

Chapter 8

Learning from Kentucky's Failed Accountability System

George K. Cunningham

The law that mandated educational reform in Kentucky is called the Kentucky Education Reform Act (KERA), and it was originally implemented as a response to a decision rendered by Judge Ray Corns of the Franklin Circuit Court in the *Rose vs. Council for Better Education*.¹ Judge Corns ruled that the Kentucky General Assembly had failed to provide an efficient system of common schools as required by the state constitution. In his ruling, Kentucky's entire legal framework for education was ruled unconstitutional. He also ruled that the system of school financing was inefficient and discriminatory. He did not restrict himself to these matters alone. His ruling included these additional requirements:

Lest there be any doubt, the result of our decisions is that Kentucky's *entire system* of common schools is unconstitutional. There is no allegation that only part of the common school system is invalid, and we find no such circumstance. This decision applies to the entire sweep of the system—all its parts and parcels. This decision applies to all

the statutes creating, implementing, and financing the system and to all regulation, etc., pertaining thereto. This decision covers the creation of local school districts, school boards, and the Kentucky Department of Education to the Minimum Foundation Program and Power Equalization Program. It covers school construction and maintenance, teacher certification—the whole gamut of the common school system in Kentucky. (215)

As is true in every state, there are many in Kentucky who conclude that public schools have failed to deliver an appropriate education to its students. There has been a long series of attempts to reform the state's education system. The most obvious problem in Kentucky has been the low proportion of students who graduate from high school. Of course, if the rate of high school graduation is low, so also will be the proportion of students obtaining college degrees. Kentucky has always been one of the poorest performers in these two categories and remains so today.

Although Kentucky has never lacked for ambitious plans for improving education, it has always lacked the political will and money to implement them. Judge Corns' ruling changed all of that. Some of what would later become KERA was already in the form of proposed legislation, including such provisions as higher standards, alternative assessment, and cash incentives. This legislation passed in the state senate but failed in the house largely because of the \$75 million in costs and skepticism about whether an acceptable assessment could be found. With the groundwork completed for the most part, the state supreme court decisions ensured the implementation of these programs. Ironically, while the legislature balked at a \$75 million price tag, KERA has now already cost billions of dollars. The cost of the Kentucky Instructional Results Information System (KIRIS) alone, for the 1995–96 school year, has been estimated by Lawrence Picus² to be somewhere between \$120 million and \$254 million.

With the requirement in Judge Corns's rulings mandating a complete rewriting of all statutes referring to Kentucky's schools, the governor and legislature had an unprecedented opportunity to improve the quality of the state's schools. The question that had to be faced was how to go about accomplishing this goal. What they did not do was seek a legislative consensus for the

many parts of KERA. This caused problems at the time, and it continues to plague the system today.

According to Paul Blanchard, a political science professor at Eastern Kentucky University, in passing the KERA legislation, the representatives of Governor Wilkinson and the legislative leaders followed a nondeliberative process.³ Decisions were made privately and anyone who called for public debate soon found himself or herself suppressed. The legislation was passed with a large number of road improvements and pet projects for legislators included. Although one of the stated purposes of KERA was the elimination of patronage, it was passed with a series of arm twists, threats, and rewards for compliant legislators. The result was a quick passage, but little real commitment of legislators to the educational goals of KERA other than the political support that their party leaders demanded.

David Hornbeck, Governor Wallace Wilkinson, his assistant Jack Foster, and the Democratic leadership of the house and senate created KERA. This group became the membership of the Task Force on Education Reform appointed by the general assembly in July 1989. The final report of the task force was adopted on March 7, 1990. This final report became the Kentucky Education Reform Act (KRS 158.6451).

KERA makes sweeping changes in the state's educational structure. The justification for these changes was the belief that Kentucky languished near the bottom of all states in most categories of educational performance. For example, the task force cited the fact that the percentage of Kentucky citizens with a bachelor's degree was among the lowest of any state. In 1998, Kentucky still ranked forty-eighth in this category, hardly an improvement. Kentucky's ACT scores remained flat throughout the 1990s during the period of KERA implementation. In the meantime, the national averages went up slightly. In 2002, the average ACT scores in Kentucky declined. This pushed Kentucky's performance even further below the national average than it was in 1992.

In most categories, National Assessment of Educational Progress (NAEP) scores in Kentucky have increased slightly, but not as much

as other states, and as a result, Kentucky has fallen even further behind other states since 1992. The only bright spot in Kentucky's NAEP scores comes from an increase in 1998 fourth-grade reading scores. Although this increase has been repeatedly cited as an indication of the success of KERA, these increases can more reasonably be attributed to a higher rate of exemptions from the NAEP assessment. Students with individual educational plans (IEPs) that contain restrictive accommodations for testing are not allowed to participate in the NAEP assessment. Because the Kentucky assessment focuses only on school accountability, it has extremely generous rules for accommodations. For this reason, school principals try to maximize the number of students eligible for special education services. This classification makes them eligible for accommodations that lead to higher scores. This also makes them ineligible for participation in the NAEP assessment. In 1998, 10 percent of students were excluded from NAEP participation. Only 4 percent were exempted during the previous testing. The increased number of these exemptions best explains the increases in fourth-grade reading performance between 1994 and 1998.

The implementation of KERA led to changes in nearly every facet of the Kentucky educational system. Numerous additional programs were mandated by legislation, many at great expense. These included a more equitable distribution of funds for school districts, the requirement that the first three years of primary school not be differentiated by grades, the implementation of school-based decision making, expanded preschool programs, a reorganized department of education, extended school services, and several others. The most expensive of these innovations and the one that has had the greatest impact on instruction in the classroom is the accountability system, formerly called KIRIS (the Kentucky Instructional Results Information System), now renamed the Commonwealth Assessment Test System (CATS).

The Kentucky Instructional Results Information System (KIRIS)

KIRIS was developed in the early 1990s, at a time when there was considerable excitement about new ways of assessing stu-

dents or new ways of using old techniques. These new ideas chiefly involved the replacement of multiple-choice items with performance assessments (also called authentic testing). The original legislation required that KIRIS be entirely performance-based by 1996. For the purposes of the legislation, portfolios were considered a performance assessment. The Kentucky Department of Education (KDE) and the legislators who were supporting this assessment methodology understood that it would not be practical to make the initial forms of the test entirely performance-based because the appropriate testing technology had not been developed. The earliest version of the school index was based on constructed response items, portfolios, and Performance Events. The Performance Events were an initial attempt at performance assessment, and they were expected to eventually replace the constructed response items. The school accountability indexes also included a nonacademic score based on dropout and attendance statistics. Multiple-choice items were administered and scored, but they were not included in the school indexes. Because they were not being used, the KDE eventually stopped administering them in the mid-1990s, only to be forced by the legislature to reintroduce them in 1998. The math portfolios and the Performance Events were eliminated in 1996.

Legislative Changes in Kentucky's Accountability System in 1998

The Kentucky General Assembly met in the spring of 1998 amidst expectations that they were going to make major alterations in or actually eliminate the Kentucky accountability system. Teachers across the state, who were in a position to be keenly aware of the deficiencies in the assessment, supported legislation that would have eliminated KIRIS. The fight started in the senate, and its members, particularly those on the Education Committee, were inundated with telephone calls, most of them from teachers, urging them to fix or eliminate the system. The committee reported out a bill (SB 243) to the full senate that would have greatly scaled back KIRIS. It passed with just one dissenting vote. The house of representatives, on the other hand, passed a bill that demanded far

fewer changes (HB 627). In the ensuing conference committee meetings, a compromise bill (HB 53) was crafted based on HB 627 and SB 243. This bill was approved by the legislature and signed into law by the governor. HB 53 mandated only a few changes but did provide a mechanism for a more ambitious restructuring of the accountability system.

The most obvious change and one that represented a victory of style over substance was the change in the name of the accountability system, from the Kentucky Instructional Results Information System (KIRIS) to the Commonwealth Assessment Test System (CATS). The new name was unveiled in front of posters celebrating the University of Kentucky Wildcats' winning the NCAA basketball championship. The signs said simply, "Go CATS."

The authority for deciding whether the test should change was given to the Kentucky Board of Education (KBE). Board members were to be advised by three committees: the Education Assessment and Accountability Review Subcommittee (EAARS), made up of eight members of the general assembly; a School Curriculum Assessment and Accountability Council (SCAAC); and a National Technical Advisory Panel on Assessment and Accountability (NTAP). In addition to its advisory role, the EAARS was responsible for reviewing regulations. The NTAP was given the responsibility for determining whether the CATS tests were of a sufficient level of reliability and validity to permit scores to be reported on transcripts.

During acrimonious debates about whether to change, eliminate, or leave KIRIS untouched, the staff of the KDE and the KBE led the opposition to changes. They were supported by the two major state newspapers, the *Louisville Courier Journal* and the *Lexington Herald Leader*, and the Pritchard Committee, a private foundation devoted to the promotion of educational reform in Kentucky. The SCAAC was given the leadership role in determining the direction of the changes. Since it included the commissioner of Education, the chair of the state school board, and the director of the Pritchard Committee—all opponents of the movement to change KIRIS—it should come as no surprise that

this committee did not recommend many substantive changes in the KIRIS tests. The changes that have been implemented as KIRIS made the transition from KIRIS to CATS are not apparent to parents, students, and teachers. Students continue to answer constructed-response and multiple-choice questions and submit writing portfolios as they did with the previous assessment system.

The CATS assessment is administered to fourth- and fifth-grade students in elementary school, seventh- and eighth-graders in middle school, and tenth-, eleventh-, and twelfth-graders in high school. It assesses reading, mathematics, science, social studies, arts and humanities, and practical living and vocational skills. The primary item format since the inception of KIRIS has been constructed-response questions. Written expression has always been assessed with writing portfolios. Multiple-choice items, which have been extensively piloted in the past but never counted in the school index, now contribute 33 percent to each subject matter index. Furthermore, HB 53 requires that the results from the Comprehensive Test of Basic Skills (CTBS-5), a standardized achievement test, be included in the computation of the school index. The KBE and the KDE were reluctant to do this. In order to fulfill the legal requirement that CTBS scores be included, they assigned it a weight of 5 percent. A small proportion of a school's score is based on graduation, rates, and retention. This input, called the nonacademic index, is weighted 5 percent for elementary schools and 10 percent for middle and high schools. It actually adds almost nothing to the variability in schools' accountability index because almost every school already receives nearly all of the 100 points allocated to it.

Instead of being evaluated as right or wrong, students' responses to the constructed-response items and the on-demand writing prompts and their performance on portfolios are placed in one of four categories: Novice, Apprentice, Proficient, and Distinguished. These are collectively referred to as the NAPD scale. With KIRIS, the points associated with each of these categories were as follows: 0 points for Novice, 40 points for

Apprentice, 100 points for Proficient, and 140 points for Distinguished. Student scores were translated into school accountability scores that could range from zero to 133.6. It is 133.6 rather than 140 because the nonacademic score can be no higher than 100.

One of the problems that the CATS revision was intended to address was the wide range in performance encompassed by the Novice and Apprentice categories. Most students fell into one of these two categories, with few considered Proficient and almost none Distinguished. In order to provide finer discriminations, the Novice and Apprentice categories were further divided. The Novice category was divided into Medium Novice (13 points) and High Novice (26 points). The Apprentice category was divided into Low Apprentice (40 points), Medium Apprentice (60 points), and High Apprentice (80 points). Students rated Proficient are awarded 100 points as before, and a Distinguished rating was still assigned 140 points. The effect of these changes was to increase ratings on the school accountability index.

A high level of academic achievement in Kentucky is operationally defined as an average score of proficient on KIRIS, which is equivalent to an accountability score of 100. All schools were originally supposed to achieve this score by 2012. The timeline has now been extended to 2014.

Because of the vast differences in student populations, consequences were not based on the absolute level of student performance. It was felt that this would have been unfair because of differences in school populations. Some schools would have been expected to do well because of the high educational and economic level of their parents, whereas others, because of deficiencies in these same areas, could be expected to perform poorly. KIRIS was designed to correct for that tendency by rewarding or punishing schools based on improvement rather than absolute performance. This was accomplished by establishing baselines for each school and goals that had to be achieved each two years (a biennium) that were called "thresholds." There were many problems with this system. For example, mathematical errors built into its design meant schools that successfully reached their assigned threshold

every year would not come close to a score of 100. Expected performance was based on performance in the previous two years, and consequences were determined by performance in relation to that expectation. As a result, good performance in a biennium would lead to high thresholds and inevitably to poor performance in the next biennium. Conversely, poor performance in a biennium would lower the goals and make it easy for a school to be rewarded the following biennium. Schools designated as being "in crisis" were not schools with a history of poor performance. They were instead schools that performed exceptionally well in one biennium and thereby were saddled with an impossible goal for the next biennium. The schools most likely to be rewarded were those that had easy improvement goals because of bad performance the previous biennium.

With the revision of KIRIS into CATS, new thresholds are no longer recomputed every two years. Instead, a straight-line model is being employed. The average from 1999 and 2000 school years are used as a baseline. A separate chart is created for each school with a line drawn from the school's baseline score in 2000 to a score of 100 in 2014. This is the Meeting Goal line because schools at or above this line receive cash rewards at the end of each two-year period. A second line, called the Assistance line, is drawn from the baseline to a score of 80 in 2014. Schools that score below the Assistance line must undergo a Scholastic Audit. Schools with scores between these two lines are considered to be Progressing and they do not get rewards or an audit. Schools are expected to increase their performance in equal increments in each two-year biennium at a rate that will lead to their attainment of a score of 100 by 2014. As long as a school's accountability score does not fall a standard error below the Assistance line, it is considered to be making appropriate progress. The purpose of the straight-line model was to eliminate the seesaw effect that occurred when a school was highly successful one year, leading to an unachievable goal for the next year and likely unsatisfactory performance. By going to a straight-line system, they eliminated that problem but introduced a new one. The new problem is one that the old KIRIS design was intended to elim-

inate. Under the KIRIS system, schools were expected to show the greatest improvement when they were furthest from 100 and lesser improvement as they approached their goal. It was assumed that it would be easiest for a school to show large improvement when they were far from their goal and increasingly difficult as they approached the goal of 100. With CATS, the same amount of improvement is required in each two-year period regardless of a school's position in relation to 100.

When KIRIS was implemented in the early 1990s, it differed from the standards-based reform found in other states by imposing high rewards for successful schools and severe consequences for schools with poor performance. Teachers in Reward schools received cash bonuses of up to \$2,500. Unsuccessful schools were labeled as being In Crisis and they faced severe sanctions. The staff in In Crisis schools was placed on probation and a Distinguished Educator assigned to supervise reclamation of the school. These Distinguished Educators were given sweeping powers. They could order teachers to change the way they were teaching, and if the teachers did not comply, they could be fired. Although the power was rarely invoked, the mere threat had a chilling effect on staff members. Teachers were often afraid to question any suggestions made by the Distinguished Educator. The poisoned climate led to an increased rate of resignations in the affected schools. Parents of students in In Crisis schools had to be notified by mail that their school was a failure, and they were to be given the opportunity to transfer. With the change to CATS, the designation of In Crisis and the consequences to the teacher were eliminated. Although the threatened consequences were understandably unpopular with teachers and they advocated their elimination, during the eight years this rule was in effect, only nine schools were ever labeled In Crisis. Most of these schools were guilty of no more than doing too well the previous cycle and therefore being cursed with an impossible-to-achieve threshold. No teachers were ever dismissed based on the recommendation of a Distinguished Educator.

The primary effect of HB 53 on the consequences was to eviscerate them. Schools that reach their goals get reward money, and

the amount can be sizable. The school council decides how the money is to be used, and they can distribute it to the faculty if they choose. The sanctions for schools that fail to reach their goals have been greatly weakened. To be identified as a school that is below their Assistance line can certainly be embarrassing, but it is a lot better than being labeled In Crisis and maybe closed. A school performing below its Assistance line is required to undergo a Scholastic Audit to determine whether the school needs Commonwealth Improvement Funds or the assistance of a Highly Skilled Educator. It can be anticipated that most principals will conclude that the funds will be more useful than the advice of a consultant. The accountability system has been ameliorated in another way. Although schools are supposed to have reached an accountability index of 100 by 2014, the way CATS is structured, a school could be a standard error below at 80 and still be considered to be making satisfactory progress. A school could be rewarded by having its students obtain an average score of High Apprentice with none achieving a score of Proficient, the supposed goal of Kentucky's accountability plan.

Conspicuous by its absence in HB 53 is any mention of performance assessments. The KBE and the KDE finally realized that the performance assessments in past iterations of KIRIS did not work, and they were eliminated. They also recognized that labeling a constructed-response formatted test a *performance assessment* was dishonest. On the other hand, writing portfolios, which have proved to be the least reliable of any of the previously used measures, continue to be included as part of a school's accountability score.

Combining the results of multiple-choice tests with those from constructed-response items poses another challenging technical problem. Although the contractor, McGraw-Hill, has considerable experience and expertise in this area, the publisher concedes that the assumptions upon which this scaling methodology is based are wildly implausible.

One of the principal complaints about KIRIS voiced by teachers was that schools were being evaluated based on cohorts of students from different years. Since KERA was first implemented,

teachers have asserted that these cohorts can differ dramatically. Eighth-graders in one year might contain many high-achieving students, whereas in the following year the students may be much weaker. Teachers have urged the adoption of longitudinal comparisons because these differences between cohorts have made comparisons between years unfair. HB 53 mandated the use of longitudinal comparisons. This legislation also required the inclusion of credit for successful student performance in sanctioned events with an established protocol of adjudication, such as band contests. The committees given responsibility for the design of CATS were given considerable latitude in the structure of CATS, and they decided that it would be too difficult to include either band contests or longitudinal results in a school's accountability index. Information about school performance in band contests appears only on the school report card.

Restandardizing CATS

Setting cut-points, the score that differentiates between passing and failing, is the most difficult aspect of standards-based assessment. The task becomes more difficult as the number of cut-points that must be set increases. With KIRIS, distinctions had to be made among the four categories of Novice, Apprentice, Proficient, and Distinguished. In their training, scorers are given verbal descriptions and examples of what the various standards are supposed to mean, and they make the ultimate decisions about each student response. The number of distinctions increased with CATS because different levels were designated within the Novice and Apprentice categories. Efforts are made to "moderate" the grader's standards for scoring student answers to make them more consistent. The scores are also adjusted statistically to correct for differences in difficulty among items.

In spring 2001, the KBE decided to change the way the grading standards for CATS were set. The scoring system had always used absolute standards because KERA demanded a high level of academic achievement for all Kentucky students regardless of how the typical students were performing at the time. It had always been recognized that Kentucky students had a long way to

go before they reached the high levels expected of them, but KERA was supposed to bring students up to the point that they were Proficient by 2014.

During the first six years of KIRIS, the scores of elementary schools increased about 15 points, middle schools about 6, and high schools about 11. As shown in Table 8.1, by 1998, this brought elementary schools up to a little less than 50, high schools to a point a little above 50 and middle schools to 44.

An examination of the results of KIRIS over the years revealed that the improvement that had been achieved in the first six years resulted primarily from having students move from the Novice to the Apprentice category, not to the Proficient category, the stated goal of KERA. Much of the improvement from Novice to Apprentice came from inducing students who were leaving their answer sheets blank and being labeled Novice, to write something down. Graders are allotted only seconds to grade each answer, so filling up a page, even if the content was not very good, could bring some of these Novice scores up to the Apprentice level. The most dramatic change in Table 8.1 is the large increase in performance in 1999 and 2000. These increases do not represent any real improvement in student performance. They are instead the result of changes in scoring that occurred with the introduction of CATS.

TABLE 8.1 Average KIRIS Accountability Scores Across Levels

<i>Year</i>	<i>Elementary</i>	<i>Middle School</i>	<i>High School</i>
1992	33.4	37.5	40.0
1993	35.7	37.4	34.9
1994	40.9	41.8	43.3
1995	47.1	44.5	44.6
1996	45.2	41.0	43.3
1997	49.0	45.6	50.4
1998	48.8	43.9	51.3
1999	60.0	50.0	60.0
2000	61.0	51.0	61.0

Although the modified scoring associated with CATS provided an initial boost, it was not sufficient to bring students up to the KIRIS goal of 100 points or even to the 80 points that were defined as the goal under CATS. It had become obvious that having students in a school average 80 points was unrealistic. The solution was to turn CATS into a norm-referenced assessment. With the standards-based system upon which the KERA assessment was originally based, students are not compared with one another; they must reach an externally established standard. There is no guarantee that students will reach the desired high level of performance. Norm-referenced scaling presumes a normal distribution of student achievement in which half of all students will be above and half below the mean. The adoption of norm-referenced scaling made it possible to redefine Proficient as statistically average. The statistical goals of KERA are much more easily achieved using this system, and furthermore, no actual increase in student achievement is required.

To use norm-referenced interpretations of test performance properly, test items need to have a level of difficulty that ensures that student scores are spread across the distribution with some students at the top and some at the bottom. CATS is a difficult test on which almost all students are in the lower half of the possible distribution of scores. In some subject matter areas and grade levels, only a few students perform above the Apprentice level. If norm-referenced assessment was to be adopted, the difficulty of the items should have been adjusted so that student performance would be closer to a normal distribution. The changes in scaling were done behind the scenes with little public discussion, and it is unlikely that many educators, much less the public, understand the scaling issues. Making the difficulty of the items appropriate to the scale would have signaled an important change in CATS, and this is something that the KDE and the KBE were not anxious to do. As a result, a norm-referenced scaling scheme has been imposed on a very difficult test intended to measure absolute standards. The result is an assessment that has far worse psychometric characteristics than it would have had it been properly constructed. It is also an

assessment for which a student's answer can be considered Distinguished even if it is incorrect.

Changing the standards on an accountability system employing constructed responses graded on a four-point scale (a polytomously scored test) is much more complex than doing so with a multiple-choice-based system in which items are scored as either right or wrong (dichotomously scored). With this latter system, to make it easier for students to pass without changing the items, all that needs to be done is set the cut-scores lower. With a polytomously scored, predominantly constructed-response test such as CATS, the definitions of Novice, Apprentice, Proficient, and Distinguished need to be changed to make them more easily attainable.

When the KBE decided to set new performance standards for CATS, they had to select a method. The Angoff method is the most widely used. It is implemented by having judges examine each item and decide what the probability is that a minimally competent student would get the item correct. It sets absolute standards, and for this reason, it would not solve the problem that the KBE needed to solve, which was that Kentucky students did not show improvement when compared with these sorts of standards. Furthermore, the Angoff method is appropriate for use only with dichotomously scored tests.

The KBE decided to pilot-test three methods of standard setting. The resulting study was a large-scale, complex, expensive study. It was their intention to select the best method or combination of methods from among the three. The CTB Bookmark, the Jaeger-Mills, and Contrasting Groups methods were compared. The CTB Bookmark method developed by McGraw-Hill, the contractor for the Kentucky assessments, is the most widely used of the three and has the advantage of being appropriate for use with both dichotomous and polytomous items. To implement the CTB Bookmark method, the contractor placed all of the items, including both constructed-response and multiple-choice, on a continuum, according to their difficulty. Each constructed-response item appeared four times to represent the four types of responses that corresponded with the NAPD scale. Judges,

teachers, and other educators were asked to decide where on the continuum the cut-point between each NAPD level should be made. With the Jaeger-Mills method, judges reviewed each student response (content/grade specific) associated with a spring 2000 scale score. The median scale score of the responses judged to be at the dividing point between each level was used to determine each cut-point. The Contrasting Groups method used teachers to identify students they considered to be in each of the NAPD levels. The student's actual performance was then related back to these teacher appraisals. The result from each of the three methods was given to a committee established to sift through the results and provide the KBE with a number of options for setting the cut-scores. The CTB Bookmark method worked best and seemed to comport with what the panel expected. The Jaeger-Mills method was awkward and difficult to interpret, and the Contrasting Groups method yielded wildly high scores.

Although the panel was only supposed to provide a set of options to the KBE from which they were to choose, they went ahead and actually set the standards. Their recommendations were allowed to stand, and the standards were applied to the 2000 results and disseminated. This proved embarrassing because the KDE had already posted the results of the 1999 and 2000 administration using the original scaling methods. The contrasts were dramatic. This release of results undercut the claims of the KDE and the KBE that the new standards were adopted because the old standards could not be applied to CATS.

Table 8.2 shows the percentage of students in either the Proficient or the Distinguished category for each of the subject areas assessed in the elementary schools. Table 8.3 shows the same for middle school students, and Table 8.4 for high school students. As can be seen, there was no real progress in moving students into the higher categories in 1999 and 2000 when the old standards were used. When these new standards were applied retrospectively to the 2000 data, in some subjects there are only modest changes, whereas in others the differences are dramatic.

At one time, it seemed unlikely that Kentucky schools would ever reach the goal of 100 by 2012 or even the more modest,

TABLE 8.2 Percentage of Elementary Students in the Proficient or Distinguished Categories

<i>Content Areas</i>	1993	1994	1995	1996	1997	1998	1999	2000	2000*
Reading	8	11	30	31	41	33	32	32	57.0
Math			18	14	19	20	21	25	34.0
Science	2	2	5	3	6	6	5	5	37.0
Social Studies	8	13	18	13	13	15	13	14	42.0
Arts and Humanities	1	1	1	2	3	3	5	5	15.6
Practical Living and Vocational Studies	2	3	3	3	4	6	6	7	50.0
Writing on Demand			3	3	3	5	2	5	
Writing Portfolio			16	13	16	17	22	23	

*Percentage of students in Proficient and Distinguished categories using the new revised standards.

TABLE 8.3 Percentage of Middle School Students in the Proficient or Distinguished Categories

<i>Content Areas</i>	1993	1994	1995	1996	1997	1998	1999	2000	2000*
Reading	11	19	13	13	18	15	13	12	53
Math			30	28	34	29	33	37	28
Science	2	1	2	1	1	1	1	1	37
Social Studies	10	16	21	13	15	12	10	12	31
Arts and Humanities	6	8	7	6	8	6	7	8	37
Practical Living and Vocational Studies	4	5	5	4	6	7	8	7	40
Writing on Demand			4	2	4	7	6	8	
Writing Portfolio			15	11	14	13	10	11	

*Percentage of students in Proficient and Distinguished categories using the new revised standards.

TABLE 8.4 Percentage of High School Students in the Proficient or Distinguished Categories

<i>Content Areas</i>	1993	1994	1995	1996	1997	1998	1999	2000	2000*
Reading	4	12	11	9	32	28	29	33	30
Math			16	23	28	27	33	33	29
Science	4	9	12	10	14	13	12	14	28
Social Studies	6	19	19	13	23	29	30	31	26
Arts and Humanities	2	4	1	2	1	4	4	5	21
Practical Living and Vocational Studies	2	4	4	2	7	6	7	7	53
Writing on Demand			2	1	5	23	9	13	
Writing Portfolio			19	20	21	22	24	24	

*Percentage of students in Proficient and Distinguished categories using the new revised standards.

redefined goal of 80 by 2014. Although there has been no real change in student achievement, by adding additional levels within Novice and Apprentice and adopting norm-referenced scaling, these goals are now attainable, at least for some schools.

Philosophy

In order to judge the effectiveness of educational reform in Kentucky, it is necessary to consider its purpose. KIRIS/CATS is not based on a consistent philosophy. It is instead the product of several conflicting philosophies. These philosophical disagreements reflect the deep divisions among educators across the nation.

At the same time that Kentucky courts were demanding changes in the state's educational system to make them more financially equitable, the business community began pressuring the governor and legislature to do something about the poor quality of the job applicants they were encountering. To politicians and business leaders, the solution to this problem could be found in the use of accountability based on testing. They were unaware of the controversies surrounding different instructional philosophies. The staff of the KDE had a deep commitment to the principles of progressive education, and they were willing to strike a bargain with the political and business advocates of education reform. They would implement a form of assessment to be used for accountability as long as they controlled the format of the assessment and the type of instruction that would be supported. Business and political leaders know little about instructional methods and are willing to accept the proposition that old and ineffective methods are actually new and promising. Student-centered instructional methods were not widely employed in Kentucky before the implementation of KERA, and it is not easy to see how progressive education beliefs can be made compatible with standards-based reform. It is unfair to require teachers to implement methods that are known to be ineffective in increasing student achievement, and then evaluate the teachers based on their students' performance.

Historically, the purpose of instruction in this country has been increasing student academic achievement. This is not the purpose

of progressive education, which prefers to be judged by standards other than student academic performance. The Kentucky reform presents a paradox, a system structured to require increasing levels of academic performance while supporting a set of instructional methods that are hostile to the idea of increased academic performance.

Evaluating the Technical Qualities of KIRIS

Describing the technical characteristics of the Kentucky assessment systems in a compendious form is quite difficult. First, both KIRIS and CATS are incredibly complex. They have many parts, some of which have received almost no publicity. There are aspects of these systems about which only a handful of people in the state have any knowledge. Some of these aspects play a critical role in determining which schools are labeled successful and which are to be called failures. In designing this assessment program, numerous implementation decisions had to be made, and the wisdom of each of these requires consideration. A complete elaboration of these issues would require at least a book and perhaps more than one volume.

When the system was being designed in 1990, the Kentucky Department of Education and Advanced Systems for Measurement and Evaluation (ASME) were under enormous pressure to have the system up and running quickly. They had to assemble the system expeditiously without the luxury of time to consider alternatives. Their decisions were strongly influenced by distrust for conventional measurement doctrine, and they were under pressure to embrace the popular alternative assessment techniques of the day. They also believed that the first priority of the system was improving instruction. Its technical qualities were relegated to a secondary role.

Formal criticisms of the technical qualities of KIRIS have been documented in four reports, each written by nationally recognized experts with impeccable credentials. Each concluded that there were serious problems with KIRIS. No formal studies of the technical qualities of CATS have been published as of yet.

The first report was released on February 16, 1995. It was conducted under the auspices of the Kentucky Institute for Educational Research (KIER), which contracted with the Evaluation Center at Western Michigan University in Kalamazoo, Michigan.⁴ The report is titled *Evaluation of the Development and Implementation of KIRIS Through December 1994*. The director of the Evaluation Center was Daniel Stufflebeam, one of the nation's leading experts on evaluation. This report is commonly known as the *KIER Report*. The second report was released on June 20, 1995, and was prepared for the Office of Educational Accountability (OEA) of the Kentucky General Assembly. This report is referred to as the *OEA Report*.⁵ The panel that prepared the report included Ronald Hambleton from the University of Massachusetts, Richard Jaeger from the University of North Carolina at Greensboro, Daniel Koretz from the Urban Institute, Robert Linn from the University of Colorado at Boulder, Jason Millman from Cornell University, and Susan Phillips from Michigan State University. The *OEA Report* concluded that KIRIS was so fatally flawed that it could not legitimately be used for making any decisions. The third report is called the *Catterall Report* and is titled *Kentucky Instructional Results Information System: A Technical Review*.⁶ The Legislative Review Committee of the Kentucky legislature commissioned this report. It was intended to provide a basis for legislation that was expected to implement changes in KIRIS during the 1998 legislative session. The *Catterall Report* echoes the serious concerns voiced in the previous two reports. A fourth report written by Daniel M. Koretz and Sheila I. Barron was published in the fall of 1998. It focuses on the validity of the KIRIS accountability scores. It was produced by RAND and is titled *The Validity of Gains in Scores on the Kentucky Instructional Results Information System*.⁷

Reliability

The computation of the reliability of conventional achievement and aptitude tests is relatively simple, and most of the technical manuals that accompany standardized tests are overflowing with

these coefficients. An examination of Buros's *Mental Measurement Handbooks*, which provides technical information and critical reviews of all major aptitude and achievement tests, includes few that have weaknesses in this area. Reliability is the *sine qua non* of test construction, and it is not difficult to create highly reliable tests.

Whereas reliability is usually established for the scores of individual students, for the KIRIS/CATS assessment, it must be based on school scores. It is much more difficult to establish reliability for school scores than for individual scores. A test is reliable to the extent that it is characterized by only small amounts of error. There is a multiplicity of different sources of error variance associated with Kentucky's tests. Not only is there variability in students across items, but there also is variability in students within schools. There is also the error associated with the use of graders to evaluate student responses and error caused by the use of the twelve different test forms used with KIRIS and the six used with CATS.

Neither the KDE nor ASME has ever published properly computed reliability coefficients for either the individual student scores or the accountability index. The public has been told that the reliability of the school scores was acceptable but that the reliability of individual scores was not. For this reason, decisions about individual students cannot be made based on these scores. Reliability coefficients for the Accountability Scores assigned to schools have been reported, but they are based on an incorrect application of generalizability theory, which has resulted in inflated coefficients.⁸

The Reliability of Change Scores

The reliability of the accountability indexes would be important only if schools were evaluated based on the magnitude of these scores. With KIRIS/CATS, it is the difference between a school's accountability index and past scores that is used to assess schools, and it is the reliability of these differences that must be established. The difference between the two is called a "change score," and it is axiomatic in measurement that the reliability of change

scores will always be lower than the reliability of the two scores upon which they are based. The KDE has always avoided any mention of problems with the reliability of change scores in their technical manual and other publications. Officials from the KDE have denied the significance of problems surrounding the reliability of change scores in several newspaper articles. The *OEA Report*⁹ cites the *Standards for educational and psychological testing*¹⁰ (1985) to confirm that the reliability of the KIRIS change scores can be expected to be lower than the reliability of the accountability indexes themselves.

Validity

The accountability scores from the KIRIS assessment have shown some increase over the years since its implementation. The average scores for each year across the three levels are provided in Table 8.1. The critical validity issue is whether these increases represent real improvement in academic achievement or reflect other factors, such as teachers preparing students for specific items, changes in test difficulty, or enhanced test-taking skills. These issues are addressed in all four of the external evaluations of KIRIS, and in each, it was concluded that the preponderance of evidence indicates that the increases do not reflect real improvement in overall academic achievement. Each year when the scores have been reported, the Kentucky Department of Education has announced with great fanfare that they provide proof that the reforms embodied in KERA have been successful in increasing student achievement. There is no evidence, other than testimonials and anecdotal reports, to support this position. The *OEA Report* responded to these periodic KDE announcements with the following analysis:

. . . the reported gains in scores on KIRIS substantially overstate improvements in student achievement. Indeed, it is not clear whether any appreciable generalizable gains in achievement have been produced in some grades and subjects. The external evidence to which KIRIS scores can be compared fails to reflect the gains shown on KIRIS.¹¹

The RAND report focuses on the question of whether the increases in the accountability scores represent improvement in

overall student achievement.¹² RAND researchers compare student performance on KIRIS with reading and math performance on the NAEP test and eleventh- and twelfth-grade performance on the ACT test. They conclude that the increases in KIRIS accountability scores are not reflected in similar increases in scores on these external tests. In addition to the external evidence they cite, they collected internal evidence that they believe establishes that students perform better on reused items than on new items. They interpret this to mean that teachers are focusing their instruction on improving student performance on specific items rather than improving overall student knowledge. What they fail to emphasize is the degree to which increases in student performance across the years are mainly the result of three factors: (1) success in getting students to move from the Novice to the Apprentice levels; (2) the increase in the number of points awarded at the Novice and Apprentice levels; and (3) the rescaling of CATS.

The RAND report makes a number of suggestions for preventing the erosion of validity that RAND researchers have identified as having occurred with KIRIS. RAND researchers cite the likelihood that teachers and educators will do everything in their power to obtain higher scores when goals are overly ambitious and high stakes are based on student performance. They acknowledge that much of the inflation in state scores is unavoidable.

Lessons That Have Been Learned from Kentucky's Attempts to Establish Its Accountability System

All but one state has adopted statewide content standards and implemented an assessment program to determine whether students are achieving these goals. Kentucky was one of the first states to initiate a high-stakes accountability system, and from the beginning it showed a willingness to commit enormous resources to the reforming of its educational system. Some of the lessons that other states may be able to learn from Kentucky's experiences are described here.

I. Reexamine claims that all students can perform at the same high level.

An underlying assumption of KERA is that all students, including special education students with IEPs, can perform at the same high level. Other state reform programs make a similar claim. This marks a fundamental difference between traditionalism and progressive education. In the early twentieth century, when progressive education first gained influence, one of its major tenets was that students differed in their academic ability and could not all benefit from the traditional curriculum that was then in place. Progressives promoted the use of standardized tests to classify and track students. Traditionalists opposed the use of these tests and the tracking of students. They asserted that all students could learn at the same high level, which provides a contemporary justification for adopting the same high standards for all students.

To fully understand the policy, it is first necessary to parse the phrase “all students can learn at the same high level.” If this phrase is intended to mean that all students can answer high-level questions equally well or all students can learn at the same rate, this assertion is false. Responding to high-level questions is a function of intelligence, and students differ in their possession of that trait. Students who are more intelligent also learn more quickly. This does not mean that all or at least most students cannot learn the same content, with the caveat that some academic content is beyond the ability of students at the low end of the intellectual continuum.

Kentucky and other states have gotten into trouble when they have gone too far with their assertion that all students can function at the same high level. Kentucky even requires this high level of performance from students diagnosed as requiring special education. States need to be realistic in their demands on students. Certainly, some states have overdone their tendency to excuse poor performance and give up on students too early. On the other hand, an assessment should not be structured in such a way that it repeatedly places students in circumstances where their failure is guaranteed.

2. Do not base state educational reform solely on the assessment of high-level thinking skills.

Many states, including Kentucky, have adopted policies that require their assessment to emphasize high-level thinking skills rather than mastery of the subject matter. This occurs for three reasons: (1) the content standards of many states are not sufficiently detailed to permit the measurement of actual achievement, (2) publishers prefer to include items assessing high-level thinking skills on standardized achievement tests, and (3) progressive educators prize process over content.

Quality of State Content Standards

Constructing high-quality content standards is difficult and expensive, and many states fall far short of the ideal. If a state's content standards are vague, cover only a select number of years, or do not provide adequate specificity, it will be difficult for the test publisher, with whom the state has contracted, to create a test that assesses content. Instead, test publishers can create the illusion of content validity by writing items that use content to assess high-level thinking skills. A bright student will be able to correctly answer the questions using reasoning skills even if he or she does not know the content well, but a student who does know the content and who is not blessed with great reasoning ability will get them wrong. The result is a test that appears to have content validity, but does not have construct validity because it is measuring higher-level thinking skills rather than achievement.

Test Construction Methods

Test publishers often rely on items that assess high-level thinking skills when faced with inadequate content standards as described above. There is a second reason why publishers like to include this sort of item on their tests. The psychometrists that create these tests generally believe that the most important characteristic of a test is internal consistency reliability as measured by Coefficient Alpha. They focus on this type of reliability because it is easy to compute and widely accepted, and tests with this quality are easily

constructed. An examination of almost any technical manual for a standardized test will reveal that most of its pages are devoted to reporting Coefficient Alpha coefficients.

The highest Coefficient Alpha values are obtained when a single, internally consistent construct is measured. These conditions prevail for reading comprehension tests, for example. Other content, such as science and social studies, assessed on state tests is more multidimensional and tends to yield lower reliability coefficients. Since higher-level thinking skills represent a unitary construct, a test that includes many such items will be internally consistent. A social studies test constructed to focus on higher-level thinking will be more reliable than one that measures the mastery of social studies content. Tests also may end up with a lot of items measuring high-level thinking skills even if this was not the test publisher's intent. When the publisher conducts item analyses of its pilot tests, they use procedures that select items based on how much they contribute to reliability. Items measuring high-level thinking do this better than those that measure content and are therefore more likely to survive the item analysis process.

Progressive Education Philosophy

Progressive education philosophy tends to reject conventional assessment methods. When an assessment must be implemented, progressives tend to favor the assessment of higher-level thinking skills rather than requiring students to memorize or learn facts. They recognize that students with lower ability struggle with conventional tests, and they think tests that do not measure content and instead measure higher-level thinking skills will be fairer. In actuality, lower-ability students have even more trouble with such tests.

KERA and Higher-Level Thinking

The Kentucky Education Reform Act is predicated on the belief that all students and therefore all schools can perform at the same high level. This is not something that its creators wished were true; it is what they sincerely believed was true. This belief is

incorporated into the structure of the assessment system in the strongest possible way. It is manifested in the original goal of having all schools achieve an accountability score of 100 by 2012 and the revised goal of 80 points by 2014.

The items used with KIRIS and now CATS are primarily constructed-response items. The items are quite difficult and generally require the application of higher-level thinking skills. In many cases, the items are far too difficult for students to even begin to make a response. This leads to a restricted range in student performance and a consequent diminution in reliability.

The futility of such an unrealistically high goal becomes evident when the released items from the test are examined. Some of the eleventh- and twelfth-grade items are so hard that students in graduate school would have a difficult time responding correctly to them. The idea that all high school students, including those who have been identified as needing special education accommodations, can eventually be brought up to a level of functioning where they can successfully answer these questions seems overly optimistic.

There is a strange disconnection between the designers and supporters of accountability systems such as those that have been implemented in Kentucky and the mainstream of cognitive science. In a letter to the editor that appeared in the *Louisville Courier* newspaper, Wilmer Cody, the former (1998) Commissioner of the Kentucky Department of Education, articulates the underlying assumptions of KERA as follows:

Finally, a large body of research demonstrates that the most important factors governing how well children do in school have nothing to do with perceived differences in individual potential. On the contrary, children do well because of instructional leadership, a clear focus on academic achievement, high expectations, the quality of professional development, curriculum alignment, teacher skill, the effective use of learning time, and parental involvement.¹³

Although some of these factors are important and may play a role in how much students learn, it would be difficult to locate competent research that would demonstrate that these factors could overcome a student's lack of academic potential. There is

extensive scientific evidence that contradicts the assertions made by Wilmer Cody.

3. Clearly delineate the content to be covered.

If teachers are to be held accountable for what their students have learned, they need to be given a clear description of exactly what students are supposed to learn. The logical sequence would be to adopt state content standards before developing the tests. Ideally, states need to go even further and specify the level at which students are to master the content. These content and performance standards provide the basis for a state's accountability system. In Kentucky, the process was reversed. The tests were developed first, and only later was the content defined.

When the KERA legislation was first implemented, six Learner Goals were specified. In addition, the KDE was supposed to create standards, which would identify what students should learn and determine the content to be included on the tests. At the same time, there was a sense of urgency about the need to get the testing program started. As a result, test items were written before standards were established. Since the program's inception, the standards have been chasing the tests. There have been six separate sets of standards published. These standards, in the order of their release, are the six Learner Goals, the Learner Outcomes, the Transformations, the Academic Expectations, the Content Guidelines, and the Core Content for Assessment. Each of these was intended to be the final word on what students were supposed to learn and the basis for the KIRIS assessment. These standards differ among themselves in terms of content and philosophy and provide minimal guidance for teachers endeavoring to prepare students for the KIRIS/CATS assessment. Throughout the implementation of KERA, the high standards implicit in the assessment have remained fixed in the test itself, but have never been clearly delineated in the published standards.

Kentucky assesses students three times in each subject, once in elementary school, once in middle school, and once in high school. Because of this schedule, the KDE decided to define only three sets of content standards. There is much that students need

to learn in the grades between those being assessed, so the test authors are forced to make assumptions about what students should have learned in the years between those being assessed. They also must assume that students have mastered the lower-level skills, which is difficult to do if they are not spelled out. A more precise assessment of students will arise from a test that includes items that cover a range of difficulty, and this is hard to do if the content standards are restricted to just one year. Although it is a lot of trouble, states need to have content standards for every year. This allows them to construct properly sequenced tests.

The Kentucky content standards have another problem, one that is shared by many other states, particularly states with comprehensive and demanding content standards. The usual method of creating content standards is to put content specialists in charge. For example, the committee responsible for writing the math content standards might include specialists in math from the state department of education, math teachers from the public schools, school of education faculty, and math professors from arts and sciences units at state colleges and universities. All of these specialists share one important idea: A love of math, a recognition of its importance, and the belief that every educated person needs know a lot about this subject. As a result, they create content standards in math that are ambitious and require every student to be able to perform at a high level in math. Implicit in these content standards is the assertion that a considerable part of the school day should be devoted to instruction in this topic. The content standards for social studies, science, and arts and humanities are assembled with similar committees, as are the standards for the other subject areas. The authors of content standards in these content areas are no less committed to the field in which they specialize and the importance of its coverage in depth than those whose task it is to write the math standards. They would never agree that only a small proportion of the educational day should be devoted to their field in order to leave more time for math, science, or whatever. What this means, in practice, is that the content standards for each area are written in isolation, with each committee ensuring comprehensive

coverage of their field. What is ignored are the limits imposed by the length of the school day. There is simply not enough time to include everything that these different specialists believe is important. This problem is exacerbated in Kentucky by the inclusion of arts and humanities and practical living/vocational studies in addition to the usual reading, math, science, and social studies. The leadership needed to tone down the requirements of some fields in order to have a reasonable range of content is just not there. The breadth of content listed in Kentucky's Core Content for Assessment, like the content standards of many other states, is so broad that it could never be covered.

4. Ensure that the instructional philosophy promoted by the department matches the instructional philosophy that underlies the state's educational reform.

The Kentucky Education Reform Act (KERA) was presented to the legislature as a conventional standards-based education reform, similar to the programs now adopted in forty-nine of fifty states. It was intended to define what students needed to know, to assess them to see if they had learned this content, and to hold principals and teachers accountable for their performance. The governor, the legislative leaders, and business leaders assumed that such a system would force students and teachers to work harder, which would lead to an improvement in student academic achievement. They strongly believed that a good school was one in which students demonstrated a high level of academic achievement. What they did not realize was that by placing control of Kentucky's educational reform in the hands of the KDE, they were empowering an administrative body infused with a progressive education philosophy that was the antithesis of the traditionalist philosophy of KERA.

Although a rejection of academic achievement as a criterion for judging the worth of schools may seem strange to noneducators, most of the established national organizations devoted to teacher training likewise reject academic achievement as an unworthy goal for public schools. For example, the state of Kentucky, like many other states, requires that teachers graduate

from an education school that is accredited by the National Council for Accreditation of Teacher Education (NCATE). A careful reading of the NCATE standards and the voluminous supporting materials that accompany them will fail to identify any commitment to student achievement.

Kentucky has invested an enormous amount of money, along with teacher and student time, in a system that is devoted to increasing student achievement by imposing high-stakes standards. At the same time, it has handed control over the system to a department of education that is committed to the belief that academic achievement is not important. One of the most controversial aspects of the original KERA legislation was the requirement that the first three years of schools be ungraded. This meant that students who in most school systems would be separated into first, second, and third grade had to be grouped together in a primary class. This requirement was controversial and unpopular, and it was eventually modified to the point that it became an option rather than a mandate. Robert Slavin, in a review of the literature on nongraded elementary schools, co-authored with Roberto Gutterrez (1992), found inconsistencies in the achievement of students in nongraded classes.¹⁴ They explained the inconsistency by attributing it to differences in instructional methods employed in successful and unsuccessful programs. Successful ungraded classes coupled direct instruction with effective methods of cross-age grouping. Those that were unsuccessful used student-centered instruction rather than direct instruction. The review asserts, "Individualized instruction, learning stations, learning activity packets, and other individualized or small group activities reduce direct instruction time with little corresponding increase in appropriateness of instruction to individual needs" (369). It also states that "to the degree that nongraded elementary schools came to resemble the open school, the research finding few achievement benefits to this approach takes on increased importance" (368). In the studies he reviewed, direct instruction methods were consistently favored over progressive education approaches. At about the same time that this review of literature was published, the Kentucky Department of Education

distributed a manual to all elementary school principals titled *Best Practices in Ungraded Classrooms*. In this document, they listed what schools should and should not do when implementing the ungraded primary program. They urged schools to employ the instructional practices that Slavin found were ineffective and cautioned against adopting the practices that he recommended.

5. Avoid use of the term “criterion-referenced.”

Criterion-referenced assessment (CRT) was first proposed for use with mastery learning by Robert Glaser.¹⁵ At the time, mastery learning was considered a wonderful new instructional technique that would revolutionize education in the United States. This instructional method required that everything students were to learn be defined with behavioral objectives. Students were to learn these objectives in an empirically derived optimum sequence. Student progress was to be reported in terms of a listing of objectives mastered, an approach he labeled “criterion-referenced” testing. James Popham is credited with popularizing the term during this time.

It turned out that CRT was not easily adapted for use in schools. One problem, among many, was that defining instructional goals using behavioral objectives was difficult for any but the most concrete content such as math. Although the use of CRTs to explicitly describe what a student has learned declined, use of the term “criterion-referenced testing” continued. Its use continued because the public tends to view CRTs in a positive light while remaining skeptical about norm-referenced tests (NRTs). Apparently, CRT does well in focus groups. Politicians, departments of education, and publishers like to describe their assessment as “criterion-referenced testing” even though what they are doing bears no relationship to what Glaser and Popham meant by the term. The inconsistency between existing practice and the original definition of criterion-referenced testing was resolved by changing the definition. The ultimate authority regarding correct definitions in measurement is the *Standards for Educational and Psychological Testing (Joint Standards)* published

by the American Psychological Association, the NCME, and the AERA.¹⁶ Here is how a criterion-referenced test is defined in that publication: "A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparisons to cut scores" (174). What this definition purports to do is contrast CRTs with NRTs, but according to this definition, any test that sets a cut-score can be considered a criterion-referenced test. Since cut-scores can be appended to any test, an NRT is easily turned into a CRT. As a result, the distinction between a CRT and an NRT has been blurred. It is probably best at this point to view the use of the term "criterion-referenced" as a public relations strategy rather than a functional definition of an assessment procedure. Almost all state testing programs are called criterion-referenced, regardless of their format.

6. Use norm-referenced measurement for school comparisons and standards-based assessment for students.

In setting up a reform program, states have two choices. They can use a norm-referenced or a standards-based approach. An evaluation of student performance based on a determination of how students compare with each other is referred to as norm-referenced assessment. Norm-referenced scales have a characteristic bell-shaped curve and are calibrated using means and standard deviations. Each point on a norm-referenced scale can be associated with a fixed proportion of test takers. Cut-scores designate the percentage of students who will pass and the percentage who will fail. Points on a scale that are defined in terms of percentages of students are called percentiles. The main disadvantage to this approach is that passing or failing is unrelated to absolute levels of performance. No matter how well or how poorly the overall group performs, these proportions are maintained. The most important advantage to norm-referenced tests is that scales created in this way have properties that permit them to be treated mathematically. The science of assessment is based

on norm-referenced assessments, and much more precise comparisons can be made using these methods. Item analysis techniques have been developed that adjust item difficulty in ways that push a distribution into the bell-shaped curve that maximizes reliability. The more sophisticated techniques of item-response theory require a norm-referenced approach.

Standards-based assessment interprets test scores by comparing them with absolute standards. Such comparisons can be more meaningful than merely comparing a student with other students. A determination of the desired level of student performance is designated ahead of time, and once this cut-score is established, passing or failing does not depend on how other students have performed. This means that the percentage of students who pass will not be known until the test is administered. Standards-based assessment is much more compatible with the underlying principles of standards-based reform than norm-referenced assessment. There are two primary disadvantages to the use of standards-based assessment: (1) tests developed in this way tend to have undesirable mathematical properties, and (2) setting cut-scores can be difficult.

Ideally, a standards-based test should consist of a set of items written at a difficulty level that assures that a student who is functioning at the appropriate academic level will get the items correct and the student who is not will get them wrong. A test constructed in this way will efficiently discriminate between those who have mastered and those who have not mastered the content. Standards-based assessment requires the implicit assumption that students can be divided into these two categories. It requires that the majority of students be at the extremes rather than in the center, as you would expect if the scores were normally distributed. Even if this assumption is satisfied, such a test will have dreadful psychometric qualities and scores from such a test should not be added together to create a school score. Kentucky, like many other states, creates elaborate school indexes based on this type of assessment. Because CATS has been changed to a norm-referenced test, there is no reason to maintain the item difficulty structure implemented when the test was stan-

dards-based, and there are even stronger reasons to change it. Other states have chosen a different path and performed extensive item analysis procedures on tests that are supposed to be standards-based. It is a mistake to use norm-referenced techniques to develop a standards-based test and wrong to use the assumptions of standards-based testing to create a norm-referenced test. When these sorts of mistakes are made, the quality of the test suffers, which usually means lower reliability. Although the standards-referenced approach has intuitive appeal, particularly in the evaluation of individual student performance, it provides a poor basis for comparing the performance of schools. It is better to use a norm-referenced approach to compare schools for accountability purposes.

7. Performance standards should have a rational basis.

Setting performance standards (assigning cut-points) is one of the most difficult problems in all of measurement. The job is even more of a challenge when there is a need for multiple cut-points, which is the case with CATS. Furthermore, CATS consists of a mixture of dichotomous and polytomous items, which makes the process of setting cut-scores even more difficult. Not only is there no accepted method of setting performance standards, there is serious doubt about whether a workable method for setting cut-scores even exists. The consensus expert on this topic was Richard M. Jaeger (now deceased). He was selected to write the chapter in the *Third Edition of the Educational Measurement Handbook*¹⁷ (1989) that focused on standard setting. In reviewing the literature on standard setting, he cites the empirical research of

. . . Poggio, Glasnapp, and Eros (1981),¹⁸ showing that test standards depend heavily on the methods used to derive them, and results reported by Jaeger, Cole, Irwin and Pratto¹⁹ (1980), showing that test standards vary markedly across types of judges used with a single standard-setting procedure. Linn et al. (1982)²⁰ conclude that thousands of students would be declared competent or incompetent in most statewide competency-testing programs on the basis of methodological decisions that have nothing to do with their abilities.

Both Glass (1978)²¹ and Shepard (1979, 1980)²² note that competence is by virtually all conceptions, a continuous variable. Setting a cutoff score that supposedly divides students into two distinct categories, the competent and the incompetent, is therefore unrealistic and illogical. Shepard argues strongly against the use of any single method of standard setting, and Glass would have us abandon competence testing altogether. (492)

Lest you note the dates associated with these conclusions and assume that better methods must be available by now, consider the 1999 report from National Research Council.²³ It explains that the National Assessment Governing Board (NAGB) has expended considerable resources studying how best to set the performance standards for the National Assessment of Educational Progress. Because NAEP scores are reported in terms of the achievement levels of “below basic,” “basic,” “proficient,” and “advanced,” the cut-points between them must be designated. There have been a series of reports criticizing the way these cut-points are set (Linn et al.²⁴ and Pellegrino et al.²⁵). These reports have had as their particular focus NAGB’s use of the Modified Angoff method. Although this method has been deemed inappropriate for setting the cut-points between achievement levels on the NAEP, no one has come up with a better way of doing it. The NAEP test is purely norm-referenced. Attaching cut-points to a norm-referenced test does not work well because its characteristic bell-shaped curve is not appropriate for this sort of standard setting. In 1978, Glass made what is still considered to be the definitive statement on standard setting, “To my knowledge, every attempt to derive a criterion measure [cut-scores] is either blatantly arbitrary or derives from a set of arbitrary premises” (258).²⁶

Kentucky uses graders to place responses to constructed-response questions into the appropriate category. In doing this, they are making decisions about whether student performance is Novice, Apprentice, Proficient, or Distinguished. This approach does not work for multiple-choice items. For multiple-choice items, a decision must be made regarding the number of items that must be correctly answered in order to place a student into a

particular category. McGraw-Hill has devised methods for combining constructed-response and multiple-choice items and placing them on the same scale. This method is called the CTB Bookmark method, and it is currently being used to scale the Kentucky assessment. Like other standard-setting methods, it requires judges to make arbitrary decisions about the cut-point between acceptable and unacceptable performance. Because it uses information about students' previous performance on the items, it is essentially a norm-referenced approach. Although it may seem ideal to be able to set absolute standards without reference to average performance, such a practice can lead to indefensible performance standards.

Advocates of standards-based reform are committed to the adoption of policies that will lead to higher academic performance. At the same time, it is important for states to set realistic standards. Ideally, standards should not be set so high that students cannot possibly meet them or so low that they become meaningless. Setting performance standards in this way is extraordinarily difficult. Standard setting is complex, and quite often, states end up with standards that were not what was intended. In most cases, this means cut-points are set too high.

While assigning cut-scores to tests constructed using norm-referenced methods may seem antithetical to standards-based reform, the use of standards-based assessment assumes that it is possible to make reasonable expectations of how students should perform prior to seeing the results of the student performance. Experience has shown this to be an implausible assumption. What tends to happen is that when reasonable people set minimal standards for what students need to know, they tend to overestimate acceptable student performance to a considerable degree. These overly ambitious performance standards stem from the same processes that lead to unrealistic content standards. Like content standards, performance standards tend to be set by subject matter specialists, either active teacher or university faculty. It is natural for them to believe that their own field is of utmost importance and to want students to aspire to the highest levels of performance in their areas. As is true with the setting of content

goals, experts setting performance standards tend to narrowly focus on their own standards and ignore the time needed for other subject areas. When these standards are applied to student performance, the failure rate tends to be too high. When this occurs, the pressure from the community can be overwhelming, and as a result, the cut-scores are adjusted. The revised cut-scores end up being made on the basis of typical performance, which renders the decisions norm-referenced anyway.

8. Use a multiple-choice rather than a constructed-response/essay format.

Although the initial version of KIRIS, introduced in 1992, included multiple-choice items, performances on these items were not included in school accountability indexes. Eventually it was decided that the costs of having them written and scored, along with the amount of student time they required, could not be justified if they were not going to be used to compute school accountability indexes. The selection of performance assessments and constructed response items rather than a multiple-choice format was not based on sound measurement principles. It was based on hostility toward conventional assessment practices and a commitment to “cutting-edge” assessment methods. Multiple-choice items were viewed as an old-fashioned way of assessing students. This also was a time of great excitement about the potential of authentic testing, and it was the intention of the designers of KIRIS to have this assessment eventually be based only on performance assessments and portfolios.

When performance tests proved to be impractical, unreliable, and too expensive, the test had to depend almost entirely on constructed-response items. Although the problems associated with performance assessments also characterized the portfolios, the portfolios had such a strong constituency that their use has continued. In 1998, the Kentucky state legislature, in HB 53, mandated the use of multiple-choice items despite the objections of the KBE and the KDE. Multiple-choice items now contribute to a third of each content area score.

The most efficient, reliable, and economical way to make the sort of decisions that KIRIS was intended to make is with multiple-choice items. Multiple-choice and constructed-response items serve different purposes. Multiple-choice items are most useful for making comparative decisions about student and school levels of achievement. Constructed-response items are useful for communicating what students know and what they do not know. On the large-scale assessments administered in most states, including Kentucky, schools receive information only about student scores. The results of the spring administration are not released until the end of the following September, too late to be of much use to teachers. Ironically, the constructed-response format requires far more time (and expense), which makes their use in providing timely information about individual student performance impractical. Even if the results could be provided in a timely fashion, information about how each student performed on specific items is not provided, so the most important advantage to the constructed-response format is lost. For the purpose served by the Kentucky tests and those of most other states, multiple-choice items are far better.

The decision by the authors of KIRIS/CATS to use a constructed-response format rather than rely on the more commonly used multiple-choice format was not based on effectiveness or even pragmatic considerations. Instead, the decision was based on misinformation and/or distrust of conventional standardized testing. Some of the reasons given for not using multiple-choice items are as follows:

- They can only measure the recall of facts and isolated pieces of information.
- They cannot be used to assess higher-level thinking processes.
- They represent an old-fashioned method of assessment that has since been replaced by more modern assessment techniques.

- They do not provide a fair measure of the achievement of non-Asian minorities and economically deprived students.

None of these assertions can withstand careful analysis. Not only can multiple-choice items be used to assess facts, dates, names, and isolated ideas, but they also can provide an effective measure of high-level thinking skills. Although they are not appropriate measures of creativity and organizational ability, they can be used to measure virtually any other level of cognitive functioning. Because economically disenfranchised and minority students do poorly on standardized tests and standardized tests are made up of multiple-choice items, there is the mistaken belief that it is the item format that is the problem. The gap between the performance of minority and nonminority is actually greater for constructed-response than multiple-choice items.²⁷

Multiple-choice items have several important advantages over constructed-response items. First, they are much less expensive and more reliable. They are more reliable because it is possible to include more items when multiple-choice items are used.

Item difficulty is a major headache in the implementation of standards-based systems. Even small deviations from the ideal can lead to tests that are either too easy or too hard. Inappropriate item difficulty can have an enormous impact on the number of students who pass or fail an assessment. The difficulty of items must be carefully controlled from year to year or no comparisons across years will be meaningful. When multiple-choice items are coupled with norm-referenced assessments, item difficulty presents few problems. The difficulty of multiple-choice items is easily manipulated by adjusting the similarities of the distracters to the correct answer. The more similar the distracter, the more difficult the item is, and the more different the distracters, the easier the item is. Because difficulty can be easily manipulated, it is possible to maximize variability and achieve high test-score reliability. Because students are being compared with each other and a different average is established each year, differences in test difficulty from year to year present few problems. On the other hand, problems with item difficulty are exac-

erbated when constructed-response items are used. The only way to control the difficulty of constructed-response questions is to change their content.

The most important type of validity associated with the large-scale assessments of achievement is content validity. This form of validity refers to the fidelity with which a test can assess instructional objectives. Tests made up of constructed-response items can contain only a limited number of items because each item requires a lengthy response. As a result, tests made up of constructed-response items will be less valid than those that utilize the multiple-choice format.

Multiple-choice and constructed-response items are compared in an article published in the *Journal of Educational Measurement*. The article was written by R. Lukhele, David Thissen, and Howard Wainer and titled "On the Relative Value of Multiple-Choice, Constructed-Response, and Examinee-Selected Items on Two Achievement Tests."²⁸

The test they chose to study was the Advanced Placement (AP) testing program of the College Board. This test is a good choice because the training of the examiners and the sophistication of the scoring methods used with the constructed-response items are "state of the art." Whatever defects the study uncovers in the scoring of the constructed-response items cannot easily be attributed to flaws in the training of the examiners or the methods they employed.

The article begins by making two important points: Constructed-response items are expensive, and the information that can be obtained from these items is similar to what can be obtained from multiple-choice items. The authors state:

Constructed-response items are expensive. They typically require a great deal of time from the examinee to answer, and they cost a lot to score. In the AP testing program it was found that a constructed-response test of equivalent reliability to a multiple-choice test takes from 4 to 40 times as long to administer and is typically hundreds of thousands of times more expensive to score. (234)

With respect to the uniqueness of the information provided by constructed-response items, they state:

The primary motivation for the use of constructed-response formats thus stems from the idea that they can measure traits that cannot be tapped by multiple-choice items—for example, assessing dynamic cognitive processes. (235)

Their conclusion was as follows:

Overall, the multiple-choice items provide more than twice the information than the constructed-response items do. Examining the entire test (and freely applying the Spearman-Brown prophesy formula), we found that a 75-minute multiple-choice test is as reliable as a 185-minute test built of constructed-response questions. Both kinds of items are measuring essentially the same construct, and the constructed-response items cost about 300 times more to score. It would appear, based on this limited sample of questions, that there is no good measurement reason for including constructed-response items. (240)

On the basis of the data examined, we are forced to conclude that constructed-response items provide less information in more time at greater cost than do multiple-choice items. This conclusion is surely discouraging to those who feel that constructed-response items are more authentic and hence, in some sense, more useful than multiple-choice items. It should be. (245)

When the purpose of a statewide accountability assessment is to make decisions about the effectiveness of schools, a multiple-choice item format is obviously preferable. From a technical, measurement perspective, there is really no choice. Of course, one option is to compromise and include both multiple-choice and constructed-response items. This is what Kentucky and several other states have chosen to do. The problem with using both item types is that it creates scaling problems because of the difficulties associated with combining the two item types. It is more difficult to do this with standards-based than norm-referenced assessment programs.

9. Do not use performance assessments in large-scale assessments.

The KERA legislation specifically mandated an assessment based on the use of performance assessments and portfolios, but at the time of the law's enactment, there was no established methodol-

ogy for implementing such a program, and there were no other states that could serve as models. Instead of starting with an entirely performance- and portfolio-based assessment, the initial strategy was to create an assessment that depended primarily on constructed-response questions and portfolios, along with some Performance Events, which were a first step in the direction of performance assessment. As experience in the use of performance assessments increased, KIRIS was supposed to become more dependent on their use. The use of constructed-response and multiple-choice items was expected to decline until the test included only alternative forms of assessment.

The Performance Events had some of the properties of a performance assessment. Proctors were sent to each school to place students into groups. Each group was given one performance task, which could be in math, science, or social studies. In some cases, the task required the manipulation of concrete objects or the creation of a product, but usually the task was in the form of a conventional essay question. The Performance Event tasks were constrained by practical limitations. They had to take place in a room, usually the cafeteria, within a fixed amount of time, with all graded responses in the form of an individual essay. Only these individual responses were included in the accountability indexes of schools.

The Performance Events were expensive and presented many logistical headaches. Proctors and graders had to be hired and trained, schools were disrupted while the assessment took place, and the special equipment and/or supplies required for some items had to be assembled. The proctors could never be sure what supplies would be available at the schools. In some cases, even the expectation that hot water would be available proved overly optimistic.

Although one-tenth of the KIRIS budget was devoted to the Performances Events, their technical qualities were so dismal that they not only made no positive contribution to the reliability of the accountability scores, their inclusion actually lowered it. The biggest problem with using performance assessments in a standards-based accountability system, other than poor reliability,

is the impossibility of equating forms longitudinally from year to year or horizontally with other forms of assessment. In Kentucky, because of the amount of time required, each student participated in only one performance assessment task. As a result, items could never be reused from year to year because of the likelihood that students would remember the tasks and their responses. This made equating almost impossible. Despite technical consultants devoting a great deal of attention to this problem and the use of several different equating schemes, no approach seemed to work.

An indication of the instability of the Performance Events scores can be seen in the eighth-grade Performance Events scores in math. On the 1992–93 test, the Performance Events score in math was 44.68; on the 1993–94 test, it was 40.69; and on the 1994–95 test, the score was 2.62. This is not a typo or error in computation. It is an indication of the incomparability of scores across years. The correlation between the math Performance Events scores and the math constructed-response scores for the years between 1993, 1994, and 1995 are .1882, .3991, and .3777 respectively, while the correlation between the math constructed-response scores and arts and humanities constructed-response scores are .6881, .6117, and .6853. Obviously, the math constructed-response items were measuring something quite different from the math Performance Event scores. This lack of compatibility made combining the math Performance Events and math constructed-response items into a single math score, as was being done, indefensible. It would have made more sense, from a measurement perspective, to combine the math constructed-response scores with the arts and humanity scores than with the math Performance Events.

In the early spring of 1996, it was decided that the 1995 Performance Events scores could not be reported because of questions about their legitimacy. Nevertheless, the KDE approved the administration of these tests for the spring of 1996. In August 1996, the Kentucky Department of Education promulgated an emergency regulation that not only eliminated the

Performance Events from future administration, but also removed their scores from the baselines and accountability scores for the current biennium. In order to comply with the legislative mandate that the assessment be primarily performance-based by 1996, the KDE insisted that constructed-response items were a form of performance assessment. The 1998 legislation that modified KIRIS into CATS eliminated any reference to performance assessment.

After the Performance Events were eliminated, there was concern in the legislature about the enormous expense incurred for the administration of an assessment that could not be used. The Office of Education Accountability hired the firm of Coopers and Lybrand to conduct an audit to determine whether Advanced Systems in Measurement and Evaluation should reimburse the state for test materials that were unusable. The auditors were unable to complete this task because the alterations in the contract surrounding the elimination of the Performance Events were based on verbal agreements, and there were few written records. Change orders could not be located, and there was no way to match contract deliverables with the amounts invoiced. The issue was eventually dropped.

10. Avoid matrix sampling.

Matrix sampling requires the assembly of a large pool of items that covers everything in the content standards. A number of different tests are created, with each made up of different subsets of questions. Across the sample, the breadth of the content is covered, but each student is assessed on only part of this content. The use of matrix sampling increases content validity, but it does so at the cost of reliability and the loss of information about individual student performance. The overall reliability of the assessment is lower than it would be for nonmatrix sample tests, and it is far too low to support reporting individual student scores. Furthermore, individual scores lack content validity.

When the National Assessment of Educational Progress was designed, it was intended to evaluate the American education

system as a whole. Only later was it used to compare states, and it was never intended for use in making comparisons below the state level. The restrictions on the levels at which scores could be reported that characterize matrix sampling were useful in obtaining the cooperation of local school district officials. They were reassured because the use of matrix sampling meant that scores could be reported only at the state level. Schools and school districts could agree to participate without fear that the results would be used to hold them accountable.

When the KIRIS assessment was designed, Advanced Systems in Measurement and Evaluation, the original contractor, was told to make the design of the assessment similar to the NAEP tests. Using the NAEP as their model, ASME included matrix sampling in the design of KIRIS. An assessment can be constructed to have an acceptable level of content validity through a process of randomly sampling content. The use of matrix sampling with KIRIS precluded the use of individual student scores. When students realize that they are not to be held accountable for their individual performance, their motivation lags and the validity of all scores, including those for individual schools, is harmed.

Tests must be equated from year to year and the inputs from tests over different content areas and different formats must be equated in order to make meaningful comparisons. Under the best of circumstances, when a large number of multiple-choice items are administered to a large sample of students, accurate equating is difficult to achieve. When constructed-response items are used with a standards-based system, equating becomes even more difficult. Adding the different forms that are required for matrix sampling makes the process nearly impossible.

For the above reasons, the use of matrix sampling is not recommended. The advantages of increased content validity are canceled by the loss in reliability. What is most surprising is that when Kentucky's assessment was revised in 1998, the matrix-sampling model was retained. This may have been the result of a misunderstanding of the nature and prerequisites of content validity or a commitment to the belief that individual scores should never be used.

11. Portfolios are inappropriate for large-scale assessments.

With KIRIS and now CATS, writing achievement is assessed using a writing portfolio and an on-demand writing task that is included with the constructed-response questions. Writing achievement is assessed at grades four, eight, and twelve. For the KIRIS portfolio assessment, students included five writing samples from their language arts class and a sixth selection from another class. In the interest of reducing the amount of class time devoted to assessment, CATS requires fourth-grade students to submit only four selections, and middle and high school students to submit five selections, which includes a selection from another class. Decisions about what is to be included in the portfolio are made by the student.

The writing portfolio is intended to serve two purposes: (1) to evaluate teachers, principals, and schools and (2) to improve student writing skills. Unfortunately, it is difficult to accomplish both purposes well. Advocates of the use of writing portfolios assert that their use increases the importance of writing and provides extensive practice for students. To do this effectively, teachers need to work closely with students. A good teacher will spend more time with weaker students and less with better students. If teachers extensively assist students in the preparation of their portfolios, the portfolios will no longer reflect student writing ability and will be an invalid measure of these skills. Extensive teacher assistance, particularly when it focuses on the lower-performing students, also suppresses variability and thereby lowers reliability. If policies are instituted that restrict the amount of assistance teachers are allowed to provide, the instructional value of portfolios will be diminished.

In Kentucky, teachers help students assemble their portfolios, which are graded at the student's school. Although some principals arrange the grading of portfolios in such a way that teachers do not grade their own student's portfolios, the ethical standards published by the KDE do not prohibit this practice. Giving teachers the responsibility for assisting in the assembly as well as

the evaluation of portfolios, then using the results of this evaluation to measure teacher effectiveness creates a conflict of interest.

Portfolio scores have always been the most unreliable of any of the input data used for computing school accountability scores. Anomalies in the scoring of the portfolios become obvious when the breakdown of KIRIS scores across schools is examined. Many small schools have shown dramatic increases that are difficult to defend as reflecting legitimate increases in student writing performance. The average portfolio scores for some schools have made gains as great as 50 points, say from 20 to 70 (on the KIRIS/CATS 1 to 133 scale). It is possible for teachers to compromise the validity of the portfolio in two ways: (1) teachers can grade the portfolios of their own students too leniently, or (2) teachers can provide too much help to students in the writing of the selections included in the portfolio.

As the school administrator, the principal is responsible for both ensuring the integrity of the assessment and, at the same time, making sure that his or her school has adequate scores to obtain rewards and avoid punishment. Since KERA was implemented, principals have understood that the writing portfolios are the one aspect of the assessment that is completely controlled by teachers and the one that can be most easily manipulated. Rather than urging teachers to be objective in their evaluations, principals are more likely to pressure them to get higher scores. The more pressure that is placed on teachers to increase portfolio scores, the less effective portfolios will be as an assessment tool.

Reports from audits of portfolio scoring indicate the degree to which teachers are under pressure to be generous in their grading. Despite enormous efforts aimed at ensuring that teachers correctly grade portfolios, the audited results of randomly selected portfolios are startling. An examination of Table 8.5 reveals the percentage of scores assigned by auditors that differed from those assigned by teachers. Only at the Novice level did auditors grade the portfolio higher as often as lower. At the other three levels, the auditors almost always assigned lower scores than the teacher.

TABLE 8.5 Percentage of Auditor Scores That Differed from Those Assigned by the Teacher

<i>Category</i>	<i>4th Grade</i>	<i>8th Grade</i>	<i>12th Grade</i>
Novice	22	4	10
Apprentice	10	25	22
Proficient	24	44	44
Distinguished	71	94	91

In response to concerns about the level of assistance students received in the creation of portfolios entries, the KDE published a set of ethical guidelines in 1996, which were approved as regulations. The 1996 guidelines included a blanket statement that teachers could not make corrections on student work. Because of controversies surrounding this regulation, the standards were rewritten in 1999. Instead of referring to student work in general, the new document focused on the entries that were to go into the writing portfolio. This is not as much of a change as it seems because almost anything a student writes can be included in his or her portfolio. From among all possible entries, the student chooses the four or five pieces that are to be included. Because one selection must come from outside of the language arts class, almost anything a student writes in any class is a candidate for inclusion in his or her portfolio. The 1999 rules allow a teacher to indicate the location of spelling, grammar, and punctuation errors, but prohibit any direct corrections. This means that a teacher can tell students that they have misspelled a word, they have made an error in grammar, or used inappropriate punctuation, but cannot provide the correct usage. Once a student is told that he or she has made an error, it is up to the student to figure out how to correct it. Furthermore, the ethical guidelines say that students are not supposed to obtain help from anyone else, including parents or peers.

These rules are intended to serve a purpose that goes beyond the protection of the integrity of the assessment system. They

also make it unethical to teach writing using any approach other than the *writing process method*, an extension of the whole-language method of teaching reading. The director of the Kentucky Writing Program describes this policy as “best practice in writing.” Advocates of these approaches believe that reading and writing are entirely natural processes and that students acquire these skills most easily with minimum interference from teachers. It is assumed that learning to read and learning to write are the same as learning to speak. Children learn to speak naturally, needing no prompting from parents. Correcting a young child’s speech can interfere with this natural process and supposedly can cause speech defects. The writing process approach assumes the same for writing. It is asserted that correcting a child’s writing will inhibit the child and prevent him or her from becoming a good writer.

Writing process is not so much a method of teaching writing as it is a philosophical justification for why students should learn to write on their own. There is also the belief among advocates of the writing process method that the most important outcome of writing activities by students in school is the opportunity for them to express their feelings; that is, advocates of the writing process method believe all writing is personal. The role of writing as a means of communication is deemphasized. Avoiding criticisms of student writing is believed to encourage the expression of feelings by students.

One of the strongest recommendations of the OEA panel²⁹ was that the portfolio scores be removed from the accountability index. The OEA report states, “Evidence about the adequacy of the measurement provided by the portfolios is limited but sufficiently negative to indicate that the portfolio scores are not at this time appropriate for use in the KIRIS high-stakes accountability system.” That suggestion has been rejected by the KDE because of their purported value for instruction. Again according to the OEA report, “Evidence about the impact of the program [portfolios] on instruction is limited, largely anecdotal, and inconsistent.”

The use of writing portfolios as a way to assess writing ability should be strongly discouraged. The experience in Kentucky provides evidence that they cannot be scored reliably and that the rules that must be implemented to protect their integrity end up having a strongly negative effect on the teaching of writing. It is not too much of an exaggeration to say that Kentucky's students are poorer writers as a result of the inclusion of writing portfolio scores in school indexes.

Concluding Comments About Lessons from Kentucky

The Kentucky accountability system has been in existence since 1992 and cost almost a billion dollars in its first eight years.³⁰ Despite the enormous commitment of resources, there is scant evidence supporting the success of Kentucky's educational reform. The only indicators of improvement in student achievement are KIRIS/CATS accountability scores, and these seem to be the result of changes in the scaling of the test.

One reason for the failure of the Kentucky accountability system is the confused and contradictory theoretical basis for the assessment. The KERA legislation and KIRIS itself were based on a traditionalist approach to education. Traditionalists believe that all students should be taught a standard liberal arts curriculum and that all students can learn this material at a high level or at least should be given the opportunity to do so. Associated with traditionalism is a commitment to conventional instructional methods, the recognition that students must be required to learn some content, the value of hard work, and commitment to high standards. It also places teachers in a central role in the education process. First and most fundamentally, traditionalists are committed to the belief that the level of academic achievement as determined by academic achievement tests can be used to judge the effectiveness of a school.

The KDE, which is charged with the implementation of the assessment system, has adopted a progressive education philosophy.

Progressive educators advocate the adoption of a student-centered classroom in which students choose their own school activities. They also believe that the proper role of the teacher is that of a guide or consultant who can help students reach their own goals. Most significantly, they believe that students differ in academic ability and that not all students can succeed at the same level. They oppose the assessment of academic achievement associated with standards-based reform because they believe it is unfair to students with low ability. Even more important, they do not believe that the assessment of academic achievement is a legitimate way to evaluate the quality of schools and teachers.

The adoption of a progressive education policy by the KDE has led to the paradox of a system in which schools are required to improve their student's academic achievement as measured by KIRIS or CATS or face unpleasant consequences of the CATS accountability system while being required by the KDE to adopt progressive instructional strategies. The mandated strategies were never intended to increase academic achievement. At the same time, low-performing schools are discouraged from adopting instructional methods that have been shown to be effective in improving academic achievement.

Because of these conflicts in educational philosophies, Kentucky is squandering the most important benefit of standards-based reform, its capacity to pressure educators into adopting effective instructional strategies. If there are meaningful consequences attached to performance, it might be expected that teachers would seek the best ways to improve their student's performance. An examination of the instructional material used by the Distinguished Educators who are sent in to help low-performing schools is rife with references to cooperative learning, developmentally appropriate practices, self-directed learning, learning styles, and multiple intelligence—the language of progressive education.

The second reason that KERA, KIRIS, and CATS have failed to improve academic achievement in Kentucky is the structure of the assessment system itself. Too often, ideology was substituted for sound psychometric practice in the construction of the test.

The ultimate purpose of the test was to compare schools. The test was never intended to evaluate the performance of individual students. The best way of comparing schools is the tried and true multiple-choice, norm-referenced test. This is the way most states conduct their standards-based reform. The authors of the KIRIS/CATS assessment were too clever for this. They wanted to create an assessment that was like no other. They also wanted a test that included the project approach so beloved by progressive educators, that is, performance tasks and portfolios. As a result, on a per-pupil basis, Kentucky has the most expensive testing system of any state. At the same time, the system is characterized by poor reliability and validity.

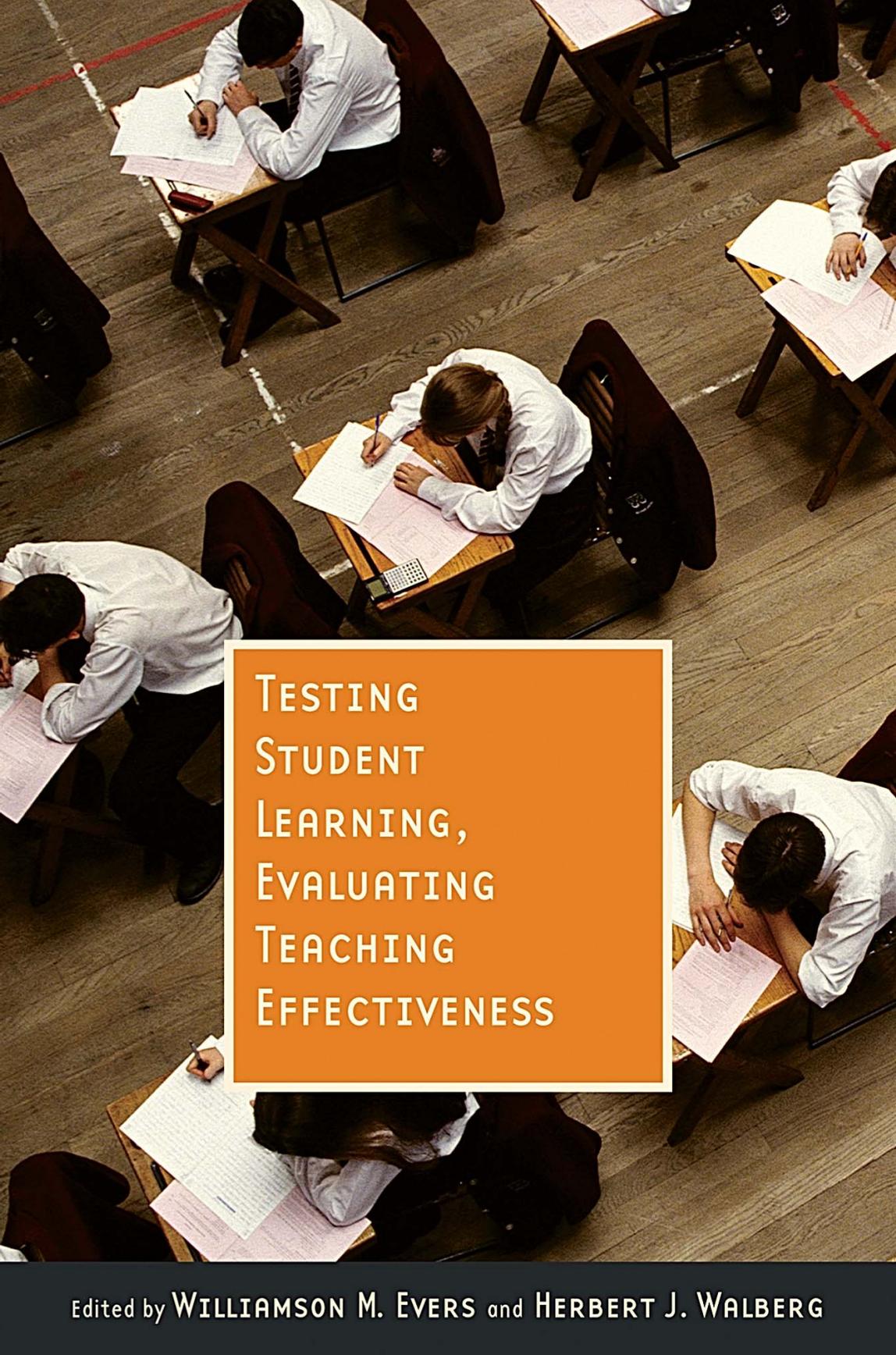
After twelve years, despite constant vociferous criticism and a series of expensive, highly critical reports written by distinguished experts and after numerous panels and committees have cited a litany of shortcomings and suggested extensive changes that needed to be made, the system persists largely unchanged. There were great expectations for change and seeming agreement between the legislative branches and the governor's office in the spring of 1998. After the dust cleared and the laws were passed, the essential elements remained in place. Bureaucracies are characterized by inertia, and changes come slowly if at all. Many Kentucky citizens, all over the state, in and out of education, are convinced that this system is not working. They also believe that it is having a deleterious impact on education in their state. The system's supporters have been able to fend off their critics for twelve years, and by manipulating the scaling so that schools appear to be doing better, they have prevented meaningful changes in the system.

Notes

1. *Rose v. Council for Better Education, Inc.* 790 S.W. 2d 186 (Ky. 1989).
2. Lawrence Picus, "Estimating Costs of Alternative Assessment Programs: Case Studies in Kentucky and Vermont" (paper delivered at the annual meeting of the American Educational Research Association, Chicago, March 1997).
3. Bert Combs, *Creative Constitutional Law: The Kentucky School Reform Law* (Lexington, Ky.: Pritchard Committee for Academic Excellence).

4. The Evaluation Center, Western Michigan University, *Evaluation of the Development and Implementation of KIRIS Through December 1994 (KIER Report)* (Frankfort, Ky.: The Kentucky Institute for Educational Research, January 1995).
5. Ronald K. Hambleton, Richard M. Jaeger, Daniel Koretz, Robert Linn, Jason Millman, and Susan E. Phillips, *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991–1994 (OEA Report)* (Frankfort, Ky.: Office of Educational Accountability, 1995).
6. James Catterall, William Mehrens, J. M. Ryan, E. J. Flores, and P. M. Rubin, *Kentucky Instructional Results Information System: A Technical Review* (Frankfort, Ky.: Kentucky Legislative Research Commission, 1998).
7. Daniel Koretz and Sheila I. Barron, *The Validity of Gains in Scores on the Kentucky Instructional Results Information System* (Santa Monica, Calif.: RAND, 1998).
8. Western Michigan University Evaluation Center, *KIER Report*.
9. Office of Educational Accountability, *OEA Report*.
10. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1985).
11. Office of Educational Accountability, *OEA Report*.
12. Koretz and Barron, “The Validity of Gains in Scores on the Kentucky Instructional Results Information System.”
13. Wilmer W. Cody, “Commissioner Cody Comments on Education,” *Louisville Courier Journal*, 13 December 1998, A10.
14. Robert Guitterez and Robert E. Slavin, “Achievement Effects of the Nongraded Elementary School: A Best Evidence Synthesis,” *Review of Educational Research* 62, no. 4 (1992): 333–76.
15. Robert Glaser, “Instructional Technology and the Measurement of Learning Outcomes: Some Questions,” *American Psychologist* 18 (1963) 519–21.
16. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1999).
17. Richard M. Jaeger, “Certification of Student Competence,” in *Educational Measurement: Third Edition*, ed. R. L. Linn (New York: MacMillan, 1989), 485–514.
18. John P. Poggio, D. R. Glassnapp, and D. S. Eros, “An Empirical Investigation of the Angoff, Ebel, and Nedelsky Standard-Setting Methods” (paper delivered at the meeting of the American Educational Research Association, Los Angeles, April 1981).

19. Richard M. Jaeger, D. Irwin, and D. Pratto, *An Iterative Structured Judgment Process for Setting Passing Scores on Competency Tests: Applied to the North Carolina High School Competency Test in Reading and Mathematics* (Greensboro, N.C.: University of North Carolina at Greensboro, Center for Educational Research and Evaluation, 1980).
20. R. L. Linn, George Madaus, and J. Pedulla, "Minimum Competency Testing: Cautions on the State of the Art," *American Journal of Education*, 91 (1982): 1-35.
21. Gene V. Glass "Standards and Criteria" *Journal of Educational Measurement* 15 (1978): 237-61.
22. Lorrie A. Shepard, "Setting Standards," in *Practices and Problems in Competency-Based Instruction*, ed. M. A. Bunda and J. R. Sanders (Washington, D.C.: National Council on Measurement in Education, 1979).
23. National Research Council, *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress* (Academic Press, 1999).
24. Robert L. Linn, Daniel M. Koretz, Eva L. Bake, and Leigh Burstein, *The Validity and Credibility of the Achievement Levels for the National Assessment of Educational Progress in Mathematics* (Los Angeles: Center for the Study of Evaluation, University of California, 1999).
25. James Pellegrino, Lauress Wise, and Nambury Raju, "Guest Editors' Note," *Applied Measurement in Education*, 11, no. 1 (1998): 1-7.
26. Gene V. Glass, "Standards and Criteria."
27. George K. Cunningham, *Assessment in the Classroom* (London: Falmer Press, 1998).
28. R. Lukhele, David Thissen, and Howard Wainer, "On the Relative Value of Multiple-Choice, Constructed-Response, and Examinee-Selected Items on Two Achievement Tests," *Journal of Educational Measurement* 31, no. 3 (1994): 234-50.
29. Office of Educational Accountability, *OEA Report*. 4-3, 4-38.
30. Lawrence Picus, "Estimating Costs of Alternative Assessment Programs: Case Studies in Kentucky and Vermont."



TESTING
STUDENT
LEARNING,
EVALUATING
TEACHING
EFFECTIVENESS

Edited by **WILLIAMSON M. EVERS** and **HERBERT J. WALBERG**

Testing Student Learning, Evaluating Teaching Effectiveness

*The Hoover Institution gratefully acknowledges
the following individuals and foundations for their
significant support of the*

Initiative
on
American Public Education

KORET FOUNDATION
TAD AND DIANNE TAUBE
TAUBE FAMILY FOUNDATION
LYNDE AND HARRY BRADLEY FOUNDATION
BOYD AND JILL SMITH
JACK AND MARY LOIS WHEATLEY
FRANKLIN AND CATHERINE JOHNSON
JERRY AND PATTI HUME
BERNARD LEE SCHWARTZ FOUNDATION
S.D. BECHTEL, JR. FOUNDATION

Testing Student Learning, Evaluating Teaching Effectiveness

Edited by
Williamson M. Evers and Herbert J. Walberg

HOOVER INSTITUTION PRESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

The Hoover Institution on War, Revolution and Peace, founded at Stanford University in 1919 by Herbert Hoover, who went on to become the thirty-first president of the United States, is an interdisciplinary research center for advanced study on domestic and international affairs. The views expressed in its publications are entirely those of the authors and do not necessarily reflect the views of the staff, officers, or Board of Overseers of the Hoover Institution.

www.hoover.org

Hoover Institution Press Publication No. 521

Copyright © 2004 by the Board of Trustees of the
Leland Stanford Junior University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission of the publisher.

First printing 2004

11 10 09 08 07 06 05 04 9 8 7 6 5 4 3 2 1

Manufactured in the United States of America

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48–1992.

Library of Congress Cataloging-in-Publication Data

Testing student learning, evaluating teaching effectiveness /
Williamson M. Evers and Herbert J. Walberg, editors.

p. cm.

Includes bibliographical references and index.

ISBN 0-8179-2982-7 (pbk. : alk. paper)

1. Educational tests and measurements—United States.
2. Educational accountability—United States. 3. Teacher effectiveness—United States. I. Evers, Williamson M., 1948–
II. Walberg, Herbert J., 1937–

LB3051.T4425 2004

371.14'4—dc22

2004001558

Contents

Introduction and Overview	vii
<i>Williamson M. Evers and Herbert J. Walberg</i>	
Part One: Setting the Stage	I
1. Examinations for Educational Productivity	3
<i>Herbert J. Walberg</i>	
2. Why Testing Experts Hate Testing	27
<i>Richard P. Phelps</i>	
Part Two: Constructive Uses of Tests	79
3. Early Reading Assessment	81
<i>Barbara R. Foorman, Jack M. Fletcher, and David J. Francis</i>	

4. Science and Mathematics Testing: What's Right and Wrong with the NAEP and the TIMSS?	127
<i>Stan Metzenberg</i>	
5. Telling Lessons from the TIMMS Videotape	161
<i>Alan R. Siegel</i>	
Part Three: Constructive Tests for Accountability	195
6. Portfolio Assessment and Education Reform	197
<i>Brian Stecher</i>	
7. Using Performance Assessment for Accountability Purposes	221
<i>William A. Mehrens</i>	
Part Four: State Testing Policies	243
8. Learning from Kentucky's Failed Accountability System	245
<i>George K. Cunningham</i>	
9. Accountability Works in Texas	303
<i>Darvin M. Winick and Sandy Kress</i>	
Appendix: Conference Agenda	323
Contributors	325
Index	331