability system.