REVIEW OF THE MEASUREMENT QUALITY OF THE KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM, 1991-1994

- Final Report -

Panel Members

Ronald K. Hambleton, Chairperson
University of Massachusetts at Amherst

Richard M. Jaeger
University of North Carolina at Greensboro

Daniel Koretz
The Urban Institute

Robert L. Linn
University of Colorado at Boulder

Jason Millman
Cornell University

Susan E. Phillips
Michigan State University

A Report Prepared for the
Office of Educational Accountability,
Kentucky General Assembly

June 20, 1995

Table of Contents

Preface i

Executive Summary 1

Chapter 1 - Introduction 1-1

1.1 Background to the Study 1-1

1.2 Statement of Purposes of the Measurement Review . . . 1-4

1.3 Limitations of the Measurement Review. 1-8

1.4 Focus of the Study 1-11

Chapter 2 - Developing the Cognitive Tests 2-1

2.1 Central Question and Principal Findings 2-1

2.2 Step 1. Specifying the Targeted Goals 2-4

2.2.1 Framework Specificity 2-5
2.2.2 Framework Congruence 2-10

2.3 Step 2. Writing the Assessment Exercises 2-12

2.4 Step 3. Evaluating the Exercises 2-13

2.4.1 Step 3A. Rating the Items 2-13
2.4.2 Step 3B. Item Analysis 2-18

2.5 Step 4. Selecting Items 2-20

2.6 So What Are These Cognitive Tests KIRIS Has Created? . 2-23

2.6.1 Judgments Based on Item Scrutiny 2-23
2.6.2 Data Relevant to Testing Hypotheses 2-26

Chapter 3 - Are the Accountability and Assessment Scores Reliable? 3-1

3.1 Central Questions and Principal Findings 3-1

3.2 The Concept of Reliability and Its Application to KIRIS 3-4

3.3 Analysis of Dependability of Accountability Decisions 3-9

3.4 Analysis of Reliability of Assessment Uses of Scores 3-23

Chapter 4 - Kentucky's Portfolio Assessments: Are the Scores Reliable and Valid? 4-1

4.1 Central Questions and Principal Findings 4-1

4.2 The Operation of the KIRIS Portfolio Program: Key Points 4-5

4.3 The Reliability of the KIRIS Writing Portfolio Scores 4-9

4.3.1 Consistency of Ratings 4-10
4.3.2 Student-Level Score Reliability 4-16
4.3.3 School-Level Reliability of KIRIS Portfolio Scores 4-19

4.4 The Validity of KIRIS Writing Portfolio Scores 4-21

4.4.1 The Validity Implications of Lack of Standardization 4-22
4.4.2 Validity Implications of Contributions from Others to Portfolios (and Performance Events) . . 4-23
4.4.3 Validity Evidence From the Scoring of Portfolios 4-25
4.4.4 Convergent and Divergent Evidence 4-27
4.4.5 Concordance Between Writing Portfolios and On-Demand Writing 4-32

4.5 Effects of the Portfolio Program on Instruction . . . 4-38

Chapter 5 - Are the Scores Comparable Across Years? 5-1

5.1 Central Questions and Principal Findings 5-1

5.2 Purpose of Equating 5-4

5.3 Levels of Equating 5-5

5.4 Equating with Item Response Theory Models 5-6

5.5 Fit Statistics 5-9

5.6 Professional Standards 5-9

5.7 Application to KIRIS 5-10

5.7.1 Positive Aspects of KIRIS Equating 5-12
5.7.2 1991-92 Linking Procedures for Transitional Assessments 5-12
5.7.3 1992-93 Linking Procedures for Transitional Assessments 5-17
5.7.4 1993-94 Linking Procedures for Transitional Assessments 5-21

5.7.5 Performance Events 5-25
5.7.6 Writing Portfolios 5-25
5.7.7 Moving the Grade 12 Assessment to Grade 11 . . . 5-25
5.7.8 Equating Errors 5-26

Chapter 6 - Are the Performance Standards Defensible? . . . 6-1

6.1 Central Questions and Principal Findings 6-1

6.2 Some Essential Background on the KIRIS Accountability Index and its Use 6-9

6.3 How Performance Standards were Set for KIRIS 6-12

6.4 Setting Performance Standards for Matrix-Sampled Items in 1991-92 6-19

6.5 Setting Performance Standards for Performance Events in 1991-92 6-25

6.6 Setting Performance Standards for Open-Ended Items in 1992-93 and 1993-94 6-26

6.7 Does the Location of Performance Standards Matter? . . 6-27

Chapter 7 - Are The Reports of Assessment and Accountability Results Clear and Understandable? 7-1

7.1 Central Questions and Principal Findings 7-1

7.2 Relevant Test Standards 7-5

7.3 KIRIS 1993-94 Assessment Curriculum Report 7-7

7.4 KIRIS Biennium I: School Accountability Report . . . 7-10

7.5 Student/Parent Reports 7-11

7.6 Briefing Packet-1993-94 KIRIS Assessment Results . . . 7-15

7.7 1993-94 Interpretive Guide 7-18

7.8 Celebrate the Progress - News Packet 7-19

Chapter 8 - Is There Evidence That Education Has Improved in Kentucky? 8-1

8.1 Central Questions and Principal Findings 8-1

8.2 Evaluating Gains in KIRIS Scores 8-4

8.3 NAEP Trial State Assessment 8-8

8.4 ACT Scores for Students Taking ACT and KIRIS 8-11

8.5 CTBS/4 and CAT 5 Scores From District Testing Programs 8-15

Chapter 9 - Conclusions and Recommendations 9-1

9.1 Purpose of the Review and General Conclusion 9-1

9.2 Major Findings 9-2

9.3 Recommendations 9-7

Appendix A - Legal Issues

Preface

A six-person Technical Review Panel was formed by the Office of Educational Accountability (OEA) of the Kentucky Legislature to conduct a review of the measurement quality of the Kentucky Instructional Results Information System (KIRIS) as designed and implemented in the Kentucky public schools between the fall of 1991 and the spring of 1994. The Panel's work began in the fall of 1994 with a review of pertinent documents about KIRIS, the development of a work plan, and a planning meeting in early December of 1994. The research was conducted between December of 1994 and June of 1995 and resulted in the preparation of this Report.

The Panel hopes that the Report and especially our main conclusions and recommendations are useful to the Kentucky Legislature and the Kentucky Department of Education in restructuring and designing the educational system to provide the very best education possible for children in the state.

The Technical Review Panel is appreciative of the efforts of a number of persons and staffs for their assistance in providing information used in the preparation of this Report. Ed Reidy and Neal Kingston and their staffs at the Department of Education were always responsive to the Panel's informational needs and always willing to supply documents, background materials, and data, and to discuss technical issues and findings. Their comments on an earlier draft of this Report were especially helpful.

Richard Hill, Stuart Kahl, and staff members from Advanced Systems in Measurement and Evaluation were responsive to our

informational needs and provided many insights and technical details which were useful to our Panel in completing the work.

The Panel is also grateful for the excellent co-operation and assistance we received from Penney Sanders, Valerie Forti, and Doug Terry of the OEA. They were invaluable in providing documents and timely consultation on the Panel's work.

Finally, the Panel wishes to acknowledge the contributions of Sheila Barron of the Urban Institute, who assisted with the Panel's analyses of KIRIS and ACT data, and Peg Louraine from the University of Massachusetts at Amherst who provided administrative and clerical support to the panel and assisted in the production of the Final Report.

REVIEW OF THE MEASUREMENT QUALITY OF THE KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM, 1991-1994

-EXECUTIVE SUMMARY-

Purposes of the Review and General Conclusion

In September of 1994 a national panel of six measurement specialists was appointed by the Office of Education Accountability of the Kentucky General Assembly to address the following question:

> Is the measurement quality of KIRIS sufficient to support the intended uses of the KIRIS results and the actions taken by the Kentucky Department of Education and the legislature? And, to the extent that shortcomings in KIRIS are identified, what changes would need to be made to improve the accountability and assessment system?

After reviewing large numbers of curriculum and technical documents and assessments, conducting a number of relevant analyses, and carefully considering our findings, the Panel is in unanimous agreement that KIRIS is seriously flawed and needs to be substantially revised. The Panel is not suggesting that the educational reform movement in Kentucky is a failure or that educators are not working hard to implement the goals of KERA. We are not suggesting either that the educational reforms taking place in the areas of curriculum design and instruction and teacher in-service training are not worthwhile. The Panel is saying that the accountability and assessment system has major flaws which need to be corrected as Kentucky moves into the second accountability cycle. This broad conclusion does not arise from any single piece of evidence but rather is based on the Panel's review and analysis of (1) assessment development procedures, (2) the accountability index and its reliability, (3) portfolio scoring, (4) the linking or equating of assessments from one year to the next, (5) the procedures used to set performance standards, and (6) evidence addressing the impact of KIRIS on student learning. Appropriate corrections or modifications along the lines of the recommendations below should allow KERA to come closer to achieving its full potential.

Background

In 1990 the Kentucky legislature passed a major education bill that was intended to substantially redesign educational goals, services, and programs in the state. The Kentucky Education Reform Act (KERA) required that new and ambitious goals for education be set and that curricula, instruction, and school

2. The Panel's impression is that the Department of Education is willing to improve KIRIS and there is evidence that improvements have been made over the last four years when problems have been identified (e.g. improvements in portfolio scoring, and assessment development). Even some of the preliminary suggestions from our Panel have already been accepted and incorporated into KIRIS.

3. The Department of Education and Advanced Systems have demonstrated the capability of developing essential material and training documents in the areas of KIRIS implementation, curriculum development, assessment development, and portfolio implementation and scoring.

4. The Department of Education has recognized the central role of teacher training and has devoted significant effort to the delivery of essential teacher training.

Recommendations

The Panel's major recommendation is that KIRIS needs to be restructured and redesigned if it is to accomplish its objectives. Main recommendations follow and more specific recommendations appear in other parts of our Final Report:

1. The portfolios should not be used at this time in the accountability index.

2. Any assessment data used for accountability purposes should be scored externally to the schools in which the data are collected.

3. An ethical code of conduct for assessment activities should be developed and implemented. The Department of Education should also develop an in-house capability for auditing school procedures and sanctioning confirmed breaches of assessment security.

4. The amount of validation work on the assessments should be expanded. Additional construct validation evidence is needed to support the various uses and interpretations of the performance assessment data.

5. The design for equating assessments should be strengthened and the ad hoc procedures eliminated. The Department should use a scientifically sound and rigorous approach to assessment equating because this activity is absolutely critical to the integrity of the total KIRIS system.

6. Performance standards should be re-established and full documentation of the process should be provided. The Department is strongly advised to avoid the use of ad hoc, unjustified statistical linkages in establishing standards.

7. The various score report forms should be field tested prior to their use to insure that communication of important information is clear and understandable. Also, effective communication of results must include caveats about the results (e.g., alternative interpretations and cautions) and indications of the measurement error in the data to reduce the likelihood of overinterpretation and misinterpretation.

8. Material provided to the press and the public should provide alternative interpretations of the results when such alternative interpretations are plausible. Results released in the media should contain important caveats such as the lack of generalizability of findings to new sets of tasks.

9. There is a great need to establish routine auditing procedures on all aspects of KIRIS including assessment development, standard-setting, equating, etc. Because of the high-stakes nature of KIRIS and the resulting potential for inflated gains in scores, it is essential that mechanisms be established for ongoing auditing of observed gains on KIRIS.

10. Item assessment formats should be used which contribute to the validity of the educational assessments. This means that multiple-choice items should have a role to play and will be valuable in enhancing content validity, the reliability of school and student scores, score equating, and score reporting.

11. The documentation in technical areas of equating, standard-setting, and score reporting, needs to be substantially improved to facilitate review and replication.

12. There has been a shift toward process at the expense of content in the curricula and this shift needs to be reconsidered. Our Panel does not have a view that the current situation is wrong. We simply feel that this situation needs to be reviewed to be sure that the impact on instruction, while presumed by the Department of Education to be positive, is, in fact, positive. In addition, the implications of this shift away from content for the adequacy of measurement--for example, for the accuracy of the estimates of change upon which KIRIS focuses--should be more fully evaluated.

In summary, the Panel feels that the Kentucky Department of Education set unrealistic expectations for itself and its contractor to deliver an accountability and assessment system that would meet all of the essential Test Standards within a period of several years. The result is a system which has many technical shortcomings and, therefore, is not meeting the accountability and assessment needs of the Commonwealth as represented in KIRIS. Progress has been made in the last four years and the Department, its contractor, and everyone who has worked on KIRIS and its implementation deserve credit for the efforts and accomplishments which have been made. On the other hand, considerably more progress is needed to establish KIRIS as a technically sound accountability and assessment system. The Panel hopes that our conclusions and recommendations will be valuable to the Department of Education as it sets out on the second accountability cycle.

Reference

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington: APA.

CHAPTER 1

Introduction

1.1 Background to the Study

In 1990 the Kentucky legislature passed an education bill that was intended to substantially redesign educational goals, services, and programs in the state. The Kentucky Education Reform Act (KERA) required that new and ambitious goals for education be set and that curricula, instruction, and school administration be reorganized in the state to insure that all students in the Kentucky public schools meet these new goals. Schools would be held accountable for student learning through a system of rewards and sanctions.

KERA affects public education at every level and in a comprehensive way with the goal of improving student learning and performance. A major component of the education reform initiative in Kentucky is the Kentucky Instructional Results Information System (KIRIS). KIRIS is the accountability and assessment system of KERA which (1) produces annual results on the performance of schools, districts, and the state, in relation to the new goals of Kentucky education in the areas of reading, mathematics, science, social studies, and writing, (2) holds schools accountable for achieving the new goals of Kentucky education, (3) results in schools being given rewards and sanctions based upon student performance at selected grades (grades 4, 8, and 12) in relation to expected levels of performance on both cognitive as well as non-cognitive outcomes,

and (4) provides student performance results in five subject areas to students, parents, and their teachers.

Among the main features of KIRIS are that (1) the rewarding and sanctioning of schools based upon student performance makes KIRIS a "high-stakes" accountability system for schools, (2) school accountability depends mainly on the assessment of cognitive outcomes in the areas of mathematics, reading, social studies, science, and writing performance but also on non-cognitive variables such as school attendance, and (3) assessment is performance-based which means that students are expected to construct responses to problems and tasks and produce performances and projects rather than select answers to problems (such as is the case with multiple-choice items).

These three features of KIRIS as a set as well as others in KIRIS are relatively new in American education and not easy to implement because they represent a major shift in educational philosophy, curricula, instructional methods and approach to assessment, These features are also controversial as might be expected in any major change in educational policy. Controversy might be expected, for example, when schools become accountable for the performance of students in their schools, regardless of the composition of the student body in their schools at the time of the assessment, and sanctions may be placed on them when there is no evidence of improvement from one school year to the next.

Clearly, there is a need to evaluate KIRIS as the first accountability cycle has been completed and the first set of rewards and sanctions has been made. Because of (1) the

importance of KIRIS to the Kentucky educational system, (2) the newness of the philosophy, concepts, and assessments in KIRIS, (3) the use of new curricula with the emphasis on higher level thinking skills, and (4) the controversial aspects of some parts of the educational reform package, evaluations are needed to guide the decision-making process concerning KERA and KIRIS, and their implementation in Kentucky.

Several studies have recently been completed which address various aspects of KIRIS such as (1) legislator attitudes about the educational reform initiative (Horizon Research International, 1994); (2) consistency of KIRIS with the legislative mandate; understanding and confidence of stakeholders; involvement of teachers and principals in design and development; accuracy, accessibility, and clarity of documentation; and impact of KIRIS accountability policies on students, teachers, and schools (The Evaluation Center, Western Michigan University, 1995); and (3) a comparative study of the predictive validity of KIRIS scores and the American College Test (ACT) with college grades. To date, there has not been a major measurement review of KIRIS. The main questions which needs to be answered are: Is the measurement quality of KIRIS sufficient to support the intended uses of the KIRIS results and the actions taken by the Kentucky Department of Education and the legislature? And, to the extent that shortcomings are identified, what changes would need to be made to improve the accountability and assessment system?

1.2 Statement of Purposes of the Measurement Review

A six-person Measurement Review Panel was formed in September of 1994 and met in December of 1994 with the Department of Education and the OEA, and approved a work plan. The main measurement questions which guided the work of the Panel along with brief descriptions of the questions themselves follow:

1. Are the cognitive tests of the Kentucky Instructional Results Information System (KIRIS) built in a sufficiently sound way to support the accountability and assessment tasks of the Kentucky Educational Reform Act (KERA)? To answer this question, the Panel reviewed the education goals in Kentucky along with the efforts to operationalize those goals into instructional initiatives and then considered the assessments which were produced to address the initiatives. How were the assessments developed, and can the development of assessments be improved? To the extent that the educational assessments fall short in quality, then the whole accountability and assessment system becomes questionable.

2. For the school accountability system: How accurate is the classification of a school into one of the accountability categories (Eligible for Reward, Successful, Improving, Decline, In Crisis)? Evidence regarding this question can be obtained by determining the expected size of the changes in a school's accountability index from one assessment to another of the same kind. Little confidence in the accountability system should exist if the classification of a school depends to a great extent on the particular tasks which are selected for the assessment, on the particular persons who rated the students' work, or on the cohort of students enrolled in a given year.

3. For the assessment uses of scores: What is the likely margin of error in reported scores? Scores are reported in KIRIS for both schools and students. To what extent might student scores change simply due to the choice of raters to grade student work, or the selection of assessment tasks? Without a reasonably low level of error in the scores due to the selection of tasks and assignment of scorers, any use of the scores would be limited. For example, it only makes sense to describe a student as "Proficient" when there is reasonable confidence that, on a new assessment with content chosen within the same curriculum framework and of the same difficulty as the first, a student would

have a high probability of being classified as "Proficient" again.

4. Are the scores from the writing portfolios reliable and valid? The development of writing skills is a central part of the education reform movement in Kentucky and therefore writing assessment is a main part of the accountability and assessment system. Are the portfolios scored accurately enough for the purposes to which KIRIS puts them, and are they implemented in ways that will produce valid inferences about gains in student performance?

5. Are KIRIS scores equivalent across administrations? Equivalent measures are especially important when the focus of an accountability system is on growth or change over time. To the extent that assessments from one year to the next are not statistically equivalent and equivalent in content, comparisons of achievement from one year to the next will be seriously misleading.

6. Are the performance standards defensible? KIRIS is a criterion-referenced assessment system and this means that performance standards are set to define levels of performance or proficiency in relation to a well-defined set of educational outcomes. Students are classified as Novice, Apprentice, Proficient, and

Distinguished, and the percentage of students in each category over time is the primary component in the accountability index which drives school rewards and sanctions. It is critical, therefore, that a defensible method for setting performance standards be adopted and then be implemented properly.

7. Are the reports of accountability and assessment results prepared for policy-makers, administrators, educators, students, and parents clear and understandable? To the extent that any or all of these groups may be misinformed because of the types of report forms used or scales on which results are reported, KIRIS will fall short of its goals. It is impossible for persons who are misinformed to use the wealth of accountability and assessment information provided by KIRIS.

8. To what extent do any gains in KIRIS scores reflect real improvements in student learning? Large gains in KIRIS scores have been reported by the Kentucky Department of Education. However, experiences with other test-based accountability systems suggest caution: Gains in scores on high-stakes assessments do not always indicate commensurate gains in student achievement. Spurious gains can be produced by a variety of factors, such as breaches in security and

teaching students to be better test-takers. The Panel examined limited data from other assessments -- in particular, the National Assessment of Educational Progress and American College Testing college admissions tests-- to explore the meaning of the large gains in KIRIS scores during the first biennium.

The eight questions above provided the framework for the review that was carried out by the Panel. In addition, Professor Susan Phillips provided a review of KIRIS from a legal perspective. That work and recommendations which were generated from Dr. Phillips' review are contained in Appendix A.

Prominent in the Panel's work are the 1985 Test Standards developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, & NCME, 1985). These professional standards for the construction and use of educational and psychological tests provide an excellent basis for reviewing KIRIS. Where technical standards were not present for aspects of KIRIS, the Panel applied its collective professional judgment and also referred to the excellent set of guidelines for evaluating performance assessments prepared by Linn, Baker, and Dunbar (1991).

1.3 Limitations of the Measurement Review

The Panel's review was limited to measurement aspects of KIRIS (i.e., aspects of KIRIS pertaining to the development and

uses of educational performance assessments in accountability and assessment). A full program evaluation with concern for all aspects of KERA and KIRIS and with special concern for the impact of KIRIS on student learning would have been desirable but would have gone well beyond a measurement review. Still, a measurement review is valuable because the measurement aspects impact considerably on the educational policies which drive KIRIS. For example, to the extent that (1) the assessments themselves are not sound (i.e., do not measure the desired learning outcomes), (2) classifications of schools are inconsistent, (3) portfolios are unreliably scored, (4) student and school performance are dependent on some arbitrary aspects of the process (e.g., particular choice of tasks and scorers), or (5) the performance standards are set inappropriately, then, it is difficult if not impossible to implement the educational policy described in the new legislation. A technically sound accountability and assessment system is needed to fully and successfully implement KERA.

Many important measurement aspects of KIRIS could not be satisfactorily addressed by the Panel with the time and resources which were available. For example, the Panel was not able to observe teachers in the classrooms, or talk to administrators and teachers about the strengths and weaknesses of KIRIS. Such information would have been valuable. Fortunately, some work along these lines was completed by The Western Michigan University evaluation team (The Evaluation Center, Western Michigan University, 1995). Also, the Panel was not able to

review the assessments, related materials such as training documents, to the extent that would have been desirable. For example, the Panel would have benefitted from observing the training of scorers and scoring of portfolios and the setting of performance standards. Future evaluation efforts might be planned at the outset of the second accountability cycle to enable the collection of data which could influence the scope and direction of KIRIS.

The Panel did not look at the non-cognitive indicators in the Accountability Index. The main reason was that variability on these indicators for schools over time is relatively low, hence, the non-cognitive indicators have only a small impact on changes in the Accountability Index. At the same time, the Panel recognizes that there may be excellent reasons for retaining the non-cognitive variables in the Accountability Index despite their low impact. They represent one important class of indicators of school success.

Also the Panel did not have the time to carefully review the Accountability Index itself with the weights and how they were derived. The Panel notes in passing that as school accountability scores approached 100, any gains will be expected to become considerably less reliable and therefore less useful in rewarding and sanctioning schools. (This is because the sizes of the expected gains may be about the size of the errors associated with the gain scores.) On the other hand, the Panel did commit a substantial amount of time to the question of the consistency or reliability of the school accountability index.

1.4 Focus of the Study

The Panel's measurement review was carried out within the context of the educational reform movement in Kentucky. The system is a dynamic one with changes being made to correct problems and shortcomings as they are identified. Four good examples of areas where mid-course corrections were made were in the areas of assessment development, portfolio design and scoring, assessment score equating, and score reporting. The Panel agrees with the general directions of the improvements while at the same time, it must be recognized that the results obtained in 1991-92 provide the baseline for assessing school growth. Any changes made in KIRIS in 1992-93 and/or 1993-94 complicate considerably the valid assessment of growth since the baseline results may no longer be strictly comparable to the results obtained in 1993-94.

Accountability versus Assessment

KIRIS has two main components: (1) the accountability system, and (2) the assessment system. The accountability system concerns the use of cognitive and non-cognitive indicators of student performance to reward and sanction schools. The assessment system involves the collection of cognitive data to report to schools and students about performance and learning. Students, for example, are placed in performance categories, and schools are informed about the percentage of students by grade and subject who are in the various performance categories. Such information is intended to be used at the school level to improve instruction and provide timely feedback to students. At the same

time, the assessment results can be aggregated across students, subjects, and grades within a school and used in the calculation of the school accountability index.

Because the assessment information obtained from students is used in very different ways in accountability and assessment, the basis of evaluation which is specific to use is different also. As a result, whenever necessary, such as in the discussion of portfolios, the Panel makes clear distinctions between the two components or uses of the data. It is quite possible, for example, for portfolios to be useful for assessment purposes but far less useful for accountability purposes, or vice-versa. The situation is analogous to a carpenter who would find a hammer, for example, quite useful for one purpose such as putting a nail through a board, but almost completely useless for cutting the same board in half. Often, assessment instruments will serve one purpose better than others and therefore assessment instruments need to be evaluated for each intended use. According to the Test Standards (AERA, APA, & NCME, 1985), assessment instruments must be evaluated for each purpose for which they are used.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington: APA.

Horizon Research International. (1994). A survey of legislators on the Kentucky Instructional Results Information System (KIRIS). Frankfort, KY: Office of Educational Accountability.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

The Evaluation Center, Western Michigan University. (1995, January). An independent evaluation of the Kentucky Instructional Results Information System. Frankfort, KY: The Kentucky Institute for Education Research.

CHAPTER 2

Developing the Cognitive Tests

2.1 Central Question and Principal Findings

Are the cognitive tests of the Kentucky Instructional Results Information System (KIRIS) built soundly enough to support the assessment and accountability tasks of the Kentucky Educational Reform Act (KERA)? A yes or no answer would not be a fair characterization of the tests because they are neither all good nor all bad. We answer the question in relative terms and point out strengths and weaknesses.

In general the testing process and documentation became stronger over the three years. (An exception, in the Panel's opinion, was the decision not to include the multiple-choice items in the accountability formula.) The Technical Review Panel has mixed feelings about this decision. The Panel feels discouraged because the first year, when the Panel found the test development process most suspect, is extremely important. It gives educators often enduring first impressions of KIRIS since, as the provider of the base year data, it counts for half of the gain score index on which the accountability decisions rest, and since the first year results are the only basis for standard setting. The Panel also feels encouraged, however, because improvement in the test development process can strengthen future biennium assessments and is consistent with what is reported to

be the view in July 1994 of "the great majority" of the Kentucky legislators that KIRIS is "heading in the right direction."[1]

The Panel found most of the test specifications (frameworks) under-specified in their ability to communicate to Kentucky educators what learning and abilities their students will be accountable for. The Panel believes the 1993 and 1994 distribution of the Transitions and the Content Guidelines, respectively, the release of some of the base year exercises, and the language used in the academic expectations will contribute positively to informing educators about the measured outcomes. These materials appear to be helpful in describing learning outcomes and curricula. We suspect that this increased awareness of the measured goals is in part responsible for the widespread increase in KIRIS scores over base year performance. (See also Chapter 8.)

The Panel's review of the item-writing and item analysis process was hampered by a lack of detail about such information as selection and qualifications of advisory group members, how the rating sessions were conducted, what precisely were the questions put to the raters, findings from the small-scale field tryouts, etc.

The Panel did find a typical number of items that appeared somewhat flawed. The statistical properties of the items, however, especially the ability of the items to discriminate more able students from less able ones, were substantial.

---

[1]Horizon Research International, A Survey of Legislators on Kentucky Instructional Results Information System (KIRIS). Louisville: July 1994, p. 11.

Not all academic expectations are assessed and those that are included are weighted unequally. The Panel believes that these deviations from complete coverage and equal weighting are acceptable, but would have preferred that the intended weighting be communicated to the schools.

The items appear to measure by design generalizable skills and cross-covering themes and to de-emphasize discipline-based knowledge. To the extent KIRIS drives curriculum and instruction in Kentucky, traditional subject matter will give way to a process oriented emphasis.

The Panel recommends:

1. Use of the multiple-choice items so that the subject matter frameworks will be covered comprehensively.
2. More documentation of test development procedures and greater attention to the synchronicity between item wording and scoring criteria.
3. Conduct or sponsor studies to investigate to what extent, if any, unwanted factors are being measured by the assessment tasks.

The question on test soundness is extremely important to address. The Standards for Educational and Psychological Testing (APA, AERA, & NCME, 1985) require an affirmative answer:

> Tests and testing programs should be developed on a sound scientific basis.... (Standard 3.1)

Further, it is hard to imagine that KIRIS and, to a somewhat lesser extent, Kentucky's educational reform effort could be successful if the tests were poorly constructed. The tests are intended to model widespread and substantial change in what is

taught and how it is taught. The tests are intended to yield fine-grained measures of improvement capable of triggering rewards and sanctions and thus the motivation to change the way education is delivered in the state. These demanding instructional (assessment) and accountability tasks highlight the importance of developing the cognitive tests soundly.

Test development goes through four major steps:

1. The goals (desired outcomes) are specified explicitly.
2. Test exercises measuring these outcomes are constructed.
3. The exercises produced in step 2 are evaluated by (a) having judges rate the acceptability of the exercise against several criteria, and (b) submitting the responses of a sample of students to statistical analysis, and
4. Selecting and assembling the acceptable items for each form of each test.

As these steps are completed, test developers are mindful of what they are measuring and how they can demonstrate it.

2.2 Step 1. Specifying the Targeted Goals

Just as home builders use blueprints to guide their work, test constructors use specifications of outcomes sometimes called frameworks to guide their item-writing efforts. A good artisan could fashion a home without benefit of plans and an item writer could construct exercises without a detailed framework but in both cases the final product would reflect the idiosyncrasies of

the constructor and might not match the home owner's or test sponsor's wishes and expectations.

The Panel addresses here two questions about the frameworks: (1) Are the frameworks sufficiently explicit, and (2) Are the frameworks congruent with other goal delineations for Kentucky?

2.2.1 Framework Specificity

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) emphasize the need for clear specifications:

> Standard 3.2
>
> The specifications used in constructing items or selecting observations and in designing the test instrument as a whole should be stated clearly. The definition of a universe or domain that is used for constructing or selecting items should be described.
>
> Standard 3.3
>
> Domain definitions and the test specifications should be sufficiently clear so that knowledgeable experts can judge the relations of items to the domains they represent.

The six learning goals in KERA focus on: (1) using basic communications and math skills, (2) applying the core concepts in the discipline, (3) becoming self-sufficient, (4) becoming responsible group members, (5) developing thinking and problem solving ability, and (6) integrating and applying knowledge from all subject matter fields. In a step toward clarification, these goals were broken down into 75 valued outcomes by Kentucky's Council on School Performance Standards. These outcomes, in turn, were adopted by the State Board for Elementary and Secondary Education.

For the first test year, 1991-92, the test frameworks were a combination of the valued outcomes and the frameworks from the National Assessment of Educational Progress (NAEP). The contribution of the two sources for the frameworks differed across subjects.

The framework for reading consisted primarily of a matrix of two comprehension categories by four reading passage types. However, the comprehension categories contain 13 specific abilities total. In addition, 14 categories of reader awareness and one of reader response are listed. By consulting the framework and supplementary information, one might, for example, expect the student to be asked to visualize while reading a literary passage.

The mathematics framework is a 7 by 4 content by process matrix, the content coming from the valued outcomes and the four process categories from the National Council of Teachers of Mathematics. However, the content for each grade is further spelled out in the 1991-92 Technical Report (Kentucky Department of Education, 1993). By consulting the framework and supplementary information, one might, for example, expect 8th grade students to be assessed on their conceptual knowledge about limits.

The social studies framework is a matrix of eight content areas (seven taken from the valued outcomes) by two cognitive skills (knowledge-comprehension and application of higher order thinking skills). By consulting the framework, one might, for example, expect the student to apply higher order thinking skills

to a question about the relationship of geography to human activity.

Six valued outcomes about the themes of science, four science content areas taken mostly from NAEP, and the two process categories used in social studies define the matrix framework for science. The content areas are each further divided into three or four subtopics. By consulting the framework, one might, for example, expect the student to be asked to demonstrate comprehension of patterns in meteorology.

Writing did not appear to have a framework per se. The on-demand writing situation consisted of the choice of two prompts, a generous time limit, and holistic scoring using the criteria of idea development, organization, support, sentences, wording, and mechanics.

These five frameworks remained the same throughout the first accountability cycle. In the second year of the first accountability cycle, two new frameworks were added, arts and humanities and practical living/vocational studies to supplement the baseline measures for the second accountability cycle. The heart of these two frameworks is a listing of valued outcomes -- seven for the arts and humanities and ten for the practical living/vocational studies areas.

Are these five frameworks explicit enough? One could answer "No" and argue that given the high-stakes purposes of KIRIS, it is especially important for educators to know what learning they are accountable for at a level of detail that allows them to modify their instruction accordingly. How can that happen when

so many of the cells of the frameworks do not vary across grade levels? Doesn't that imply that the outcomes are too general to guide instruction? Also, the lack of specificity calls into question what the equating means and casts doubts on change measures for individual schools. Further, tests like these that yield criterion-referenced interpretations require more rigorous definitions of frameworks than do norm-referenced achievement tests. (See the comment section of the Test Standards for Standard 3.2.)

One could answer "Yes" the frameworks are explicit enough and, as the Kentucky Department of Education (1993) does, argue:

> The KIRIS program does not focus on specific concepts, topics, and skills to be mastered at each grade level; such narrow objectives, consistent with traditional objective tests, encourage instruction based on factual recall and skills in isolation. Kentucky's assessment program targets larger concepts or themes, the connecting of ideas, and the application of knowledge, skills, and reasoning to solve problems. (1991-92 Technical Report, 1993, p. 13)

However, the Kentucky Department of Education (1994), at least in early 1994, felt as the Panel does, that the content specifications were not explicit enough:

> The content being assessed is not well enough defined and the relationship among the learning goals, academic expectations, the curriculum frameworks, the program of studies, course outlines, released items and performance standards is not clearly described. This makes it difficult for teachers to focus their instructional efforts and to explain to parents what is being taught and assessed. (p. 10)

The Department advocated the development and dissemination of content guidelines.

Six months later, Content Guidelines was published in draft form. For each of the seven subject areas or subject area pairs

assessed by KIRIS, the guidelines indicate subtopics included within an academic expectation. (An earlier publication, Transformations, lists appropriate activities for Grades 4, 8, 12 for each academic expectation included in Goal 1.)

The Panel notes that not all frameworks are equally ambiguous. Of the group, perhaps the reading domain is best described and the social studies, arts and humanities, and practical living/vocational studies domains are the least well described. Similarly, not all academic expectations are equally clear.

Consider academic expectation 2.20, which pertains to the social studies, "Students understand, analyze, and interpret historical events, conditions, trends, and issues to develop historical perspective." Although 14 subtopics are listed for Grade 8, few are very detailed. Examples are: change over time, multiple perspectives and viewpoints, roles of individuals in history. Some topics, e.g., cause and effect, are the same for two or all three grade levels.

For contrast, consider the 15 subtopics at the 8th grade level for goal 2.32, "Students demonstrate strategies for becoming and remaining mentally and emotionally healthy." Examples are: recognizing and understanding mood swings which occur during puberty, refusing pressure from peers and coping with potential rejection, recognizing factors that affect our decisions.

The Guidelines are a positive contribution, helping to clarify the content to be assessed, but they have room to

improve. The practice of releasing the common items also helps to communicate the content of the assessment.

It is the Review Panel's conjecture that an important contributing factor to the result that 95% of the schools improved their KIRIS test performance over that of their base year[2] is because the educators at the school level were better acquainted with the assessment tasks during years two and three. Other states (e.g., New Jersey) have shown dramatic increases in student achievement on different but equated yearly tests when detailed content frameworks become widely available. The Panel cautions that such large gains scores are unlikely to continue and that should the assessment focus or formats change from what educators may now be gearing up to, it is quite possible that performance would show a decline.

2.2.2 Framework Congruence

Four sets of goals could be said to contribute to the frameworks: the six goals of KERA, the 75 valued outcomes, 57 academic expectations, and the goals implicit in the NAEP frameworks. The Panel considers here whether or not they "speak with the same tongue." Our short answer is "yes." Briefly, the KERA goals are general enough to include the others; the valued outcomes were constructed specifically to put flesh on KERA's bones; the academic expectations are an excellent editing of the valued outcomes referencing goals 1, 2, 5, and 6; and the NAEP

---

[2]See, Lexington Herald-Leader, February 8, 1995, p. 1.

material can best be seen as adding useful detail to a heavily process-committed set of outcomes.

The academic expectations were adopted by the State Board for Elementary and Secondary Education and are a much clearer statement of the valued outcomes they replaced. The Panel did note five occurrences where the intent of the valued outcomes appeared to be distorted by the substitute academic expectation, but closer scrutiny of the valued outcome's rationale, related processes/concepts, sample assessment tasks, and general performance criteria affirmed that the academic expectation was true to the spirit of the supplanted valued outcome.

As an example of both a suspect translation and of the increased clarity, consider Goal 6.1. The value outcome is worded:

> Students address situations (e.g. topics, problems, decisions, products) from multiple perspectives and produce presentations or products that demonstrate a broad understanding. Examples of perspectives include: economic, social, cultural, political, historic, physical, technical, aesthetic, environmental, and personal.

The academic expectation for Goal 6.1 is:

> Students connect knowledge and experiences from different subject areas.

The greater clarity of the academic expectation is obvious, but we were concerned that multiple perspectives that different groups in our society hold was being slighted. Except for the word "personal" at the end of the value outcome, there is no hint of a perspective except one that is discipline (subject areas) based. For example, the general performance criteria for the valued outcome are the extent to which students:

- identify critical issues associated with the situation.
- use information from more than one discipline.
- research appropriately and included data from reliable sources.
- evaluate the consequences of solutions, conclusions, or actions from different perspectives.

2.3 Step 2. Writing the Assessment Exercises

A typical pattern is for Advanced Systems [the contractor] to write initial drafts of items based on brainstorm-produced suggestions from the content advisory committees. Item drafts are reviewed by the committees.

> Because of the short timelines between the planning and actual administration of the 1991-92 KIRIS assessment, many of the multiple-choice and some open-ended test items used in the first year were those that had been used by both Advanced Systems [the contractor] and NAEP in previous testing programs. (1991-92 Technical Report, 1993, p. 25)

The types of items or assessment formats used in the assessments are: open-response, multiple-choice, performance events, and portfolio. In this chapter, the Panel will consider examples of only open-response items.

Multiple-choice items are not illustrated here because, as mentioned above, they do not enter the accountability formula. The Panel believes, however, that whatever the other advantages and disadvantages of multiple-choice testing, including multiple-choice questions would make it more practical to provide reasonable coverage of the KIRIS subject area domains.

Performance events are not illustrated here because they are so few in number and do not contribute substantially to the

accountability index. The Panel would be remiss, however, if it did not mention the excellent Performance Event Design Manual for the KIRIS Assessment (March 9, 1995 draft).

Finally, the portfolio assessment is not discussed here because Chapter 4 is devoted to this component.

2.4 Step 3. Evaluating the Exercises

At this point in our review, the Panel looks at how the KIRIS items were evaluated, question by question. Two general methods are employed; the first consists of judges rating the items with respect to some desirable characteristic, and the second consists of analyzing the responses of students to learn about the quality of a question.

2.4.1 Step 3A. Rating the Items

> The linchpins of the test development process are the content advisory committees that assure the tests and scoring guides are appropriate for Kentucky students. (KIRIS Biennium I Technical Manual, 1995, p. 53)

The committees do deal with many important tasks, including developing and reviewing frameworks, reviewing items and scoring guides, reviewing field test results, helping to select items, etc. Kentucky's committees are quite large in number: 30-40 members for each subject. The technical manuals all tell us the committees include classroom and resource teachers, university professors, Kentucky Department of Education staff, and other stakeholders. They also briefly mention their training. But the Test Standards require more.

Standard 1.7

> When subject-matter experts have been asked to judge whether items are an appropriate sample of a universe or are correctly scored, or when criteria are composed of rater judgments, the relevant training, experience, and qualifications of the experts should be described. Any procedure used to obtain a consensus among judges about the appropriate specifications of the universe and the representativeness of the samples for the intended objectives should also be described.

The Panel suspects, but can't say for sure, that the content advisory and other committees are able individuals who are in leadership positions or are representatives of their constituencies. Nevertheless, experience and qualifications of the committee members should be reported.

The Panel is less confident about the procedures by which the judgments were obtained. Judgments are sensitive to the details by which they are elicited and by the procedures by which they are aggregated and reported. The Panel has no way of knowing the strength and amount of committee member support of the individual items.

Judgments are usually obtained with respect to three kinds of item characteristics. One is the extent to which the item is seen to belong to the category or cell of the subject matter framework it is intended to reference. A second is the degree to which an item may favor, be insensitive to, or present a stereotypical view of a particular group. The third is whether the item's answer is accurate and follows good item-writing techniques.

Item by Framework Category Match. Item by category judgments are solicited, at least by the third year. "When

committee members review test questions, they sort the items by content categories defined by the content guidelines and academic expectations" (KIRIS Biennium I Technical Manual, 1995, p. 60).

The Panel does not know how the ratings took place, whether individually or for the group as a whole, whether independently or not, whether the question was "Which category?" or "Does it fit this intended category?" Without these and other basic details about the procedure, the Panel does not know how to interpret the degree of fit, which we were not given either.

Perhaps evidence that items can be placed in the proper nook and cranny of its framework is not high priority for KIRIS. Stuart Kahl (1995) of Advanced Systems argues that:

> . . . categories within a domain need not be totally mutually exclusive, but extensive overlap makes the assignment of questions to categories nearly impossible and, if done, results in meaningless categories. (p. 5)

If it is correct that items are measuring much in common, and a finding in Section 2.4.2 to follow and inspection of the items reinforce this conclusion, then all this has important implications for the education of Kentucky's students. What the test is measuring is treated in Section 2.6.

Item Bias and Sensitivity Review. For at least two years, KIRIS has made use of a bias review committee:

> The role of this committee was to ensure that the content of the test was fair and equitable for all students, and that the tests contained no material that could be construed to be racially-, ethnically-, or gender-stereotyped or biased against any group. (1992-93 Technical Report, 1994, p. 3-1)

Again, the Panel has no information how this large (n=24 in 1992-93) group operated or what their findings were. We were

impressed with a 3-page set of sensitivity guidelines (No author, 1995) that identified a variety of issues Kentuckians considered sensitive and which should be avoided or addressed carefully on KIRIS items.

Accuracy and Item Quality Review. Items should be accurate, communicate well, and, especially for multiple-choice, avoid giving away the correct answer. A review by us of the 60 1993-94 common, open-response items revealed a number of defects that we were surprised passed through the contractor and Kentucky screening processes.

Recurring Problem. A frequent defect of the KIRIS items is to say too little, or worse, to mislead the student as to what is required for a good answer. This sparseness of information can take several forms.

Questions often do not specify the number of responses needed to score the maximum number of points possible. Questions often are worded, "Discuss several . . ." (e.g., Grade 12, Social Studies, Q. 32; Grade 8, Science, Q. 18). By providing information about the number of responses required (e.g., "Discuss at least three . . ." instead of "Discuss several . . ."), the student will be clearer about what the task is.

The distinction between score 4 and score 3 responses frequently involves the amount of detail in the student's answer, yet the question doesn't always state the need to give a detailed response (see Grade 4, Social Studies, Q. 22).

An example of a misleading task description with respect to the scoring criteria is Grade 4, Reading, Q. 12. Part B of the

question asks, "If you were having a bad day, which of these ways would you want a friend to treat you and why?" Score level 2 of the rubric states, "Student response . . . uses less significant examples from the poem to explain how he/she would want to be treated," even though the question never asks for examples from the poem itself. The "why" part could be answered solely from the student's experiences. The same comment applies to Grade 8, Reading, Q. 20, where the scoring rubric for Score 4 specifies "several well-supported reasons," but the question makes no mention that this is the objective.

Item-Specific Problems. Grade 4, Social Studies, Q. 33 asks students to make inferences about Native Americans based on unearthed artifacts. According to the scoring rubric, students must draw conclusions about causality based on weak correlational evidence. For example, they must conclude that this group of Native Americans traded with other Native Americans and early pioneers based on finding the artifacts of (b.) jewelry made with sea shells from the Atlantic Ocean and (c.) metal cooking pots. There are plenty of other plausible explanations for the presence of such artifacts. For example, the tribe may have migrated from the Atlantic Ocean area, or some members may have traveled there. The cooking pots could have been found at a failed pioneer settlement, or stolen, and so forth.

Grade 4, Reading, Q. 7: The question "Why do you think the Navajos retold this story about the dove?" is not specific enough. The scoring rubric specifies that the purpose of retelling the story is "teaching children a lesson not to be too

proud." However, nothing in the story indicates that this is its sole purpose. For example, the story could also be told in order to provide an explanation for the cooing sound of the dove, its red feet and its gray feathers. The question could be made clearer if it read: "What lesson does the story teach to the Navajo children?"

Other flawed items are mentioned in Appendix A (see especially page 11).

All measurement has error, and flawed items pass the inspection of even the most careful test developers. Occasional presence of such items rarely invalidates a test. If the 60 1993-94 common open-response items are representative of the quality of the open-response items in general, then greater attention needs to be given to the item development and review process.

2.4.2 Step 3B. Item Analysis

Step 3A reviews the judgments of experts about an item's quality; in this section of our report, we review how students answered the questions for evidence of item quality.

Design. Basically a test constructor has three ways to collect item-quality data:

1. in a separate field tryout before the use of the item in an assessment,
2. embedded in the operational test as a "experimental" item, and
3. embedded in the operational test and scored.

KIRIS uses designs 1 and 3, and has plans to move from design 1 to design 2. Each design has advantages and disadvantages. The Panel feels the item tryout designs, present and proposed, are reasonable.

In implementing design 1, according to Kahl (1995),

> New items are administered to 75 to 100 students via out-of-state field testing. The actual student responses to field-tested items are reviewed by committee members to determine if items produced a range of responses, if the responses reveal general problems in understanding questions or portions of questions, and if the questions can be answered effectively in the projected testing time. (p. 4)

More sophisticated analyses would be inappropriate, in our view, given the small number of students used in the field tryouts.

In implementing design 3, a variety of item difficulty and discrimination statistics was computed. Item difficulty is a measure of how easy a test question is. For open-response items, it is the average score earned on the item's 0-4 scale. Item discrimination is whether those who get the item correct or do well on it also do well on some appropriate measure -- typically a subject area score. These statistics are described on p. 57 of the KIRIS Biennium I Technical Manual (1995, March 1 Version) and in our view were proper ones to compute.

Results. The Panel was not told the "rules" for keeping versus changing/discarding an item under design 1, but, given the qualitative element in the criteria, it would have been difficult to develop strict rules.

In implementing design 3, several very hard open-response items were detected, but even these discriminated well. Only one

of the 875 item-total correlations[3] for open-response items had a discrimination index of less than .30. The median values were estimated to be .49, .55, and .55 for Grades 4, 8, and 12 and .52, .59, .48, and .56 for Math, Reading, Science, and Social Studies, respectively. Such consistently high values signal the presence of a strong common factor that underlies the set of open-response items for a subject matter. The implications of this finding are discussed in Section 2.6.

Typically, under design 3, it is recommended that items with negative discrimination values not be counted in the scoring. Evidently, no such item "neutering" has taken place in KIRIS to date.

2.5 Step 4. Selecting Items

Steps 3A and 3B dealt with the evaluation of assessment questions, one item at a time. Here, the Panel is concerned with the collection of items that make up an assessment instrument for a given year. Specifically, the Panel is concerned with the distribution of subject matter content, balance, and weighting, with respect to the frameworks. Recall that the frameworks are to be blueprints for test construction. Test constructors call the congruence between test content and frameworks content validity, and the Test Standards require this property for tests like those found in KIRIS:

---

[3]Total is computed on the common items only. For common items, correlations have been corrected for item-total contamination.

Standard 1.6

> When content-related evidence serves as a significant demonstration of validity for a particular test use, a clear definition of the universe represented, its relevance to the proposed test use, and the procedures followed in generating test content to represent that universe should be described. When the content sampling is intended to reflect criticality rather than representativeness, the rationale for the relative emphasis given to critical factors in the universe should also be described carefully.

The Panel is unable to comment upon the intended distribution of items into their content and process categories since this distribution, with one exception, is not publicized. The exception is for type of reading passage. Literary selections, for example, are to make up 50% of the reading passages for Grade 4 but only 30% for Grade 12.

The Panel was able to comment upon the fact that not all academic expectations were covered initially, even allowing for the Board of Education's exclusion of the valued outcomes self-sufficiency and responsible group membership (learner goals 3 and 4). Depending on grade level, only 2 of 12 of the academic expectations under learner goal 1 (basic communication and mathematics skills) and two-thirds of the academic expectations under goal 2 (core concepts from the disciplines) are covered by the open-response items in the 1993/94 assessment. Some of the communication skills not covered do not typically have formal programs associated with them; entire disciplines (arts and humanities, vocational studies, and practical living) were not covered initially instead of allowing of gaps in all the disciplines. The KIRIS Biennium I Technical Manual puts the reason why more coverage did not take place this way:

> Because the scope of reform was so comprehensive, it would have been unfair to hold all schools in Kentucky immediately responsible for these academic expectations which require special programs or facilities. (p. 20)

But how about content coverage over the academic expectations of the disciplines covered -- are the open-response items well distributed over the specific expectations? Based on tables supplied by Advanced Systems,[4] the Panel would judge yes, although the distributions are not even. The observations of the classifier of the items is insightful:

1. . . . the open-response questions tend to be broad and often belong to multiple categories (i.e., address multiple academic expectations).

2. This "double counting" of questions might lead to vastly disproportionate numbers of items in an area. This is especially true in science, where large numbers of items fall into the patterns and systems categories. (Some would agree that science is the study of patterns.) Data and procedures (operations) in mathematics also tend to "capture" many areas.

3. Some categories are relatively indistinct from one another by definition -- e.g., measurement and data, democratic principles and political systems, social systems and cultural diversity.

4. Some academic expectations may "define" a smaller portion of a content domain than others -- e.g., mathematical structure.

Not all academic expectations are assessed and those that are included are weighted unequally. The Panel believes that these deviations from complete coverage and equal weighting are acceptable, but would have preferred that the intended weighting be communicated to the schools.

---

[4]Personal correspondence from Lisa Ehrlich to Jason Millman, May 15, 1995. Author unknown.

2.6 So What Are These Cognitive Tests KIRIS Has Created?

Before attempting an answer, the Panel wants to separate this question from others worth asking.

- How well are the academic expectations, both collectively and individually, covered by the cognitive tests? That is a content validity question addressed in section 2.5 above.
- How reliably and fairly do the cognitive tests identify the stronger and weaker schools? This question is addressed in the technical material to follow, particularly in Chapter 4.
- To what extent have the cognitive tests (a) affected what is taught in the schools, and (b) resulted in improved learning? The (a) part of the question is beyond the scope of this review; the (b) part of the question is addressed in Chapter 8.

Another way of asking the question that heads this section is, What can we say about a student who scores high in the assessment in subject X? We can only speculate, but our hypotheses can often by informed by close scrutiny of the measurement tasks and by testing hypotheses about what is being measured by collecting relevant data.

2.6.1 Judgments Based on Item Scrutiny

In numerous announcements, the Kentucky Department of Education has claimed that KIRIS measures higher order thinking skill rather than narrow facts. One of the justifications for

moving from multiple-choice to open-response items is the association of multiple-choice items with the recognition of facts and open-response items with producing, applying, and other skillful actions. Our inspection of the science, social science, and mathematics common items for 1993-94 agrees with this characterization.

Our inspection also reveals, especially for the science and social studies items, a striking de-emphasis of subject matter knowledge. It was almost as though the student did not have to attend science or social studies classes to be able to answer the questions. Consider this 8th grade item:

> Manifest Destiny was a belief in the mid-1800s that the United States was destined to extend its borders from the Atlantic to the Pacific. Discuss whether fulfillment of this belief was a success. Explain the effects that Manifest Destiny had on Native Americans at that time.

One Panel member, dazzled by the words Manifest Destiny, concluded that this item clearly requires knowledge of social studies (history). Closer scrutiny shows that the concept was defined for the student and that the top two score categories did not require the student to know or discuss whether fulfillment of the belief was a success. It would appear now that the item does not require subject-matter knowledge, illustrative of the recurring item-construction flaw of lack of congruence between question and answer (see section 2.4.1):

> **Answer indicates a good understanding of American history and the social and cultural implications of the Manifest Destiny concept. Student uses relevant examples and may use modern day examples as well.**

The Panel member classified 7 of the 15 social studies items, including the one about Manifest Destiny, as subject-matter based and 4 of the 15 science items as subject-matter based. Although others would undoubtedly rate the items differently, the trend away from assessing what has been traditionally taught in the classroom would show up.

In mathematics, very difficult and typically not taught reasoning problems, like the post office and factorial questions discussed in Appendix A, share the same assessment form with items about graphs and areas.

Our discussion about what the KIRIS cognitive instruments are measuring is not to advocate a particular bias. The "right" content is not a technical measurement quality question about which we claim expertise. Rather, our motivation is principally two-fold. First, we believe it extremely important that stakeholders have a good sense of what the assessment instruments measure. Second, we should heed the warning of the Test Standards:

> Standard 3.14
>
> The sensitivity of test performance to improvement with practice, coaching, or brief instruction should be studied as part of the developmental research, especially on performance tests that use an unfamiliar response mode, such as computer-administered tests.

It may be that the assessment is measuring a much narrower domain of skill than is claimed, a domain that, because it is more limited, may actually be easier to teach, be repetitive over the grades, and be less rich educationally. Learning how to write, learning how to answer the popular "What if" science

questions, learning how to give examples in answering social studies questions would be part of this domain.

2.6.2 Data Relevant to Testing Hypotheses

Analyzing existing assessment data or data collected as part of studies especially designed for the purpose is a second way to understand better what the cognitive tests are measuring.

To its credit, the Kentucky Department of Education has conducted or sponsored studies with the purpose of validating the meaning of the tests. Many state testing programs do not take this initiative. We encourage the Kentucky Department of Education to continue these efforts with particular attention to studies that will refute or support claims that the test is measuring unwanted factors.

One way to proceed is to hypothesize a factor the test is measuring, decide what would count as evidence that the hypothesis is correct or not, collect and analyze such relevant evidence, and come to a conclusion about the viability of the hypothesis.

Here's an extreme claim. The open-response science and social studies tests measure the same thing--general thinking ability. Subject matter achievement is not a factor. If this claim were true, then perfectly reliable open-response tests like the science and social studies tests should correlate perfectly. In fact, the correlations are estimated to be .63, .56, and .63 for grades 4, 8 and 12 respectively. The tests clearly include factors specific to each.

A less extreme claim would be that these open-response tests have little to do with measuring achievement in the subject matter and more to do with learning general higher order thinking skills. Relevant data are available if one assumes that the multiple-choice tests do measure subject-matter achievement. In that case, the hypothesis could be checked by comparing each of the two open-response measures (i.e., science and social studies) with the two corresponding multiple-choice tests. If the hypothesis is true, the correlations, corrected for unreliability, with the same subject multiple-choice test would be about the same as with the other one. If the hypothesis were not true, the underlined correlation of each pair shown below should be higher. Since that is the pattern of the data, subject matter achievement appears to be a contributing factor to these tests.

	Grade 4		Grade 8		Grade 12	
	M-C Sci.	M-C S.S.	M-C Sci.	M-C S.S.	M-C Sci.	M-C S.S.
Open-response Science:	<u>.62</u>	.59	<u>.61</u>	.58	<u>.71</u>	.65
Open-response Soc Stu:	.51	<u>.56</u>	.46	<u>.53</u>	.53	<u>.62</u>

Is writing and reading a factor in these tests as the Panel has suggested? If the open-response tests in science and social studies are essentially writing tests rather than tests measuring achievement in their respective subject matters, then we would expect perfectly reliable open-response item sets in a subject area to correlate more with on-demand writing ability reliably measured than to correlate with a perfectly reliable multiple-choice test in the subject matter. The assumption is

again that the multiple-choice tests do measure subject matter achievement.

The results are shown below. Although the writing on demand task usually does not correlate as heavily as the corresponding subject matter multiple-choice test, the correlations for the open-response reading test are comparable to those for the multiple-choice test. That doesn't mean reading is important, however, as the open-response reading test does require writing and shares other method variance with the science and social studies tests.

	Grade 4			Grade 8			Grade 12		
	M-C	w	r	M-C	w	r	M-C	w	r
science open-response	62	39	61	61	36	50	71	47	63
social studies	56	43	64	53	45	60	62	45	64

The Kentucky Department of Education might commission any number of studies to learn more about what its tests are measuring, that is, to help to establish the construct validity of its tests. The following are some examples intended only to be illustrative of the possibilities: coaching studies -- e.g., can students be taught in a short period of time how to do well on "what if" kinds of questions? Do students who have studied and know about Manifest Destiny, as one example, score better on the Manifest Destiny item? What is the pattern of relationships that show up when open-response items are divided into subject-matter-based and not subject-matter-based (see section 2.6.1)? As the focus expands to other than open-response measures, the possibilities for studies multiply.

References

No author. (1995, March 9). Guidelines for handling sensitive issues in Kentucky's state assessment development. In Performance event design manual for the KIRIS assessment (pp. 15-17). (Draft)

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington: APA.

Kahl, S. (1995). Specifications of content for the KIRIS Assessment.

Kentucky Department of Education. (1993). KIRIS 1991-92 technical report. Frankfort, KY: Author.

Kentucky Department of Education. (1994). Position paper on recommended changes in the KIRIS Assessment and Accountability Program.

Kentucky Department of Education. (1995, March 1). KIRIS biennium I technical manual. Frankfort, KY: Author.

CHAPTER 3

Do the Accountability and Assessment Scores Have Adequate Reliability?

3.1 Central Questions and Principal Findings

The central reliability questions for KIRIS differ for the major uses of the assessment results. The central question for the school accountability system is: How likely is it that the classification of a school into one of the accountability categories (Eligible for Reward, Successful, Improving, Decline, In Crises) would change if a second assessment of the same kind at about the same time had been conducted? Consistency of decisions for different cohorts of students, different collections of assessment tasks designed to measure the same academic expectations, and different raters of student performances needs to be evaluated to respond to this question.

For the assessment uses of scores the central question is: What is the likely margin of error in reported scores? At the school level, this requires an estimate of the probable variability of school scores on each cognitive learner goal as a result of changes in student cohort, tasks, and raters. At the individual student level, this requires an estimate of the probable margin of error in student scores due to task and rater sampling.

It is quite possible that a test that was judged to have adequate reliability for the assessment system would have inadequate reliability for the accountability system, or vice versa. In each case, adequacy depends on the seriousness of errors as well as their likelihood of occurrence. The

seriousness of errors depends, in turn on the actions that are taken based on the student or school scores. Hence, reliability and measurement error issues for accountability and assessment uses will be addressed in separate sections. Because of the unique features of portfolio assessments, reliability issues regarding portfolios are addressed along with other technical issues in Chapter 4 which is focused specifically on the portfolio results.

The major conclusions based on the analyses described in this chapter may be summarized as follows:

School Accountability

1. Although the data are not available to calculate exact numbers, as the result of measurement error, a substantial number of schools are assigned to a reward/sanction category other than the one deserved. Small schools are most often misclassified. Misclassification errors could be reduced by administering a larger number of test questions to each student. Misclassification errors could also be reduced by including the multiple-choice items in the calculation of accountability indices.
2. To date, there is only an incomplete basis for evaluating the likelihood that a school placed in a particular accountability category is misclassified. The only available analysis is sensitive to only one of the two largest sources of error and therefore underestimates the likelihood that schools will be

misclassified. Analyses reported in the March 1, 1995 draft of the KIRIS Biennium I Technical Manual do not directly address issues of misclassification of schools in the accountability system. The most relevant analyses, reported on April 11, 1995 by Kingston and Dings (1995) and summarized by Kingston (April 27, 1995), provide information on school misclassification probabilities as the result of one source of error (sampling of student cohorts) but do not include error as the result of task sampling. In other words, the latter analyses treat the collection of tasks in the assessment each year as if they were perfectly reliable.

3. Some summary claims about KIRIS reliability based on conventional indices appropriate for some test interpretations, but that are largely irrelevant to the primary issue of misclassification (e.g. reliabilities greater than 0.90), provide a misleadingly positive characterization of KIRIS reliability.

4. A policy judgment is required to determine whether the estimated error rates are acceptably low.

School and Student Assessment

5. Student scores on the KIRIS open-response tasks for each subject are generally less reliable than would typically be expected for standardized tests. Whether the level of reliability is considered adequate depends on uses made of scores. It may be considered adequate

for some low-stakes uses but not for high-stakes uses with serious consequences for individual students.

6. Current student reports do not convey information about the likely margin of error associated with scores. As required by the Test Standards, information about standard errors of measurement, possibly in the form of score bands, should be included in the reports of student results.

7. School-level assessment results are typically more reliable than individual student scores, but the March 1, 1995 draft or the KIRIS Biennium I Technical Manual gives insufficient information to be sure if this typical finding is true for KIRIS. Estimates of the standard errors of measurement associated with school scores for each subject area should include error due to sampling of students, tasks, and raters. These are currently lacking.

8. Reliability of both the student and school subject area scores could be improved by using both multiple-choice items and open-response tasks to obtain scores.

3.2 The Concept of Reliability and Its Application to KIRIS

There is always some degree of uncertainty associated with the scores on an assessment and with decisions that are based on those scores. Students know that their performance on a classroom examination, for example, can be affected by the particular topics emphasized and the specific questions that a

teacher decides to include in the examination. Performance may also differ from one occasion to another as a result of changes in both situational characteristics (e.g., assessment administration conditions, distractions, or disruptions) and personal characteristics of test takers (e.g., student health, fatigue, mood, or anxiety). Such extraneous influences on assessment scores contribute to uncertainty or "measurement error." Scores that would be obtained if all such extraneous influences were removed are referred to as "true" scores.

Public opinion polls commonly report margins of errors associated with percentages, e.g., 40% of registered voters say they will vote for candidate X with a margin of error of 3%. This is a way of conveying uncertainty due to sampling of respondents. Indications of uncertainty associated with assessment results are also needed. With assessments, however, there are several potential sources of uncertainty that need to be taken into account.

Sampling of assessment tasks and of raters who score student responses contribute to the uncertainty of individual student scores. Those are also sources of uncertainty in assessments used for measuring school performance, but the degree of uncertainty due to these influences may be reduced at the school level as a result of averaging and certain assessment design features. The particular students enrolled in a school in a given year also contribute to the uncertainty associated with school-level results when they are used to draw conclusions about school quality or improvement over time.

Investigations of reliability of assessment scores and the dependability of decisions based on those scores provide a means of evaluating the degree to which the scores and decisions are free from measurement error. Validity is the most important consideration in evaluating the uses of test scores. Reliability, however, is closely connected to validity. High reliability (which implies small errors of measurement) does not guarantee high validity, but low reliability limits the degree of validity that can be obtained. Thus, systematic analyses of reliability, the sources of measurement error, and the magnitude of errors produced by those sources are important aspects of the overall evaluation of the appropriateness and adequacy of the uses and interpretations of assessment results required in judging validity (AERA, APA, & NCME, 1985, p. 19).

The specific uses and interpretations of assessment results need to be taken into account in evaluating measurement error. Differences among cohorts of students, for example, may be an important source of measurement error when student assessment results are used to characterize groups or make decisions about schools but not when scores are used to describe individual students. On the other hand, use of multiple assessment forms in a matrix design may reduce measurement error for school-level aggregations of scores but increase the relative magnitude of error when reporting results for individual students.

Since the KIRIS accountability results depend on changes in the scores obtained by students from the baseline to subsequent years, errors associated with change scores are relevant. In

general, change scores are less reliable than measures of current status. Moreover, systematic errors associated due to imperfect equating of forms from year-to-year must be considered. Equating issues are addressed in Chapter 5. The Panel notes here, however, that equating errors would add to the uncertainty due to measurement error.

Guidance for addressing reliability and measurement error questions is provided by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985). The Standards were written before the current trend toward a greater reliance on performance-based assessments began. With the exception of Chapter 12, the Standards also focus primarily on the use and interpretation of test results for individual test takers rather than the use of aggregate results for making decisions about schools or other types of institutions. Nonetheless, many of the principles articulated in the Standards are quite relevant to the uses of assessment results in KIRIS. This is clearly the case with regard to the standards for reliability and errors of measurement (Chapter 2) which emphasize the need to attend to "those sources of error of greatest concern for a particular use and interpretation of a test" (p. 19).

Professional judgment is always required in applying the Standards for a particular use or interpretation of test results. Since the uses and interpretations of assessment results in KIRIS differ in significant ways from those most commonly illustrated in the Standards, we believe it is important to articulate the

key principles that we think are applicable to judging the reliability of KIRIS.

(1) The primary purpose of analyses of reliability and measurement error is to provide information on the likely degree of uncertainty of particular interpretations and uses of scores based on the assessment results. For example, if a school receives a 4th grade cognitive index score in reading of, say, 25, how certain is it that the index would be between, say, 20 and 30 if different reading passages had been administered or a different cohort of fourth graders had been enrolled in the school that year?

(2) As is stated in the introduction to Chapter 2 of the Standards, "the importance of a particular source of error depends on the specific use made of a test" (p. 19). For example, with regard to KIRIS, errors of measurement that might make results quite unsatisfactory for making important comparisons among individual students might be relatively unimportant (or of trivial size) for determining the reward/assistance status of a school based on the KIRIS Accountability Index. For example, the administration of different performance events to different students or the matrix sampling of open-ended items, while increasing uncertainty of comparison of individual students with each other or to a common standard, may actually decrease uncertainty of school-level scores by increasing the total number of performance events or open-ended items administered to samples of students within a school.

(3) Where assessment results are used to classify objects of measurement for purposes of differential treatment (e.g., allocation of rewards to schools), analyses of reliability and measurement error should provide a means of estimating the likelihood that the classification would change as a result of measurement error. For KIRIS, for example, there is a need to obtain estimates of the probability that a school that is identified to receive a reward (sanction) would not be so identified if assessments had been based on different collections of tasks, different raters of student performances, or different cohorts of students.

(4) "Reliability coefficients" provide less valuable information than estimates of components of error variance, standard errors of measurement, or misclassification probabilities. Indeed, reliability coefficients are of greatest use with traditional normative comparisons among individuals where variability is expected to be relatively stable--a use that is **quite unlike** the uses and interpretations of results for KIRIS where the emphasis is on comparison of results to absolute standards and substantial changes in the distribution of performance are expected as a result of the educational reform efforts associated with KERA.

3.3 Analysis of Dependability of Accountability Decisions

Chapter 13: "Reliability of KIRIS Accountability Indices" of the March 1, 1995 draft of the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995b) consists almost

entirely of reproductions of tables and text from the KIRIS 1992-93 Technical Report (Kentucky Department of Education, 1994). Specifically, Tables 13-1, 13-2, 13-3, 13-4, and 13-6 contain the information presented in Table 9-1, 9-2, 9-3, 9-4, and 9-5 of the KIRIS 1992-93 Technical Report. They were included in the KIRIS Biennium I Technical Manual in order to bring together much of the reliability results in a single place. Only Table 13-5 of the Biennium I draft report contains new information. That table reports generalizability coefficients based on four different measurement models (that, in technical terms, consider tasks and students as either fixed or random) for the 1992-93 and the 1993-94 Accountability Indices. The other tables report only for the 1991-92 and 1992-93 assessments. For those years, separate estimates for schools classified as small middle, or large are provided for generalizability coefficients and standard errors of measurement with and without the inclusion of writing portfolios on the Total Cognitive Index. Reliability estimates for individual student scores are also provided separately for open-response and multiple-choice items in 1992-93 and/or open-response items only in 1991-92 (see section on assessment use of scores below).

In addition to the results in the KIRIS Biennium I Technical Manual the Panel was provided with a February 1995 News Packet, Celebrate the Progress: 1992-94 Kentucky Accountability Results Summary News Packet (Kentucky Department of Education, 1995a, February) where some 1993-94 reliability results are reported. The panel also received some partially labeled computer printouts

reporting generalizability coefficients and variance components, a report by Kingston and Dings dated April 11, 1995 that presented decision consistency estimates based on a simulation, and an April 27, 1995 report by Kingston on accuracy of accountability decisions.

Some of the reliability analyses are complicated by the need to convert the scores on individual open-response tasks from the assigned scores of 0 to 4 to the metric of the accountability index (Novice = 0, Apprentice = 40, Proficient = 100, and Distinguished = 140). Considerable thought and effort obviously went into the development of that conversion procedure and to the development of estimates. The analyses appear to have been done competently. The Department of Education and its contractor, Advanced Systems in Measurement and Evaluation, have also been responsive to some earlier criticisms about the conduct and reporting of reliability results. The addition of results where tasks and/or students are treated as random, for example, is responsive to concerns raised in the January 1995 report of the evaluation conducted by The Evaluation Center, Western Michigan University Evaluation Center for the Kentucky Institute for Education Research (The Evaluation Center, Western Michigan University, 1995) and to preliminary feedback from this panel. The April 11 Kingston and Dings paper and the April 27, 1995 Kingston paper also reflect efforts by the Department of Education to respond to questions raised by the panel about decision consistency.

If KIRIS were a conventional testing program, and if information parallel to that reported for earlier years was added for the 1994-95 assessment, then the treatment of reliability and measurement error in the KIRIS Biennium I Technical Manual might be judged to be adequate. Where needed, sophisticated analytic tools were used to place item results on the scale of interest (e.g., the use of item response theory and the regression rescaling of open-ended 1-4 responses to the 0-140 accountability scale). The Panel would urge greater emphasis on standard errors of measurement, the inclusion of estimates of variance components, and the inclusion of school-level results separately for each cognitive area assessed.

But the school accountability component of KIRIS is not a conventional testing program and the results reported on the March 1, 1995 draft of the KIRIS Biennium I Technical Manual unfortunately do not directly address the most important issues of reliability and measurement error for the uses that are made of the accountability index.

The conventional, albeit largely irrelevant to the accountability system, orientation is evident in the last sentence of the second paragraph of Chapter 13 where it is stated: "The critical question is whether the reliability of the overall index is as high as professional judgment dictates is required for a high-stakes test--typically, above .90" (Kentucky Department of Education, 1995b, March 1, p. 211). That is NOT the "critical question" for use of the accountability index. Indeed, the emphasis on reliability coefficients greater than .90

is only indirectly related to the central policy question for accountability uses of results and a focus on the reported reliability coefficients is potentially misleading.

Much more critical than the emphasized reliability coefficients is the probability that a school in a given reward/sanction status at the end of Biennium I has been appropriately classified. Or to be more specific, the analysis of measurement error needs to provide the basis for providing policy makers and the public with an indication of the likelihood that a school identified for receipt of a financial reward based on the comparison of 1992-93 and 1993-94 results to the school's 1991-92 results would have failed to receive the reward if different samples of tasks, different raters of student responses, and/or different cohorts of students had been used to determine the school's reward/sanction status. Similarly, policy makers, educators, and the public need to have a basis for judging the likelihood that a school identified as "In Decline" may have been so identified as a result of measurement error when it did not deserve to be so designated.

As noted above, the Department has conducted additional analyses in response to the Panel's requests at its March 2-4 meeting for information about misclassification probabilities. A preliminary report on the results of some of those analyses was presented in a paper by Kingston and Dings at the annual meeting of the American Educational Research Association on April 19, 1995. Those results are also described in an April 11, 1995 draft of that paper entitled "Estimating the Accuracy of Complex

School Accountability Decisions." Selected results from the latter paper are presented in a paper entitled "Accuracy of Kentucky School Reward Decisions" dated April 27, 1995 by Kingston.

The approach used by Kingston and Dings provides estimates of percentages of schools with specified characteristics (e.g., those "that truly declined" that were "categorized as Reward"). The estimates are based on a simulation that draws random samples of students of various sample sizes from a group of all students with available scores from schools that had at least 60 students each year. Sampling was done with replacement and results for the groups of 60 students each year were used to simulate the "true" values for a school. Fifty replications were used to obtain the estimates for each school in the simulation. Samples sizes of 12, 24, 48, and 96 students were used. Analyses were conducted with and without the performance events included and with and without portfolios included. Analyses were also conducted with a single year of baseline data to mirror Accountability Cycle I and with two years of baseline data to reflect plans for Accountability Cycle II.

The simulation is a reasonable approach to estimating the probability of making certain types of errors and of correctly classifying schools that should receive rewards and those that should be labeled "In Crisis." However, the estimates only take into account uncertainty due to differences in cohorts of students. They do <u>not</u> take into account the effects of sampling of tasks. Consequently, the estimates of decision consistency

and misclassification probabilities give the appearance of more certainty of results than would be provided if the effects due sampling of tasks and raters had also been taken into account in the simulation.

It clearly would be better to include the sampling of tasks as well as the sampling of students as potential sources of uncertainty, however, we agree that the sampling of students is an important, possibly even the most important single source of uncertainty. The Panel also recognizes that conducting a simulation that would include an evaluation of uncertainty due to sampling of raters, tasks, and students would make the analytic work even more complicated than it was. As it is, the simulation represents a substantial effort on the part of the Department of Education to respond to requests for information about misclassification probabilities.

The accountability system was designed to categorize schools into one of five categories:

1. Eligible for Rewards (accountability index exceeds improvement goal by one point or more and at least 10 percent of student at the novice level move to apprentice or higher),
2. Successful, (accountability index equals improvement goal or exceeds it by less than one point),
3. Not Meeting Threshold, sometimes referred to as Improving (accountability maintains baseline or improves but does not reach the improvement goal),

4. In Decline (accountability index drops by less than five points), and
5. In Crisis (accountability index drops by five points or more).

By action of the 1994 Kentucky General Assembly the implementation of the In Crisis category was postponed and schools that would have been placed in that category were included in the In Decline category. Analyses, therefore that include the In Crisis category should be interpreted as indications of what would apply if that category had been implemented and should be useful in considering possible policies involving uses of this category in Biennium II.

The Kingston and Dings paper estimates eight percentages corresponding to the following eight questions for each condition of the simulation.

1. What percent of schools that have been categorized as **Reward** have truly improved performance?
2. What percent of schools that have been categorized as **Reward** have truly met or exceeded their Improvement Goals?
3. What percent of schools that have improved their performance have been categorized as **Reward**?
4. What percent of schools that have truly declined have been categorized as **Reward**?
5. What percent of schools that would have been categorized as **Crises** have truly had their performance decline?
6. What percent of schools that would have been categorized as **Crises** have truly had their performance decline by 5 or more points?
7. What percent of schools that have declined would have been categorized as **Crises**?

8. What percent of schools that have truly improved would have been categorized as **Crises**? (p. 13).

Several points should be made regarding these percents:

1. The so-called "true" results (e.g., truly improved performance) ignore measurement errors due to task sampling.
2. The improved designation includes the Eligible for Reward, Successful, and Not Meeting Threshold categories and the declined category includes the In Decline and In Crisis categories.
3. Schools said to have truly met or exceeded Improvement Goals includes ones where "true" performance would lead to a categorization as either Successful or Eligible for Reward.
4. The ideal would be 100 for some of these percents (e.g., percents corresponding to Questions 2 and 6) and zero for others (e.g., percents for Questions 4 and 8).
5. There is a tradeoff among these percents. For example, changing rules to increase the percent of schools that truly improved their performance that would receive rewards (percent 3) would lead to a reduction in the percent of schools that received rewards that truly improved (percent 1).
6. Although eight percentages are already a large number to consider in reaching an overall judgment regarding the acceptability of the accuracy of classification of schools, still other percentages might also be relevant (e.g., percent of schools that "truly" had their

performance decline by 5 or more points that would not be identified as In Crisis or the percent of schools that "truly" had their performance decline by less than 5 points that would be identified as In Crisis).

7. The number of classification errors would be greater if the focus was on schools in the Improving (met or exceeded baseline, but did not meet improvement goals) and In Decline (school score declined, but by fewer than 5 points) categories rather than on the Eligible for Reward and the In Crisis categories.

As previously noted, the percentages reported by Kingston and Dings are more favorable than they would be if task sampling was included as a source of error in the simulation analyses. If only one source of error is considered, however, the results suggest that one can have a high degree of confidence that schools that received rewards truly had increases in student performance on the particular tasks as scored by the particular raters each year, albeit not necessarily enough to meet or exceed their improvement goals by one point or more. Few schools that had declines in student performance would be misclassified into the reward category, and those relatively rare misclassifications would be expected to occur only for small schools. The results for the In Crisis category are less favorable. It is estimated (see Table 13 of Kingston and Dings), for example, that only one in two elementary schools with school size of 24 that would have been classified as In Crisis, "truly had their performance decline."

The April 27, 1995, paper, "Accuracy of Kentucky School Reward Decisions" reports results for only three of the eight questions and associated conditional percentages reported by Kingston and Dings. These are Questions 1 (percentage of schools categorized as reward that truly improved), 2 (percentage of schools categorized as reward that truly met or exceeded improvement goals), and 4 (percentage of schools that truly declined that were categorized as reward). For Biennium I, the summary estimates of percentages, weighted to reflect the distribution of school size, are as follows (see Table 5 of April 27, "Accuracy" report):

	Grade		
Percentage	4	8	12
1. Truly improved given reward received	100.0	100.0	100.0
2. Truly exceeded improvement goals given reward received	88.4	84.5	66.3
4. Reward received given truly declined	0.5	0.2	0.2

The ideal would be 100.0 for Percentages 1 and 2 and 0.0 for Percentage 4. Percentages 1 and 4 clearly paint a quite favorable picture. Percentage 2 is less than 100.0 at each grade, but perfection cannot be expected since schools that are just below the minimum required by their improvement goals will always have a sizable probability of being misclassified.

Although the results in the Kingston (1995, April 27) paper look quite positive, they do not tell as complete a story as can be seen in the Kingston and Dings paper (1995, April 19). For

example, percentage 5, in the Kingston and Dings paper is an estimate of the percent of schools that would have been classified as In Crisis that have truly had their performance decline. The ideal for that percent would be 100.0. The estimates weighted for school size as was done in the April 27, 'Accuracy" paper are shown below.

	Grade		
Percentage	4	8	12
5. Truly declined given that school would have been classified In Crisis	66.4	69.8	86.8

These results are clearly less favorable than those selected for the Kingston (1995, April 27) report. Whether such percentages are acceptably high is a policy judgment, but in making that judgment it will be important for policy makers to understand that the odds that a school that would have been classified as In Crisis at Grade four truly declined are only two to one. In other words, 1 in 3 schools that would have been classified as In Crisis at Grade 4 would actually have improved, or at least shown no decline. Inclusion of percentages such as 5 in the April 27, 1995, "Accuracy" report would also help highlight the fact that there will be a tradeoff among the percentages estimated by Kingston and Dings. Since the In Crisis category was not implemented in Biennium I the percentages shown above may be of less interest in evaluating 1994 results than percentages that show the number of schools classified as In Decline that truly improved. The In Crisis results may deserve greater attention in planning for Biennium II and beyond, however.

The summary table in the report by Kingston (1995, April 27) does not highlight the high likelihood of misclassification errors for small schools. The appendix gives results for schools with 12, 24, 48, or 96 students in a grade. According to Kingston (1995, Table 4) 12.5 percent of the elementary schools have less than 24 students at Grade 4. Thus, for one elementary school in eight the results are less favorable than those reported for elementary schools of size 24 in the Appendix of the Kingston report (1995, April 27), or more completely in Table 13 of the Kingston and Dings paper (1995, April 19).

A more complete analysis would not only include the influence of task and rater sampling on the misclassification errors, but would present a full breakdown of the implicit five-by-five table of true classifications (REWARD [truly exceeded improvement goal by one point or more], SUCCESS [truly exceeded improvement goal by less than one point], IMPROVING [truly met or exceeded baseline, but did not meet its improvement goals], DECLINE [score truly declined, but by fewer than 5 points], or CRISIS [scored truly declined by 5 or more points]) by the corresponding observed classification (reward, success, improving, decline, crisis). The following Table could be used to compute a variety of possible errors in classification.

True	Observed Classification					
Classification	reward	success	improving	decline	crisis	Row Total
REWARD	X11	X12	X13	X14	X15	R1
SUCCESS	X21	X22	X23	X24	X25	R2
IMPROVING	X31	X32	X33	X34	X35	R3
DECLINE	X41	X42	X43	X44	X45	R4
CRISIS	X51	X52	X53	X54	X55	R5
Col. Total	C1	C2	C3	C4	C5	

In terms of the entries in the above table, the eight percentages estimated by Kingston and Dings are as follows:

1. percent(REWARD, SUCCESS or IMPROVING given reward) = (X11 + X21 + X31)/C1
2. percent(REWARD or SUCCESS given reward) = (X11 + X21)/C1
3. percent(reward given REWARD, SUCCESS or IMPROVING) = (X11 + X21 + X31)/(R1 + R2 + R3)
4. percent(reward given DECLINE or CRISIS) = (X41 + X51)/(R4 + R5)
5. percent(DECLINE or CRISIS given crises) = (X45 + X55)/C5
6. percent(CRISES given crises) = X55/C5
7. percent(crises given DECLINE or CRISIS) = (X45 + X55)/(R4 + R5)
8. percent(crises given REWARD, SUCCESS or IMPROVING) = (X15 + X25 + X35)/(R1 + R2 + R3).

With the full table, evaluators could estimate types of misclassification errors that might be of particular concern to specific groups of schools. For example, the percent(IMPROVING given decline) would be estimated by X34/C4.

The analysis of the dependability of the accountability uses of KIRIS leads to the following conclusions:

1. Analyses, reported in April 19 by Kingston and Dings (April 19, 1995) and summarized by Kingston (April 27, 1995), provide information on school misclassification probabilities as a result of one source of error (sampling of student cohorts) but do not include error as a result of task sampling.
2. Those results suggest that almost all schools designated as Eligible for Reward showed some improvement and the majority of them exceeded their Improvement Goals.
3. Misclassification probabilities are higher for schools designated as In Crisis than for schools designated as Eligible for Rewards.
4. A policy judgment is required to determine if the error rates reported by Kingston and Dings are acceptably low. In making such judgments, it should be noted that the analyses yield results that appear more favorable than they would if the effects of sampling tasks had been included in the analyses.

3.4 Analysis of Reliability of Assessment Uses of Scores

Chapter 13 of the March 1, 1995 draft of the KIRIS Biennium I Technical Manual reports student-level coefficient alphas based on open-response items only in 1991-92, open-response items only in 1992-93, and for multiple-choice items only in 1992-93.

Parallel coefficients for the 1993-94 assessment were not included in the March 1 draft of the reliability chapter (Chapter 13) but can be found in the diagonals of the correlation matrices presented in the validity chapter (Chapter 14). The alpha coefficients are reported separately by grade and subject (Reading, Mathematics, Science, and Social Studies). Coefficient alpha is a standard index of reliability that provides an indication of the consistency of scores for different samples of items or tasks.

In 1991-92 the alpha coefficients for open-response items range from a low of .59 for Grade 4 reading to a high of .80 for Grade 12 social studies. The corresponding range in 1992-93 was from a low of .61 for Grade 4 science to a high of .87 for Grade 12 reading and Grade 12 social studies. Seven of the 12 subject-by-grade alpha coefficients for open-response items were less than .70 in 1991-92, compared to only 2 of 12 coefficients below that value in 1992-93. In 1993-94, only 1 of the 12 grade by subject alpha coefficients (grade 8 science was the exception) was below 0.70. The generally higher coefficients in 1992-93 and 1993-94 can be attributed to the fact that the number of common open-response items increased from 3 items per subject-by-grade combination in 1991-92 to 5 in 1992-93. For the 20 multiple-choice items the 1992-93 alpha coefficients ranged from a low of .62 for Grade 8 science to a high of .88 for Grade 12 reading. Only 1 of the 12 multiple-choice subject-by-grade coefficients was less than .70 and 7 of the 12 were .80 or higher. In most cases, the coefficient alpha reliabilities are lower than those

normally obtained for standardized tests used to report individual student scores in major domains such as reading, mathematics, science, or social studies.

It is concluded in Chapter 13 of the March 1, 1995 draft of the KIRIS Biennium I Technical Manual that "the current reliabilities are not sufficiently high to make student-level decisions without additional information" (1995, p. 223). As the chapter goes on to suggest, a combination of open-response items and multiple-choice items would be likely to yield improvements in the student-level alpha coefficients. Alternatively, the student-level reliabilities could be increased by adding additional open-response items to the common core taken by all students, albeit at the cost of increased testing time.

Reliability estimates are not reported for the combination of open-response and multiple-choice items, presumably because only the open-response items are used for reporting KIRIS results. The combination is used for purposes of Chapter I reporting, however, and an evaluation of the combined items (multiple-choice and open-ended) would be relevant for that use and for considering the gains in reliability that would be achieved by a change in policy to include both multiple-choice and open-response items in obtaining student-level scores. It might be noted in this regard that the State Board for Elementary and Secondary Education approved a plan for the transitional test at its June 1994 meeting to "include performance on common multiple-choice items in both student-level and school level reporting" (KIRIS 1994-95 District Assessment Coordinator

Implementation Guide, Kentucky Department of Education, August, 1994, p. 2). Unfortunately, this plan was changed in September, 1994, and multiple-choice items were dropped from the assessment.

Although the 1993-94 alpha coefficients for student-level data are not included in the March 1, 1995 draft of the KIRIS Biennium I Technical Manual, they are included in the Celebrate the Progress: 1992-94 Kentucky Accountability Results Summary News Packet (February, 1995) for mathematics and science tests (p. 17) and for student-level battery and academic index scores (p. 16). The 1993-94 battery and academic index results are also reported in Kingston's April 27 "Accuracy" paper. For math and science, the coefficients range from a low of .74 to a high of .81 across the two subject and three grade combinations. Those figures are the average across 12 forms, each consisting of 7 items (5 common and 2 matrix sampled), administered in 1993-94. The addition of the two matrix-sampled items increases the alpha coefficients compared to the corresponding 1992-93 figures based on only the five common items. Only the five common items were used for reporting, however. Hence, these reliabilities give an inflated indication of the reliabilities of student scores that were actually reported in 1993-94.

The common plus matrix sampled item reliabilities may be considered appropriate as a way of providing an indication of what to expect in 1994-95 when student scores will be based on 7 items at Grade 4 or 8 at Grades 8 and 12 (Kingston, 1995, p. 2). Using matrix-sampled items to report individual student results, however, raises additional questions about comparability of

results for students who are administered different matrix-sampled items. The average within-form reliabilities ignore errors due to lack of perfect between-form comparability of the matrix-sampled items.

Standard errors of measurement for individual student scores would be more informative than the alpha coefficients. The standard errors could be used to place error bands around the point estimates that are presented in the form of arrows in the individual student reports. This would provide a reader with a visual indication that a student who has his or her scores represented by an arrow pointing to the upper end of, say, the apprentice part of the score bar, may have a reasonably high probability of actually belonging in the proficient range, for example.

School-level assessment results are reported by subject and grade. If those results are to be used to make inferences about improvements in school performance in a subject area from one year to the next, estimates of measurement errors associated with those scores should take into account cohort, task, and rater differences. The generalizability analyses reported in the March 1, 1995 draft of the KIRIS Biennium I Technical Manual are not specific to subject area and, with the exception, of the coefficients reported in Table 13-5 (p. 221), do not reflect difference due to task or student sampling. More complete reporting of generalizability analyses, including estimates of standard errors associated with school-level scores for schools of various sizes are needed.

References

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

The Evaluation Center, Western Michigan University. (1995). An independent evaluation of the Kentucky Instructional Results Information System KIRIS. Study conducted for the Kentucky Institute for Education Research. Frankfort, KY: The Kentucky Institute for Education Research.

Kentucky Department of Education. (1994). KIRIS 1992-93 Technical Report. Frankfort, KY: Author.

Kentucky Department of Education. (1994, August). KIRIS 1994-95 District Assessment Coordinator Implementation Guide. Frankfort, KY: Kentucky Department of Education.

Kentucky Department of Education. (1995a, February). Celebrate the progress! 1992-94 Kentucky Accountability results summary news packet. Frankfort, KY: Author.

Kentucky Department of Education. (1995b, March 1). KIRIS Biennium I Technical Manual. Frankfort, KY: Author.

Kingston, N. (1995, April 27). Accuracy of Kentucky school reward decisions (Unpublished Manuscript). Frankfort, KY: Kentucky Department of Education.

Kingston, N. & Dings, J. (1995, April 19). Estimating the accuracy of complex school accountability decisions. Paper presented at the meeting of the American Educational Research Association, San Francisco.

CHAPTER 4

Kentucky's Portfolio Assessments: Are the Scores Reliable and Valid?

4.1 Central Questions and Principal Findings

Portfolio assessment plays a key and growing role in KERA. Portfolio scores were the only measure of writing performance used for accountability in the first biennium; on-demand writing assessments were administered, but a decision was made during the past school year not to use them for accountability. Thus writing portfolios had an appreciable impact on schools' accountability index scores, constituting 1/6 of the total index and 20 percent of the cognitive component of the index. Moreover, the role of portfolios in the accountability system is expected to grow. Mathematics portfolios were implemented more recently and were not used for accountability purposes in the first biennium, but current plans call for them to be used for accountability in the second biennium.

The design of KIRIS reflects two objectives: encouraging qualitative changes in instruction (such as writing throughout the curriculum and more emphasis on higher-order thinking), and holding schools to a strict accountability standard. These two goals can conflict, in that an assessment well designed for one of these purposes may be poorly suited to the other. For numerous reasons, the tensions between these goals are particularly substantial in the case of portfolio assessment, which is embedded in instruction and unstandardized in terms of tasks and administrative conditions (see, e.g., Koretz, et al., 1994). The tensions are exacerbated in KIRIS because the

accountability system (unlike, for example, Vermont's portfolio assessment program) requires precise classifications of schools based on student performance for the assignment of rewards and sanctions.

Accordingly, it is necessary to evaluate the KIRIS portfolios in terms of two different sets of criteria, both of which are discussed below:

- measurement quality, which is necessary if the accountability system is to reward and sanction the right schools; and
- effects on curriculum and instruction.

In addition, it is important to note that if the KIRIS portfolio assessment were used only for "low-stakes" purposes -- that is, if neither students nor educators faced serious consequences for portfolio scores -- less rigorous standards of measurement quality would be appropriate. In a low-stakes context, for example, one might be willing to accept greater error in scoring than one can accept if scores are the basis for important rewards and sanctions.

This chapter focuses primarily on the writing portfolio system. Only the writing portfolios were used for accountability purposes in the first biennium, which is the primary focus of this report. In addition, the Panel could not obtain mathematics portfolio data with which to carry out analyses comparable to the analyses of writing portfolio scores described below. However, the KIRIS mathematics portfolios raise many of the same issues as the writing portfolios and are slated to assume an accountability

function in the second biennium, so limited discussion of the mathematics portfolio program is included in this chapter.

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. Although improvements in the accuracy of portfolio scoring were noted in the most recent school year (1993-94; Lewis, 1995), scoring remained poor enough to undermine the validity and utility of the scores for purposes of accountability. In 1994, raters were moderately consistent in ranking students' work but inconsistent in their application of scoring standards. That is, they disagreed about the percentage of portfolios reaching each of the KIRIS standards. Specifically, the ratings of teachers from students' own schools, which are the scores used in the accountability system, appeared to be biased substantially upward. Moreover, there is currently no effective check built into the system to identify and correct biased scores.

Evidence about the reliability of the KIRIS portfolio scores (a larger question than the consistency of scoring, and one which takes other sources of error into account) is lacking. Estimating score reliability is a difficult challenge in unstandardized assessments such as portfolio systems, and that difficulty is exacerbated by certain aspects of the KIRIS design. Evidence about the validity of the scores is limited but discouraging.

The Panel believes that the evidence at hand does not justify the use of the KIRIS portfolios for high-stakes purposes. Specifically, we do not believe that it is appropriate at this time to use the KIRIS portfolio scores as a basis for rewarding or sanctioning schools. (For the same reasons, the Panel would disapprove of using the portfolio scores as a basis for high stakes for individual students.) Consistent with the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985), the Panel believes that it is the responsibility of the test user to ensure that the quality of measurement has been established, particularly when the test has high stakes. Moreover, there is evidence from other portfolio programs that raise serious questions about validity (e.g., Koretz, et al., 1994). Therefore, the Panel's view is that the KIRIS portfolio assessment should not be presumed to be reliable and valid for its intended purpose and should not be used for that purpose until its reliability and validity is empirically established.

On the other hand, the Panel remains more optimistic about the potential of the portfolio system for improving instruction and believes that this potential might justify the continuation of the portfolio program on a lower-stakes basis. The evidence to date pertaining to the instructional effects of the KIRIS portfolio program is limited, unsystematic, and mixed. But experience elsewhere (e.g., Vermont; see Koretz, et al., 1994; Stecher & Mitchell, 1995) suggests that (1) portfolio assessment

may have substantial effects on instruction, many (but not all) positive; and (2) those effects may take a number of years to become evident, because it takes teachers a long time to learn to adapt to the system.

Therefore, the Panel recommends that if the portfolio system is maintained, systematic evidence about its diverse effects be collected over a period of several more years and be made public. Moreover, even if the portfolio program is continued on a lower-stakes basis, it will be the responsibility of the Kentucky Department of Education to insure that evidence about reliability and validity be obtained.

4.2 The Operation of the KIRIS Portfolio Program: Key Points

A detailed discussion of the operation of the KIRIS portfolio system is beyond the scope of this report. However, it is important to note a few aspects of the system that potentially bear on the validity of the portfolio scores for use in the KIRIS accountability system and on the system's potential effects on instruction. Except as otherwise noted, comments here are based on the KIRIS technical manuals from various years, the Kentucky Mathematics Portfolio Teacher's Guide (referred to here for brevity as the mathematics guidelines), and the Kentucky Writing Portfolio Teacher's Handbook (referred to here as the writing guidelines).

In contrast to the on-demand portions of KIRIS, the portfolio program is designed to be integrated tightly into ongoing instruction. For example, the writing guidelines note

that "portfolios are not an 'add-on to the instructional program; instead they should be the foundation upon which the writing program is built" (p. ii). Similar language is found in the mathematics guidelines. Many of the specific attributes of the program that are discussed here (such as its lack of standardization) appear to stem directly from this intended integration with instruction. The demands of good instruction do not always correspond to the demands of good assessment, however, and in the Panels' view, the tensions between the instructional and assessment goals of the portfolio program are severe.

The portfolio program is highly unstandardized in almost all aspects of its operation. There are very general guidelines about the types of tasks that should be included; for example, the writing portfolio in grade 8 should include one piece that achieves any one or combination of seven different purposes, such as defending a position, predicting an outcome, or creating a model. The writing guidelines prohibit the inclusion of group products, while the mathematics guidelines allow one (although both in fact permit and even encourage substantial assistance from other people in performing and revising any portfolio tasks). But beyond those general guidelines, teachers and students are largely unrestricted in their selection of tasks and have substantial leeway in terms of the types of products they entail. The mathematics guidelines, for example, note that videotapes, audiotapes, and computer disks are all acceptable products. Tasks may differ among schools or over time in terms

of difficulty, novelty, time allowed for completion, amount of revision permitted, and so on.

Policies pertaining to revision of portfolio pieces are highly permissive. For example, the writing guidelines note that peers may offer suggestions for revisions and call for teachers to instruct students on how to review each others' writing. They also note that teachers may provide revision check lists. Similarly, the mathematics guidelines indicate that revisions "should be a collaborative effort among students and/or between teacher and student" (p.3). The mathematics guidelines also mention parental input but do not specify what that input should be. They do clarify that teachers "will not solve the problem,...do the mathematics or writing, or make direct corrections" (p.14).

Scoring of KIRIS portfolios is carried out by teachers in students' own schools. The reasons for that decision are specified in the writing guidelines (p. 9): (1) to enhance teachers' understanding of and ability to apply the Learner Outcomes and performance standards, and (2) to ensure consistency in the application of those standards. Training of teachers for both implementation and scoring entails a wide array of mechanisms, but the core mechanism has been a pyramidal model in which Area Coordinators train cluster leaders, who in turn train groups of 20 to 25 teachers. KDE provides extensive training materials, including diverse sets of benchmark papers. The scoring rubrics (called "holistic scoring guides") reference a variety of attributes in terms of which portfolios should be

judged, such as "purpose/approach" and "surface features" in writing and "problem solving" and "understanding/connecting core concepts" in mathematics. While raters are free to annotate their ratings with respect to these attributes, they assign only a single score for each portfolio in the Novice, Apprentice, Proficient, Distinguished scale. The individual pieces comprising the portfolios are not scored separately.

The Panel has not analyzed the KIRIS portfolio scoring rubrics in detail, but they appear to have both strengths and weaknesses. On the positive side, the rubrics focus attention on attributes of performance that are consistent with the current wave of curricular reform, such as problem solving and communication in mathematics. On the negative side, the rubrics appear complex and in some respects ambiguous. In a number of instances, very similar wording appears at different scale points or in describing different attributes. In writing, for example, the Proficient writer is "focused on a purpose," while the Distinguished writer "establishes and maintains a clear focus." In mathematics, 'drawing charts' is used as evidence of three different attributes: problem solving, reasoning, and mathematical communication. Of course, training and benchmark papers may provide clarity beyond that offered by the rubrics themselves. Some degree of generality and ambiguity in rubrics may also be hard to avoid in a system that generates highly variable products in response to largely unconstrained selections of tasks. However, these potential difficulties with the rubrics

may be germane to the errors in scoring noted later in this chapter.

4.3 The Reliability of KIRIS Writing Portfolio Scores

One critical aspect of measurement quality is the consistency or reliability of scores. As noted in Chapter 3, unreliability (that is, uncertainty) in the assignment of scores can arise because of a variety of factors, including both situational and personal factors.

In standardized performance assessment programs such as the KIRIS transitional assessment, in which tasks and administrative conditions are uniform, analysis of reliability generally focuses on two potentially important sources of error: scoring (error attributable to raters) and tasks (that is, uncertainty stemming from the sample of tasks a student is assigned). In the case of portfolios, however, numerous other sources of error come into play. For example, the following factors may vary across students: initial instructions from teachers; the amount and type of pre-teaching using similar tasks; time allotted to performance of the tasks; amount and type of assistance from teachers, peers, parents, and others in performing the tasks; opportunities for revision; and the amount and type of assistance in revision provided by others. In addition, in portfolio assessment programs, nominally similar tasks may in fact be different. For example, in the Vermont portfolio assessment program, raters found variants of common problems that differed substantially in difficulty. All of these factors could have a

sizable effect on the score obtained by a given student, independent of the student's proficiency, and therefore bear on the reliability and validity of portfolio scores.

4.3.1 Consistency of Ratings

In the case of KIRIS, there is clear evidence about the impact of one source of error: inconsistencies among scorers. Consistency of scoring between raters is often labeled "inter-rater reliability." Inter-rater reliability is only one narrow aspect of the reliability or consistency of the final scores, which can be affected by many other factors, including those listed above. For that reason, inter-rater reliability (the consistency of scoring) will generally be considerably higher than the reliability of scores. Put differently, the consistency of scoring places an upper bound on the reliability of scores, and therefore substantial inconsistency in scoring may be sufficient to make the portfolios inappropriate for high-stakes use.

In the first two years of the portfolio program, the Kentucky Department of Education reported substantial inconsistencies among raters (see KIRIS Biennium I Technical Manual, Chapter 12 (Kentucky Department of Education, March 1, 1995 draft). For example, the Kentucky Department of Education reported that in 1992-93, the average score from fourth grade teachers statewide was about 33, while the average score from summer rescores was about 20. In grade eight, the difference was roughly as large: the corresponding averages were 30 and 18.

Scores by the Kentucky Writing Advisory Committee were lower yet (KIRIS Biennium I Technical Manual, 1995, p. 199).

These discrepancies clearly threaten the validity of the accountability index, and the Panel therefore analyzed the 1994 data to determine the level of reliability of scoring at the end of the Biennium. We addressed two questions: (1) are raters consistent in their rankings of students' writing portfolios, and (2) do raters apply standards consistently in their scoring of portfolios? We compared final ratings by teachers from students' own schools (that is, the scores used in the accountability index) to scores provided in the summer review (in which students' portfolios were scored by individuals other than their own teachers). We had both portfolio scores for over 6,900 fourth-grade students and almost 4,900 eighth-grade students. We are grateful to the Kentucky Department of Education for making these data available to the Panel.

Inconsistencies in rankings and inconsistencies in applying standards can represent different types of errors in scoring. For example, simple random error in scoring will cause inconsistencies in the rankings assigned by two groups of scorers but is not likely to have much impact on the average scores assigned by each group. However, if the two groups are systematically different in their application of standards, they are likely to assign different average scores or place different percentages of students at one or more of the standards.

In the 1993-94 school year, Kentucky raters showed moderate consistency in their rankings of writing portfolios. That is,

they showed a moderate tendency to rate the same portfolios as good or bad relative to each other, but not necessarily relative to the KIRIS standards. The Spearman correlations between the ratings provided by teachers in students' own schools and by the summer scorers were .70 in fourth grade and .67 in eighth grade. (A value of 0.00 would indicate no systematic relationship between the rankings of portfolios by original raters and summer raters; a value of 1.00 would indicate perfect correspondence.) These correlations are reasonably high in comparison to various performance assessments, but some portfolio programs have reached higher levels of inter-rater reliability. For example, in the first National Assessment of Educational Progress trial of writing portfolios, inter-rater correlations ranged from .76 to .89 (Gentile, 1992). More important, the KIRIS correlations still entail considerable error: knowing the rankings provided by teachers in students' own schools allows one to predict only about half of the variance in the scores from the summer rescoring.

Despite this moderate consistency in rankings, however, Kentucky raters were not consistent in applying standards: the scores assigned by teachers in students' own schools (that is, the scores used in the accountability system) were again considerably higher than those assigned by raters in the summer rescoring. This is shown by Tables 4.1 and 4.2, which are based only on students whose portfolios were rescored in the 1994 summer review.

In grade 4 (see Table 4.1), the modal score assigned by students' own teachers was Apprentice, with approximately equal percentages rated as Novice and Proficient, and four percent rated as distinguished. Applying the same values for the NAPD scale that are used in the accountability index, this produces an average score of 48. When portfolios were rescored by others in the summer review, more students were scored as Novice or Apprentice and fewer as Proficient and Distinguished, and the average score was only 36. This 12-point disparity is roughly comparable to the discrepancy for 1993 documented there. However, the 1994 disparities in ratings were much larger in grade 8 than in grade 4 or in grade 8 the previous year (see Table 4.2). In grade 8, the rescoring placed fully half of the students at the Novice level, and the average index dropped 22 points, from 51 to 29. To put this drop in perspective, a school with a baseline index of 30 would only have to increase its index by 8 points -- roughly a third as much as the discrepancy between original and summer scores -- to be placed in the reward category.

To further clarify the extent of consistency between raters, Table 4.3 shows how original scores for eighth grade students were rescored in the summer. The rows in Table 4.3 represent the official scores assigned by teachers in students' own schools,

Table 4.1. Percent of Student Portfolios at Each Score Point and Index for Group, Fourth Grade, N=6,895[1]

	Original Scores	Summer Rescores
Novice	28	37
Apprentice	42	47
Proficient	26	14
Distinguished	4	2
Index (0, 40, 100, 140)	48	36

[1]Column entries may not sum to 100 percent because of rounding. Index values are obtained by assigning each level the score it is given in the accountability index (Novice = 0, Apprentice = 40, Proficient = 100, Distinguished = 140) and averaging.

Table 4.2. Percent of Student Portfolios at Each Score Point and Index for Group, Eighth Grade, N=4,825[1]

	Original Scores	Summer Rescores
Novice	28	51
Apprentice	38	35
Proficient	26	12
Distinguished	7	2
Index (0, 40, 100, 140)	51	29

[1]Column entries may not sum to 100 percent because of rounding. Index values are obtained by assigning each level the score it is given in the accountability index (Novice = 0, Apprentice = 40, Proficient = 100, Distinguished = 140) and averaging.

and the columns represent the scores assigned in the summer rescoring. The entries are the percents of portfolios assigned a given original score that were assigned each score in the summer, and the shaded cells are represent cases in which raters agreed. Thus, for example, summer raters tended to agree with original scores of Novice; 91 percent of portfolios scored as Novice by teachers in students' own schools were also scored as Novice by the summer raters. The agreement is much worse, however, at every other score point. For example, only 42 percent of the

portfolios initially rated as Apprentice were also rated as Apprentice in the summer rescoring, and more than half of the Apprentice portfolios (55 percent) were rescored as Novice. Similarly, 55 percent of portfolios scored as Proficient by teachers in students' own schools were rescored only as Apprentice by the summer scorers. The agreement was worst at the Distinguished level, but very few students were scored at that level by either set of raters, so this disparity probably had less impact on the accountability system than the errors at other score levels.

Table 4.3. Percent of Portfolios at Original (Official) Scores Rescored at Each Score In the Summer Rescoring Session, Grade 8

Original Scores	Summer Rescores[1]				
	Novice	Apprentice	Proficient	Distinguished	Total
Novice	91	8	1	0	100
Apprentice	55	42	3	0	100
Proficient	14	55	29	2	100
Distinguished	3	32	46	18	100

[1]Components may not sum to 100 percent because of rounding.

The implications of these findings seem clear: the scores provided by teachers in students' own schools are biased upward by a substantial amount. There appears to be no reason to suspect a downward bias in the scores from the summer rescoring, and indeed the Biennium I Technical Manual showed that in 1993, summer rescores were considerably closer than original scores to the ratings of the Kentucky Writing Advisory Committee. Moreover, the correlation between scorers, while moderate, is nonetheless small enough that the degree of bias could vary

markedly from one school to another, thus affecting the allocation of rewards and sanctions.

4.3.2 Student-Level Score Reliability

Although consistency of scoring is important in portfolio assessment, the reliability of scores -- which takes into account all important sources of error, not just those arising from scoring -- is even more important. Score reliability indicates how much confidence students, teachers, and parents should have in estimates of students' performance and is therefore important for the assessment functions of KIRIS. (In this context, it is relevant to recall that when publishers of commercial norm-referenced multiple-choice achievement tests publish estimates of the reliability of their tests, they are estimating score reliability. In those cases, the "inter-rater" reliability is nearly perfect -- barring malfunction, the machine would produce the identical score time after time -- but scores are less than perfectly reliable regardless, because of such factors as the limited sampling of tasks in the test.)

The Panel believes it likely that the reliability of KIRIS portfolio scores is considerably lower than their inter-rater reliability, and it strongly disagrees with the Kentucky Department of Education's earlier argument that portfolio tasks should be considered "fixed" for this purpose. (Treating tasks as fixed would denote that variations among tasks need not be considered in estimating score reliability.) However, it is not feasible for the Panel to estimate the reliability of the

portfolio scores, and extant research does not provide a clear suggestion about the level of score reliability one should expect from the KIRIS portfolio system.

The difficulty in estimating the score reliability of the KIRIS portfolios stems from several factors, including both the nature of the program (in particular, the wide array of sources of error that can come into play because of its almost entirely unstandardized nature), the lack of systematic information about variations in implementation, and the way in which portfolios are scored. Portfolio tasks, unlike tasks on a well-constructed standardized performance assessment, may not be independent of each other; for example, an individual teacher may tend to give tasks of a certain type, or one task may depend on another already completed. Error in inferring students' competence from portfolio tasks may arise from differences in revision, help from peers, help from parents, etc. Systematic descriptive information on variations in implementation of the portfolio program might suggest which of these (and other) sources of dependency and error are most substantial.

Unlike some portfolio systems (such as the Vermont program), KIRIS does not provide separate scores for the tasks that are included in portfolios, so one cannot examine variability of performance among them. Because of reasons such as those above, one could not in any event use variations among the tasks in students' portfolios as an indicator of task-related error in scores. One might argue, however, that variation among tasks in a students' portfolio underestimates task-related error because

of dependencies among tasks that would not arise in other forms of assessment. For example, many teachers are likely to have their own favorite tasks, types of tasks, and styles of administration, and these are likely to differ substantially among schools and over time as well. Thus, variation among tasks in a student's portfolio is likely to be considerably smaller than variation among a random set of tasks drawn from the domain (or from all the portfolios in the state) would be. Following that logic, Koretz, et al. (1993) used variability among portfolio tasks to estimate score reliability in the Vermont portfolio program, but with the assumption that the resulting estimates would overestimate the reliability of scores. This approach might help disentangle the sources of error in portfolio scoring but would not produce estimates of score reliability accurae enough to justify high-stake uses of the portfolio scores.

If portfolio scores are to continue to be used for comparisons of performance improvements among schools (even for low stakes), the Panel recommends that systematic data on implementation of the program be obtained. Given the uncertainties about score reliability noted here, the Panel does not recommend that the Kentucky Department of Education now base estimates of the reliability of portfolio scores on differences in performance across tasks in the portfolios. However, in conjunction with descriptive information about implementation of the program, analysis of variability in performance across tasks could be a useful step toward a better understanding of the

strengths and weaknesses of the program. In addition, the Panel recommends that the Kentucky Department of Education explore alternative methods for estimating the score reliability of portfolios.

4.3.3 School-level Reliability of KIRIS Portfolio Scores

KIRIS is designed to provide school-level rather than student-level scores. Therefore, it is important to go beyond the reliability of student-level scores and to examine the reliability of KIRIS scores at the school level.

As noted above, a number of factors made it impractical for the Panel to estimate the score reliability of portfolios at the student level, and the same issues precluded an estimate of score reliability and the school level. However, it was practical to obtain estimates of inter-rater reliability at the school level. For this purpose, the Panel used the sample of students whose portfolios were rescored in the 1994 summer rescoring. The Panel analyzed only grades 4 and 8 because we lacked summer rescoring data for grade 12. For each school, two means were calculated: the mean of the summer rescores, and the mean of the original scores (used in the accountability index) for those students. Because some schools had few students rescored (as few as 1 per school), we included in this analysis only schools in which 15 or more students had both original (official) and summer rescores. (This criterion led us to exclude about 10 percent of the schools

at each grade level, leaving 348 schools at grade 4 and 175 schools at grade 8.[1]

In the fourth grade, the school-level correlation between the mean original score and the mean rescore was 0.69 -- virtually identical to the correlation between original scores and rescores at the student level. In grade 8, the school-level correlation was .63, trivially lower than the corresponding student-level correlation. Thus, for the rescored sample, knowing schools' average score from the summer rescoring allows one to predict about half of the variance in school means from the original scoring that is used in the accountability index.

These moderate correlations at the school level, however, imply only that schools were ranked with a modicum of consistency by the original and summer scores; they do not imply that they received similar scores on average. In fact, just as was shown above for individual students, the average school received a much lower score in the summer rescoring than it received based on the original scores for the same portfolios. In grade 8, the average school received a mean score in the summer rescoring 22 points lower than it received on the same portfolios in the original scoring used for the accountability index. The corresponding discrepancy in fourth grade was 13 points.

The combination of a large mean difference between original scores and rescores and the moderate consistency in the rankings

---

[1]Although imposing this screen raised the correlation between original scores and summer rescores, setting a higher minimum did not improve the correlations consistently and led to the deletion of many schools.

of schools implies that some schools will be substantially advantaged and others substantially disadvantaged by the apparent bias in the scores used in the accountability system.

4.4 The Validity of KIRIS Writing Portfolio Scores

The ultimate question about the quality of the information provided by the KIRIS portfolios is their validity for their current use: Do they provide a reasonable and defensible basis for the conclusions that are based on them? (In this section, we do not consider the effects of the testing program, which some people label "consequential validity." We consider that essential question separately. Here we are concerned only with adequacy of measurement.)

Reliability can be seen as a necessary but insufficient criterion for validity. If scores are not reliable, they cannot be valid (by most meanings of the word). However, finding that scores are reliable is no guarantee that scores are valid for their intended purposes.

Evidence about the validity of the KIRIS portfolio scores is limited, and the structure of the program makes it difficult to obtain the information one would want. One limitation, for example, is the lack of data on students' performance on the individual pieces that comprise their portfolios and the lack of consistent information on performance on each of the various dimensions (such as "problem solving" and "reasoning" in mathematics, or "purpose/approach" and "idea development/support" in writing) that raters are supposed to consider in reaching a

score for a portfolio. There are no data on the specific tasks that are included in portfolios or their difficulty. Perhaps most important, there are only limited external performance criteria to which portfolio scores can be compared.

Despite these limitations, some evidence pertaining to the validity of KIRIS portfolio scores is available, and taken together, it suggests skepticism about the validity of the scores.

4.4.1 The Validity Implications of Lack of Standardization

Apart from any evidence based on the portfolio scores themselves, the nature of the portfolio program raises serious doubts about the validity of portfolio scores for comparative measurement of student performance (such as those required for the KIRIS accountability system). One critical attribute of the program in this respect is its extreme lack of standardization, which apparently stems from (and may be appropriate for) the instructional-improvement goals of the KIRIS portfolio program. Beyond some minimal guidelines, the system imposes only minimal constraints on the implementation of the program. For example, it imposes few restrictions on the selection of tasks, the amount of revision, the extent of pre-teaching, etc. Faced with these very permissive guidelines, as well as an explicit expectation that portfolio work will be tightly integrated into instruction, one would expect teachers to vary substantially in their portfolio practices. In the Vermont portfolio program, surveys

confirmed wide variations in teachers' practices pertaining to selection of tasks, revisions, etc. (see Koretz, *et al.*, 1994).

However desirable this lack of standardization may be for purposes of instructional improvement, it can seriously threaten the validity of comparisons among schools or of estimates of change over time based on portfolio scores. The lack of standardization of tasks poses an obvious problem: students with identical proficiency may obtain different scores depending on the difficulty level and novelty of the tasks they perform for portfolio purposes.

4.4.2 Validity Implications of Contributions from Others to Portfolios (and Performance Events)

As noted earlier, the portfolio guidelines are very permissive with respect to the source, amount, and type of input other individuals -- teachers, fellow students, and others -- can contribute to an individual student's portfolio. Although only one formally identified group product is allowed in the mathematics portfolio and none is permitted in the writing portfolio, the guidelines would seem to make it highly likely that the input from other individuals is both substantial and highly variable from class to class and even from student to student within a class.

This raises an obvious question about validity: whose work is represented in a portfolio? And to what degree does this vary among schools, either at any one time or over time (as a way of raising scores)?

The mathematics guidelines state "students must have total ownership of their mathematics" and that interventions from others should "enhance rather than...diminish that ownership." They also note that "at no time should students' ideas, revisions, or editing be characterized [emphasis added] as teacher-authored, peer-authored, or parent authored" (p. 11). How contributions from others is "characterized" is beside the point, however, because the KIRIS system in fact characterizes portfolio scores as an indication of the students' own performance. Rather, the issue is variations in the contributions made by others and the possibility that such variations undermine the validity of the KIRIS system's representation of portfolio scores as having their intended and claimed meaning. For example, suppose two teachers have students who are roughly equally competent as writers. Suppose also that these two groups of students received similar portfolio scores in the baseline year. (Given the lack of standardization, there is no reason to assume that similar students would in fact obtain similar scores, but the point being made here holds regardless.) Say that one teacher decides the next year to provide additional time for revisions and to provide more extensive help (within the KIRIS guidelines) for making those revisions, while the second teacher decides that her students were spending enough time on revisions and received sufficient help. All other things being equal, scores in the first class are likely to increase more than those in the second group, even if both groups stay constant in their writing proficiency.

It is important to note that contributions from others may undermine the validity of scores on Performance Events as well, although the issues in the case of performance events are narrower because the activity is more standardized. Performance events are "hybrid" tasks in which students perform an activity together and then complete the task individually. One potential threat to the validity of hybrid measures is that the performance of any one student may be influenced by the characteristics of other students in ways that could undermine relative measures of improvement. For example, suppose that higher-ability students have a disproportionate influence on group products. (Some research indicates this may sometimes be the case; see Webb, 1993.) Suppose further and that high-ability students can also be more easily taught to do well on the complex tasks involved in the Performance Events. In that case, schools with a substantial number of high-ability students will be unfairly advantaged in their efforts to raise all students' scores on Performance Events.

4.4.3 Validity Evidence From the Scoring of Portfolios

Quite apart from the general consideration that reliability places a bound on validity, the scoring of portfolios by teachers in students' own schools raises serious concerns about validity. The Panel accepts the Kentucky Department of Education's rationale that having portfolios scored in every school may be important for purposes of professional development, but doing so

raises at least two potential problems for the uses of scores called for by KIRIS.

The first problem is the possibility that some teachers will give their students inappropriately high scores. As the Kentucky Department of Education's analyses and the Panel's analyses above show, this problem has been severe in the case of the KIRIS portfolio program. Moreover, the upward bias in the scores assigned by teachers in students' own schools appears to be inconsistent across schools. This undermines the validity of portfolio scores as an external measure of improvement.

The second potential problem raised by having all teachers score their own students' work is that it requires training a very large number of teachers and makes it difficult or impossible to bar from scoring any individuals who cannot reach a reasonable level of accuracy. The pyramidal training scheme that is a response to the large number of teachers who need training raises the possibility of inconsistencies as training moves down the pyramid. Indeed, a report entitled "Regional Writing Portfolio Audit Meetings Final Report" (Kentucky Department of Education, March 28, 1994) indicated that this has been a problem in the KIRIS writing portfolio problem. This raises the likelihood of inconsistent and error-prone scoring. (In the Vermont portfolio system, the response to this tension was to specify that all teachers should score their own portfolios but that the scores used by the Vermont Department of Education would be based on controlled scoring by other teachers of random samples of portfolios drawn from each school.)

4.4.4 Convergent and Divergent Evidence

A conventional basis for evaluating the validity of a measure is to determine whether it correlates more highly with theoretically related variables than with theoretically less related variables. For example, one would typically expect one test of algebra to correlate more highly with another test of algebra than with a test of language mechanics. If the correlations did not show this pattern, one would conclude that at least one of the three tests is of low validity.

The Kentucky Department of Education provided correlations appropriate for such an analysis in Chapter 14 of the KIRIS Biennium I Technical Manual (1995, March 1 Version). In the case of KIRIS, both subject areas and test formats are relevant to comparisons such as these. For example, one might expect a multiple-choice test of mathematics to correlate more highly with another multiple-choice mathematics test than with a mathematics test (such as the KIRIS transitional test) that relies on open-response items. The Kentucky Department of Education therefore provided correlations within and across both formats (components of the KIRIS assessment) and subject areas. (The Technical Manual labels this a "multi-trait, multi-method matrix," after Campbell and Fiske [1959]; in that terminology, subject areas are "traits" and test formats i.e., KIRIS components] are "methods.")

For these purposes, the most relevant correlations are between the writing and mathematics portfolios and the other KIRIS tests of writing, reading, and mathematics. One would expect, for example, that within a single test format, writing

portfolio scores would correlate highest with an on-demand test of writing and lowest on an on-demand test of mathematics, with reading probably falling in between. Similarly, one would expect an on-demand test of mathematics to correlate more strongly with mathematics portfolios than with writing portfolios.

The evidence pertaining to the validity of portfolio scores in the Kentucky Department of Education's analysis of these correlations is ambiguous. Many of the correlations ranked in the expected order, but the differences among them were often negligible, and some ranked in the wrong order.

For example, consider the correlations for grade 8 (see Table 4.4). (These correlations reflect all students; a different set of correlations that excludes students who did very poorly on the test is described below.) The diagonal (shaded) entries, running from top left to bottom right, comparing one component to itself, are the reliability estimates used by the Kentucky Department of Education. The entries below the diagonal are raw correlations between observed scores. Thus, for example, the first two entries of the bottom row show that mathematics portfolio scores correlated .40 with multiple choice (MC) reading scores and .44 with MC mathematics sores. ("OR" stands for "open-response," and "OD" stands for "on-demand.")

The correlations above the diagonal in Table 4.4 are disattenuated (adjusted upward) to take into account unreliability of scores. Thus, the top right cell shows that when disattenuated -- that is, when the estimated effects of unreliability are removed -- math portfolio scores correlated .73

with MC reading scores; this compares to the raw correlation of .40 in the bottom left cell. Disattenuated correlations are helpful in comparisons such as these, because they reduce the possibility that correlations will be ranked in the wrong order just because of differences in the reliability of the tests. Unfortunately, the Kentucky Department of Education lacked estimates of the score reliability of portfolio scores and therefore disattenuated portfolio and on-demand writing tests using a different basis (inter-rater correlations) than were used with the other components (an estimate of intra-test consistency called coefficient alpha). This difference clouds the interpretation of the disattenuated correlations. Moreover, in a system such as KIRIS, in which decisions are made on the basis of raw scores, unadjusted relationships among measures are also important.

Reading across the second-to-bottom row, one sees that in terms of raw correlations, the writing portfolios correlate more highly with reading than mathematics (as one would expect), both for MC and OR items. The same is true of the disattenuated correlations (second-to-rightmost column). However, the differences are extremely small, in some cases negligible. Moreover, the correlations between writing portfolio scores and on-demand writing scores, which ideally should be relatively high, are lower than all but one of the raw and disattenuated

Table 4.4. Correlations Between Components of the KIRIS Assessment, Grade 8 (Reliabilities on the Diagonal, Raw Correlations Below the Diagonal, and Disattenuated Correlations[1] Above the Diagonal)

	MC Reading	MC Math	OR Reading	OR Math	OD Writing	Portfolio Writing	Portfolio Math
MC Reading	.82	.70	.80	.69	.50	.81	.73
MC Math	.56	.78	.60	.85	.45	.76	.82
OR Reading	.66	.48	.82	.73	.55	.84	.76
OR Math	.53	.64	.56	.72	.50	.83	.87
OD Writing	.37	.33	.41	.35	.68	.76	.70
Portfolio Writing	.49	.45	.51	.47	.42	.45	1.25
Portfolio Math	.40	.44	.42	.45	.35	.51	.37

[1]Extracted from Table 14-2 of the KIRIS Biennium I Technical Manual, March 1, 1995 draft.

correlations between portfolios and reading and mathematics tests. Moreover, when disattenuated, writing portfolio scores correlate more highly with MC social studies scores (not shown) than with MC reading, and almost as highly with MC science (not shown) as with MC reading. Similarly, in grade 4 (not shown), writing portfolio scores correlate as highly (more precisely, trivially higher) with OR mathematics than with OR reading.

KDE also presented a second set of correlations among KIRIS measures, in this case removing from the calculations students with such low scores that the Kentucky Department of Education presumed them not to be showing a good-faith effort. The criterion for establishing lack of good-faith effort was a score of 3 or less on either a 20-item multiple-choice test or an open-

response test with a maximum score of 20 (Kentucky Department of Education, 1995, p. 232). At least in the case of the multiple-choice components, this screen is very low, below what one would expect from chance if students simply picked answers at random. Nonetheless, this screen eliminated 37 percent of the eighth-grade students represented in Table 4.4. The resulting correlations (only raw correlations were presented) were lower than those in Table 4.4, but the basic pattern remained the same. The correlations were mostly ordered in the desired direction, but the differences among them were very small.

One additional correlation warrants note: the disattenuated correlation of 1.25 between writing portfolio and mathematics portfolio scores in Table 4.4. Correlations cannot exceed a value of 1.00 -- the value indicating a perfect correspondence between two sets of scores. (The comparable disattenuated correlation in grade 4 was 1.20; in grade 12, .88.) Thus, the values that exceed 1.00 indicate (as the Kentucky Department of Education notes in the KIRIS Biennium I Technical Manual, March 1, 1995 draft (pp. 231-232) that something is wrong with the disattenuation. The Kentucky Department of Education notes several possibilities, including an overly conservative estimate of portfolio reliabilities and confounding with factors such as teacher influence and writing ability. Of course, this finding calls into question all of the other disattenuated correlations involving portfolio scores. More important in the Panel's opinion, is that these impossible values indicate that the error

in KIRIS portfolio scores is simply not yet sufficiently well understood to have confidence in the scores.

4.4.5 Concordance Between Writing Portfolios and On-Demand Writing

In many respects, the on-demand writing scores provide the best single criterion within KIRIS against which to judge the portfolio scores because the domains and tasks are the most similar. On-demand writing did not count in the Biennium I accountability index, but the Panel's understanding is that a memorandum proposing that decision was first drafted in April of 1994 and was not presented to the Kentucky Board of Education until September, 1994. Therefore, most teachers and students should not have been aware of this pending decision when the 1994 on-demand writing test (the results of which are used here) was administered, and it is therefore unlikely that the on-demand scores were biased downward by a lack of concern about performance on that part of the assessment.

The correlations above (see Table 4.4) suggest that the writing portfolio scores show only a low concordance with KIRIS on-demand (OD) writing scores unless adjustments are made for unreliability in the scores. These correlations, however, reflect consistency in rankings, not consistency in applying KIRIS standards. Accordingly, the relationships between 1994 on-demand and portfolio writing scores are explored in more detail here. We consider unadjusted scores, which are the scores that count in the accountability index.

The comparison of OD and portfolio writing scores poses a logical dilemma, however, stemming from a seeming inconsistency in the Kentucky Department of Education's presentation of the portfolio program. Both the on-demand and portfolio writing assessments are "authentic" measures of writing. Moreover, they are both scored using the same rubric and are both scored directly onto the Novice, Apprentice, Proficient, Advanced (NAPD) scale. This scale suggests that they should support similar inferences, although some discrepancy between the scores should be expected because of the different contexts in which students are performing. That is, a score of 3 on either assessment is interpreted as "Proficient in writing," not as "Proficient in writing portfolios but not in writing on demand" or vice-versa. Put differently, if the percentage of students scoring at the third standard increased from 20 to 25 percent on either assessment, it would be presented as an increase of 5 points in the percent of students who are Proficient in writing.

On the other hand, the KIRIS portfolios are supposed to reflect students' best work. If only students' best work is considered in the portfolio system, one would expect students to score higher than they do on the on-demand writing if the same rubric were applied. But in that case, a score of 3 on the portfolio would not imply the same thing about a student's competence as would a 3 on the OD assessment, and it would therefore not be sensible to consider both scores of 3 to indicate that the student has reached "Proficient" status.

Faced with this inconsistency in presentation, the Panel's view, in conformance with the Standards for Educational and Psychological Testing, is that the specific inference based on the scores should be the basis for validation. In this case, both the on-demand and portfolio writing assessments were scored directly onto the NAPD scale and would seem to support the sasuring different constructs (different aspects of writing) and should be presented as such. It is important to note that the second interpretation would still leave the need to validate the portfolio scores and would not alter the threats to validity posed by discrepancies between original and summer portfolio scores, differences in contributions to portfolio entries by individuals other than the student being scored, differences in task selection, etc.

Data from the summer rescoring of portfolios is repeated here (albeit for a slightly different sample), because it helps clarify the discrepancies between portfolio and on-demand writing scores.

The correspondence between OD and portfolio writing was generally poor, especially in grade 4 (see Table 4.5); students generally scored higher on their portfolios than in the on-demand writing assessment. The modal student (42 percent of students) was scored Apprentice on their portfolios, and about a fourth of the students were scored Novice and Proficient. The average score (NAPD=0,40,100,140) was 48. Portfolio rescoring shifted students out of Proficient and primarily into Novice, dropping the mean score to 36, but the modal score remained Apprentice.

In contrast, in the on-demand writing assessment, 75 percent of students were scored as Novices, and the average score was 11. In other words, the minor part of the very large discrepancy between OD and portfolio writing scores in grade 4 can be attributed to scoring by teachers in students' own schools, but the larger part must be attributed to other differences between the portfolio and OD assessments.

Table 4.5. Percent of Students Scoring at Each Scale Point in Writing and Mean Score, Grade 4

	Official Portfolio	Portfolio Rescore	On-Demand
Novice	28	37	75
Apprentice	42	47	23
Proficient	26	14	2
Distinguished	4	2	0
Mean (0,40,100,140)	48	36	11

In part because of the large mean difference between portfolio and OD scores, the rate of agreement between the two sets of scores was little better than one would expect on the basis of chance. OD and portfolio scores agreed exactly for about 35 percent of students; 31 percent agreement would be expected by chance alone. Nearly all of the students who obtained a score of Novice in the portfolio assessment also received a score of Novice in the on-demand writing assessment, but that was the only scale point at which there was substantial agreement (see Table 4.6). Of the students who obtained a score of Apprentice in the portfolio assessment, 79 percent obtained a score of Novice in the on-demand writing assessment. The disparity was worse yet for students with portfolio scores of

Proficient; virtually all of them received lower scores in the on-demand assessment, and well over half (58 percent) received a score of Novice. Spearman correlations between portfolio and on-demand scores (which measure consistency of ranking and would be less affected by the large mean difference) were also moderately small. The correlation between OD scores and official portfolio scores was 0.34; between OD scores and portfolio rescores, 0.32.

Table 4.6 Percent of Students with Portfolios at Original (Official) Scores Assigned Each Score In the On-Demand Writing Assessment, Grade 4[1]

	On-Demand Writing Scores				
Portfolio Scores	Novice	Apprentice	Proficient	Distinguished	Total
Novice	92	8	0	0	100
Apprentice	79	20	1	0	100
Proficient	58	39	4	0	100
Distinguished	37	52	11	1	100

[1]Components may not sum to 100 percent because of rounding.

Agreement between OD and official portfolio scores was a bit better in grade 8 but still not strong. However, in grade 8, the summer portfolio rescores were closer to the OD writing scores than they were in grade 4, suggesting that a large share of the discrepancy between official portfolio scores and on-demand writing in grade 8 may be attributable to the scoring of portfolios. The official portfolio scores were distributed much as in grade 4: the mode (38 percent) was Apprentice, and a bit more than a fourth each were scored as Novice and Proficient (see Table 4.7). This yielded a mean score (on 0,40,100,140 NAPD scale) of 51, close to the mean in grade 4. The summer rescores placed the modal student (1 percent of them) at Novice and

produced a mean score of 29. The OD writing distribution was quite similar to the summer rescores, but with fewer Proficient, more Novice, and a mean of 20.

Table 4.7. Percent of Students Scoring at Each Scale Point in Writing and Mean Score, Grade 8[1]

	Official Portfolio	Portfolio Rescore	On-Demand
Novice	28	51	57
Apprentice	38	35	37
Proficient	26	12	5
Distinguished	7	2	0
Mean ()	51	29	20

[1]Percents may not sum to 100 because of rounding.

Given the somewhat smaller discrepancy between OD and portfolio scores in grade 8, the agreement rate between the two scores was a little better than in grade 4, although still not greatly better than chance. OD and official portfolio scores agreed for 42 percent of students; 32 percent agreement would have been expected by chance alone. Similarly, the correlations between scores was somewhat better than in grade 4, although they were still only moderate. The Spearman correlation between OD writing scores and official portfolio scores was .48. Although the rescores were closer to OD scores in terms of averages, they were not more consistently related to OD scores in terms of rankings: the correlation between them was .42.

The substantial discrepancy between official portfolio scores and rescores casts serious doubts on the validity of the official scores as they are currently presented. However, the interpretation of these findings is clouded by the inconsistency

noted above in Kentucky Department of Education's representation of the portfolio scores. If portfolio scores and OD scores are truly on the same scale -- e.g., a score of "3" denotes being Proficient in writing in both cases -- then the large discrepancies between OD and portfolio writing scores is troubling indeed. If, on the other hand, portfolio scores are expected to be higher because they represent "best pieces," then the Kentucky Department of Education should not place them on the same scale without some appropriate adjustment, and another external criterion may be needed to validate the portfolio scores.

4.5 Effects of the Portfolio Program on Instruction

Although the Panel's does not consider the current KIRIS portfolio program appropriate for inclusion in the KIRIS accountability system, it remains guardedly optimistic about the portfolio program's potential as a lever for instructional improvement.

Evidence about the impact of the program on instruction is limited, largely anecdotal, and inconsistent. In presentations to this Panel, for example, the Kentucky Department of Education provided examples of both positive effects and disappointments. Limited evidence from other programs, however, suggests two conclusions: (1) portfolio programs may have substantial potential to influence achievement, but (2) the types of fundamental instructional changes envisioned by KIRIS will come only slowly, even if the program is reasonably effective. In the

case of the Vermont portfolio program, for example, many mathematics teachers reported that it took considerable time for them to locate and test out suitable tasks and even longer for them to find ways to integrate portfolio work with the "regular" mathematics curriculum.

For these reasons, the Panel suggests removing portfolio scores from the accountability index but maintaining the portfolio program as a lower-stakes program may be a reasonable compromise between the measurement difficulties imposed by portfolios and their potential as a tool of instructional reform. If the program is continued in a lower-stakes form, however, the Panel strongly urges that its implementation and impact be systematically evaluated. Large-scale portfolio assessment is a largely untried tool of reform. It is also burdensome (for example, in terms of teachers' time) and diverts resources from other uses. Like any other large-scale and complex intervention, its implementation and effects are likely to vary greatly, showing success in some instances and failure in others. Moreover, while studies have found evidence of positive instructional effects of portfolio programs, they have not fully documented the net effects, taking into account not only what teachers do to implement and support the program, but what other activities they curtail (perhaps in other subject areas) to accommodate the time and other demands of a portfolio system. Finally, even in a lower-stakes context, the reliability of the resulting scores and their validity for their intended uses will be critically important questions. For these reasons, careful

scrutiny of the program and its effects will be essential even if it is continued in a lower-stakes manner.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Gentile, C. (1992). Exploring new methods for collecting students' school-based writing. Washington: National Center for Education Statistics.

Kentucky Department of Education. (1995, March 1). KIRIS biennium I technical manual. Frankfort, KY: author.

Kentucky Department of Education. (1994, March 28). Regional writing portfolio audit meetings final report. Frankfort, KY: Author.

Koretz, D. M., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). Can portfolios assess student performance and influence instruction? The 1991-92 Vermont Experience (CSE Technical Report 371). Los Angeles: Center for the Study of Evaluation, UCLA.

Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5-16.

Lewis, S. (1995, March 3). Presentation to the KIRIS National Technical Review Panel, Louisville, KY.

Stecher, B. M., & Mitchell, K. J. (1995, May). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving (CSE Technical Report No. 400). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Webb, N. W. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. Educational Assessment, 1(2), 131-152.

CHAPTER 5

Are the Scores Comparable Across Administrations?

5.1 Central Questions and Principal Findings

The KIRIS accountability system depends on linking performance in successive years back to the baseline year in which standards were set. Because subsequent assessments are scored in terms of the initial standards set on the baseline year assessment, the equating back to the baseline is critical for ensuring comparability of KIRIS scores. Accurate determination of changes in schools' accountability indices and reward status depend on the equivalence of KIRIS assessment scores across accountability years.

This chapter examines the question: "Are KIRIS scores comparable across administrations?" Topics include the equating of each year's assessments to the baseline scale, forms equating, the choice and implementation of equating models, and other equating issues related to the KIRIS transitional tests.

The major conclusions based on the analyses in this chapter are as follows:

1. Repeated use of ad hoc, judgmental procedures results in an accumulation of equating errors that make year-to-year comparisons of questionable validity.
2. The overall adequacy of the equating was undermined by the changes in procedures used across years, the use of inefficient designs for linking forms, and the exclusion of multiple-choice items from the equating links.

The analysis raises the following concerns which need to be addressed if KIRIS remains a high-stakes accountability system:

1. inconsistency of procedures across years;
2. use of inefficient equating designs;
3. exclusion of multiple-choice data;
4. repeated use of ad hoc, nonreplicable, judgmental adjustments;
5. separate classifications of students based on small numbers of questionably equated common items and matrix-sampled items;
6. unjustified method of equating of performance events and use of nonequated portfolio scores; and
7. combining data from noncomparable performance events, portfolio assessments, and alternative assessments for special education students with transitional assessment data.

Recommendations for improving the equating of performance assessments in the future include:

1. Either use a two-parameter IRT model and allow multiple mappings of the same raw score or, if each raw score must map to only one classification, convert to a one-parameter IRT model.
2. Begin using the multiple-choice items in the accountability index and use them to design stable form-to-form and year-to-year links.
3. Set standards for Novice, Apprentice, Proficient and Distinguished on an intact form and equate all other

forms to it. Base a student's classification in each subject on the total set of multiple-choice and open-ended items taken. Eliminate separate classifications for common and matrix items.

4. Prepare detailed content specifications and specify weights. Use these specifications to build parallel forms for KIRIS. Equating is more accurate when the assessments being equated are parallel.
5. Eliminate all ad hoc, nonreplicable procedures. Specify procedures and rules to be followed in determining adherence to model assumptions, adequacy of model fit, elimination of items from equating, and conversion of equated scores to classifications. Use the same procedures for all form-to-form and year-to-year equatings.
6. Use larger sample sizes to improve the equating of performance events.
7. Check the stability of all items remaining in an equating link. Eliminate those items with large differences.
8. Do not try to equate Grade 12 performance to Grade 11. Establish a new baseline the first year the transitional tests are administered in grade 11.
9. Reconsider the design of KIRIS assessments vis-a-vis the need for stable, accurate equating. Use an adequate number of linking items to allow for expected losses. Design the placement of common items to ensure

acceptable comparability of performance across forms and years.

5.2 Purpose of Equating

When an assessment is used to make high-stakes educational decisions about individual schools, all schools subjected to the requirement should be treated equally and fairly. This implies that the results of the assessment procedure should be comparable from one year to the next. According to Lord (1980), for an equating to be considered equitable, users of assessment results are indifferent as to which form of the assessment is administered.

Judgments about changes in the performance of Kentucky students from year-to-year depends on equating. Without equating it would not be possible to use gains in the percentage of students reaching proficient or distinguished levels as evidence of improved performance since the gain could be due to changes in the difficulty of the assessments rather than increases in student proficiency.

Testing programs that are used for high stakes decisions and that require comparisons of scores across years need new secure assessment forms for future assessment administrations. Elaborate assessment construction procedures are needed to build parallel assessment forms that are approximately equal in difficulty. Assessment equating procedures then need to be used to adjust assessment scores to a common scale so that scores on different forms and in different years are comparable.

For example, for the KIRIS assessments in the 1991-92 baseline year, standards were set which required students to achieve specified scores in each subject area in order to be classified as proficient. The purpose of equating in subsequent years was to determine equivalent "proficient" scores for the 1992-93 and 1993-94 assessments which represented comparable levels of performance. This was necessary because the assessments for the three years were not equally difficult.

5.3 Levels of Equating

The KIRIS assessments are designed so that two levels of equating are necessary: equating across forms within a year and equating across years. The first level ensures that scores on different forms of the assessment given in a single year are equivalent. For example, the 12 forms of KIRIS administered in the 1991-92 baseline year each contained different matrix items which differed in difficulty. Students who were administered forms with harder matrix items should have the same opportunity to be classified Proficient as students who were administered forms with easier items. Thus, all 12 forms must be placed on a common scale that adjusts for differences in difficulty and makes student performance comparable across forms. The same was also true for the two subsequent years in Biennium I.

The purpose of the second level of equating is to place KIRIS assessments given in different years on a common scale. This allows a fair comparison of schools from one year to the next. Because categories of school level performance are based

on gains over a two-year period, it is imperative that the scores used to calculate the gains be comparable. Because the items used one year may be more difficult than those from another year, adjustments must be made so that schools have an equal opportunity to have their students classified as proficient in the first and second years of the biennium as in the baseline year.

For example, if the year two assessment was more difficult than the baseline year assessment but no adjustment was made, students would have to demonstrate a higher level of knowledge and skill to be classified as Proficient in year two than in the baseline year. This could result in a school having fewer students classified as Proficient in year two even when student achievement had remained constant. Conversely, if the year two assessment were easier than the baseline year assessment and no adjustment was made, schools which had really made no progress could appear to have increased the percentage of Proficient students.

5.4 Equating with Item Response Theory Models

Item response theory (IRT) models are used for equating in many statewide assessment programs. The most commonly used IRT models assume a single underlying ability that governs the responses of individuals to a set of assessment items.

In the IRT estimation of ability from observed raw scores, the only distributional assumption necessary is that ability be an unbounded and continuous variable. The intent of IRT is to

remove person and item effects from the estimation of student abilities and item parameters.

The application of the most commonly used IRT models requires two additional assumptions:

1. unidimensionality of the underlying ability being measured by the assessment; and
2. adequate fit of the data to the particular item response model that is used (e.g. a two-parameter logistic model).

Unidimensionality of the ability scale implies that the assessment measures only one ability parameter (or "latent trait," "aptitude," or "achievement") for all students.

Three basic IRT models can be used to equate assessments that consist of items scored zero or one. They are designated by the number of parameters estimated for each assessment item. The Rasch or one-parameter IRT model estimates a difficulty parameter for each item that is analogous to a p-value. The two-parameter IRT model estimates a difficulty parameter and a discrimination parameter (analogous to a biserial correlation). The three-parameter IRT model estimates difficulty, discrimination, and a "guessing" parameter. The "guessing" parameter represents the probability that a student with no knowledge will correctly answer an assessment item by luck or guessing.

The one-parameter and two-parameter IRT models can be used with multiple-choice items and can be generalized to deal with open-ended items that involve more than two score levels. When used with open-ended items involving multiple score categories,

IRT models estimate the difficulty of moving from one score level to the next higher score level. Thus, each open-ended item has multiple difficulty values. The three-parameter IRT model is generally not used with open-ended items because educators believe that, unlike multiple-choice items where a student with no knowledge could select the correct answer by guessing, students with no knowledge answering open-ended items can not construct a correct answer by guessing.

When IRT models are used, there are three choices for use with dichotomous items. For two of those choices, the one-parameter and two-parameter models, there are extensions of the model for tasks with more than two scoring categories such as are common with KIRIS.

Choosing an IRT model for equating requires professional judgment about adherence to the assumptions of different models. The critical issue for selection of an equating model for open-ended items is whether more accurate results can be obtained when discrimination is part of the model than when it is not. There are tradeoffs to consider when answering this question. Adding an additional item parameter to be estimated increases complexity and cost. Larger sample sizes are needed for two-parameter models and measurement error must be small enough that real differences between items can be estimated accurately. For example, in some applications, significant numbers of discrimination parameters may appear to be out of range and be set to default values.

5.5 Fit Statistics

IRT models also provide an internal method of evaluating the fit of the data to the model. This evaluation facilitates the identification of items providing information inconsistent with the remaining items of the assessment. The issue of model-data fit is an important one since misfit may disturb the desirable measurement properties of the model.

Prior to the computation of the linking constants, common (linking) items that do not fit the model can be eliminated. The purpose for doing so is to eliminate any biases created by misfitting items that might be functioning differently than the other items in the assessment. Deletion of unstable items in the link can also function to eliminate items that might disturb the equating results. Thus, it is particularly important to consider fit statistics or stability estimates when IRT models.

5.6 Professional Standards

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985) state:

> Standard 3.8
>
> When parameter estimates of item response curves are used in test development, the item response model and calibration procedures should be specified clearly. Information about the degree to which important [IRT] assumptions (e.g., unidimensionality or equality of slope parameters) are satisfied should be presented in order to demonstrate the adequacy of the fit of the model to the data. The sample used for estimating item parameters should be described and should be of adequate size and diversity for the estimation procedure. (p. 27)

Comment

Although overall sample size is important, it is important also that there be an adequate number of cases in regions critical to the determination of item parameters. (p. 27)

Standard 4.6

When scores earned on different forms of a test . . . are intended to be used interchangeably, data concerning the parallelism of the forms should be available. Details of the equivalence study should be available, including specific information about the method of equating; . . . Periodic checks on the adequacy of the equating should be reported. (p. 34)

Standard 4.7

Content specifications that are changed from an earlier version of a test to a later version should be identified in the test manual, and an indication should be given that converted scores may not be strictly equivalent. In addition, when radical shifts in test specifications occur, either a new scale should be introduced or a clear statement should be provided with the scores to alert users that the scores are not interchangeable with those on earlier versions of the test. (p. 34)

Standard 4.9

Continuing testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale. (p. 34)

5.7 Application to KIRIS

The equating model chosen for KIRIS was a two-parameter graded-response IRT model. Only open-ended items were used in the equating. The questions of interest are whether the KIRIS data satisfied the unidimensionality, and mathematical form of this IRT model.

Unidimensionality requires each KIRIS assessment to measure a single trait. Factor analysis has often been used to provide

evidence regarding the dimensionality of an achievement assessment. Prior experience with factor analyses for multiple-choice tests indicates that reading and mathematics usually have large first factors which might be labeled reading ability and mathematics ability, respectively. Mathematics will often have several smaller factors which appear to be related to specialized areas of mathematics such as geometry or measurement. But generally the first factor is large enough relative to the smaller remaining factors to believe that it is reasonable to assume unidimensionality.

No factor analysis data is currently available for KIRIS. However, one might speculate that unidimensionality may be more problematic here. The KIRIS open-response items in all subjects require extensive reading and significant amounts of writing.

Local independence (which is equivalent to unidimensionality) is problematic for any assessment that uses the same stimulus materials for more than one question. This concern is not unique to open-response items. It is violated when multiple questions are used with a single reading passage, when multiple questions are asked about a mathematical graph or figure or when multiple questions are asked about a scientific experiment. For such situations, responses to the dependent questions are more highly correlated than with other independent questions. Careful assessment construction can minimize these effects.

Because KIRIS equating was based only on the open-ended items, guessing was not an important factor in the choice of an IRT model. However, had the multiple-choice items been included,

a choice between the two-parameter and three-parameter models would have been necessary.

5.7.1 Positive Aspects of KIRIS Equating

Several aspects of the KIRIS equating are praiseworthy. The use of the graded-response IRT model for the open-ended responses provided the advantages of IRT equating while taking into account the use of multi-point scoring rubrics for the open-ended items. The design for equating performance events using out-of-state subjects was optimal under the given constraints. Attempts to monitor portfolio scoring through rescoring checks acknowledged the potential unreliability inherent in portfolio scores. As described in Chapter 4 of this Report, monitoring also provided the basis for evaluating the potential bias in scores awarded by local teachers in comparison to independent raters.

Unfortunately, many of the benefits of these praiseworthy actions were lost because many changes were made from one year to the next. A proactive equating plan which builds on these strengths and is incorporated in the KIRIS assessment design from the beginning could significantly increase the accuracy of KIRIS equating.

5.7.2 1991-92 Linking Procedures for Transitional Assessments

For 1991-92, the common and matrix open-ended items for the 12 forms of KIRIS within a subject area were linked using a regression procedure. (Multiple-choice items were not included in the linking process.) Categorical distributions on the common

items were converted to standardized normal scores based on interval midpoints. Each of the 12 forms contained one matrix item rated on a five-point scale. Dummy-coded variables represented the five possible ratings and were the independent variables in a regression which predicted the standardized normal scores for the three common items. The 60 predicted scores for the 12 matrix items were then rank ordered and cutoffs established such that the proportion of students in each category corresponded as closely as possible to the corresponding statewide proportions for the common items. Based on these cut-off scores, predicted standard normal scores for each matrix item were assigned classifications of Novice, Apprentice, Proficient or Distinguished. Thus, each student assessed on KIRIS in 1991-92 received two classifications: one based on standards set for the three common open-ended items and one based on the regression for a single matrix item.

The linking procedure described for the 1991-92 KIRIS assessments is not consistent with the stated purpose for including matrix items. According to the 1991-92 Technical Report (Kentucky Department of Education, 1993):

> The technique of matrix sampling was used to increase the coverage of the assessment . . . . Matrix sampling . . . divides [items] among different forms of the test; each student responds to only one [of 12] form[s] including all of the [3 open-ended] common questions and a fraction [1 open-ended] of the matrix-sampled items (p. 7).

The above explanation indicates that the Department of Education wanted to assess a broader spectrum of content than could be covered by the common items. Coverage was increased by adding 12

different open-ended items, one per form. Thus, the matrix open-ended item included in each form covered something different than the content covered by the common items that each student took. Because the common and matrix items were assessing different knowledge/skills, there was no reason to believe that students would perform equally well on the matrix and common items. In fact, a student might do well on the common items but score poorly on a difficult matrix item for which the student lacked the assessed knowledge/skills.

For example, suppose a student had good math skills through first year algebra but had never taken geometry. Suppose further that the three common items covered topics through first-year algebra and the matrix item covered geometry. The student might score well on the pre-algebra and algebraic common items but miss the matrix geometry item entirely. Therefore, the student might be Proficient on the common items but only a Novice on the geometry item.

Estimation of school-level performance on a content domain can be improved by matrix sampling of items. However, equating procedures that force the distribution of student classifications to be the same for matrix and common items, as was done for KIRIS, bias the results.

The 1991-92 KIRIS linking procedures assume a linear relationship between performance on the common items and matrix items. That is, the regression procedure assumes that performance on the matrix item should predict a student's classification on the common items. Given that the matrix items

were not constructed or selected to be parallel to the common items, such an assumption is not justified.

Further, the 1991-92 linking procedure assumes that the statewide classification distribution of students (percent Novice, Apprentice, Proficient, Distinguished) should be the same for the matrix items as for the common items. Again, because the set of matrix items measures different knowledge/skills than the set of common items, and because no attempt was made to link the common and matrix items to a common scale, this assumption also is not justified. Given that different content was being measured, it is reasonable to suppose that fewer Kentucky students might be Proficient on the content of the matrix items than on the content of the common items, or vice versa.

Finally, the regression procedure used to link the common and matrix items in 1991-92 was flawed because no attempt was made to adjust for differences in difficulty across the matrix items taken by different students nor to account for the avoidable error due to basing a student's matrix classification on a single item. The relationship between performance on the common items versus the matrix items could be expected to differ depending on the difficulty of the matrix item. Low-ability students who received a hard item and high-ability students who received an easy item may have been unable to demonstrate their true skill level. In addition, classifying a student based on a single item from a large content domain (i.e. reading or mathematics) will result in large errors. Even three common items may not be enough to reliably determine a student's

classification (see Reliability in Chapter 3). This is important because significant errors in student classifications aggregate into inaccurate school-level indices.

A Possible Alternative. There are many ways to equate tests. A more parsimonious, efficient and straightforward alternative procedure for linking the 1991-92 forms in each subject area would have been as follows. Calibrate the common and matrix items for all 12 forms simultaneously using an IRT model. Designate one form as the base form and set standards for Novice, Apprentice, Proficient and Distinguished based on all items from that form. (One could use only the open-ended items but a stronger link and a more reliable estimate of student performance would be obtained if the multiple-choice items were also included.) Using the parameter estimates for the items on the base form and each of the other 11 forms, calculate the equated score on the base form scale equivalent to each possible score on each of the other forms. Use the base form scale equated scores for each form to categorize student performance.

Because all 12 forms will have been equated to a common scale, student classifications across forms will be comparable. In addition, the classification for each student in each subject will be a straightforward assignment based on all available data (common and matrix) from the transitional assessment.

Reconstructing Lost Data. Fall retesting was used to reconstruct data for schools whose 1991-92 spring assessment results were destroyed by fire while in transit. The equating used to adjust the fall results introduced an unknown additional

amount of error into the system. Although this act of nature was beyond the control of the Kentucky Department of Education, Advanced Systems, or the schools, it posed a difficult dilemma for maintaining comparability of data in the accountability system. In an appeal of its Biennium results, Huntertown Elementary School in Woodford County School District reported, "When retesting, some students took the exact same form of the test. All students had seen all the common items" (Appeal 5, Summary Appeals of Accountability Cycle I Performance Judgments, March 23, 1995, p. 6).

5.7.3 1992-93 Linking Procedures for Transitional Assessments

Equating Designs. Equating designs should be chosen before the first assessment is given. They should take into account the uses of the data and the need for strong links across forms and years. Parsimony is desirable. The least complex model that provides adequate error reduction should be chosen.

The inclusion of multiple-choice items in the common anchor increases the total number of items in the link and enhances the stability of the equating. Form-to-form and year-to-year equating is most accurate when based on adequate numbers of stable common items. Assessment forms should be designed to be parallel and to accommodate equating requirements.

The procedures used for equating should be clearly specified and consistent across years. Such procedures should be capable of independent replication.

Equating Analyses. For the 1992-93 KIRIS assessments, a different procedure was used to link forms. 1991-92 and 1992-93 open-ended items were calibrated using a two-parameter graded-response IRT model. 1992-93 items were linked to the 1991-92 using items that appeared in the assessments both years. This included all five 1992-93 common items and some additional matrix items which had been used both years. Ad hoc, nonreplicable judgmental procedures were used to identify outliers to be deleted from the set of linking items. Based on the remaining linking items, a linear transformation was established for adjusting 1992-93 items to the 1991-92 scale. Using adjusted item parameters, two ability estimates were calculated for each student: one based on the five common open-ended items and the other based on two matrix open-ended items. Based on the abilities which corresponded to the standards set for the common items in 1991-92, students received common-item and matrix-item classifications of Novice, Apprentice, Proficient or Distinguished.

Problems with 1992-93 Linking Procedures. The ad hoc procedures used to judge the stability of item parameter estimates across years for the linking items could have been more precise. Instead of relying on nonreplicable, nonquantitative methods, outliers from a best fitting line with error bands could have been used to identify unstable items.

Prior to final classification of students, an additional scoring adjustment was made. The scoring adjustment was based on 1992-93 rescoring of a sample of 1991-92 responses to linking

items. If the scoring of linking items has changed, they are no longer identical and may not be suitable for year-to-year equating.

Another adjustment to scores involved students who received perfect scores on the two matrix items. Even though IRT ability estimates indicated a category of Proficient for some forms of the assessment, a policy decision was made to designate all perfect scores as Distinguished. While this may have appeared fair on the surface, it actually resulted in differential standards across students. For example, a student who took a form with two relatively easy matrix items might be categorized as Distinguished while a student with a higher estimated ability on two harder matrix items but lacking a perfect score might only be classified as Proficient. Such situations could arise due to uneven sampling of content and difficulty level in matrix items across forms.

Finally, there was an additional ad hoc, judgmental adjustment used in 1992-93 which affected the assignment of students to performance categories. This adjustment preserved a one-to-one correspondence between total raw scores on the common items and the corresponding assigned performance level. Such an adjustment was necessary because KIRIS policy required that each raw score convert to a single performance level. But the use of the two-parameter graded-response IRT model resulted in ability estimates that could differ for students with the same raw score. This occurred because in a two-parameter graded-response model,

students receive more credit for the same score on a highly discriminating item than on an item with low discrimination.

For example, suppose the common items consisted of two high discrimination and three low discrimination items. If two students both had a total score of 12 on the five common items (two 3s and three 2s), they could have different estimated abilities and be classified at different levels if one student received the 3s on the two high-discrimination items while the other student received the 3s on two low-discrimination items. Thus, even though the two students had the same total raw score, the student with higher ratings on the more discriminating items would have a higher ability estimate under a two-parameter graded-response IRT model. This could result in one student being classified as Proficient while the other was classified Apprentice.

To force identical classifications for students with identical raw scores, the distribution of classifications by raw score from the two-parameter graded-response IRT model was arbitrarily revised so that in most cases, the classification with the largest frequency for a given raw score was assigned to all students with that raw score. However, classifications were not allowed to decrease as raw scores increased. In addition, when students at a given raw score were nearly evenly split between two classifications and there was no adjacent classification to balance the adjustment, classifications in two categories for a single raw score were retained. No rules are

available for making these adjustments so these judgments are not independently replicable.

The use of the ad hoc adjustment procedure which forced a one-to-one correspondence between raw scores and classifications was inconsistent with the results from the two-parameter graded IRT model. The same result could have been achieved by using a one-parameter graded IRT model. Because one-parameter IRT models estimate only difficulty for each item, they result in a one-to-one correspondence between raw scores and ability estimates. Therefore, using a one-parameter model, students with identical raw scores would have identical ability estimates and identical classifications. The inconsistency created by the ad hoc adjustment should be removed by either using the two-parameter results intact or converting to a one-parameter solution.

5.7.4 1993-94 Linking Procedures for Transitional Assessments

The two-parameter graded-response IRT model was used again to link the 1993-94 assessments to the 1992-93 assessments which had already been placed on the common 1991-92 scale. However, several modifications were necessary due to changes in the scoring of open-response items in 1993-94 and differential placement of potential linking items in the two assessments.

Some of the potential 1992-93/1993-94 linking items were matrix items at the end of the 1992-93 assessment and common items at the beginning of the 1993-94 assessment. Due to fatigue effects which may have affected performance on the matrix items, the decision was made to eliminate matrix items reused as common

items. Items remaining in the linking set were matrix items at the end of the assessment in both years. This decreased the potential linking set by five items; remaining items ranged from 3 to 19 across grades and subjects.

The scoring change involved the division of the 1992-93 "1" score of minimal or off-topic response into two categories in 1993-94: "0" = off topic and "1" = minimal response. No information was available on possible effects of this change on other score points but calibrations for reused items remained approximately linear. Therefore, the decision was made to base the equating only on the upper, unchanged score points.

Problems with 1993-94 Linking Procedures. Equating of the 1993-94 assessments to 1991-92 is dependent on a two-step chain: equating of 1993-94 to 1992-93 and equating of 1992-93 to 1991-92. Some direct link going back to the original year would have been desirable to provide a check on possible drift. Such links back more than one year will become more important in future years of KIRIS to satisfy Standard 4.9 quoted earlier.

One of the differences between the 1992-93 equating and the 1993-94 equating is that the rescoring study was used to make adjustments in 1992-93 but not in 1993-94. Adjustments for possible changes in scoring standards from one year to the next were based on small samples of rescored responses. Rescoring adjustments for 1992-93 to 1991-92 were based on 200 per grade for Grades 4 and 8 and 500 at Grade 12.

But in 1993-94, it was decided that no additional adjustments based on the 1993-94 rescoring of 1992-93 responses

of reused items would be made. On the basis of results of the rescoring of the 1992-93 responses by 1993-94 scorers "it was determined that any possible scoring differences from 1992-93 to 1993-94 were not major enough to warrant any changes in the 1993-94 data" (Biennium I Technical Manual, 1995 draft, p. 144). It is unclear what the precise differences in ability estimates were that were judged to be "not major enough" or how those compared to differences in the rescoring of 1991-92 responses by 1992-93 scorers that apparently were considered major enough to warrant adjustments.

Besides the scoring changes described earlier, equating accuracy may have been affected by repetition of the same ad hoc procedures discussed earlier for 1992-93. In addition, because the classification results for the common and matrix items differed significantly, a further adjustment was made to align the classification distributions for the two types of open-ended items. Given that the assessed content may have differed for the common and matrix items, it is possible that the student classification distributions could have been reflecting real differences in performance.

The Biennium I Technical Manual states:

> It was found that matrix-based ability estimates yielded a distribution of performance levels in which 50% of the students were classified as Novices. Whereas scores from previous years had yielded equal percentages across matrix and common items, only 33% of the students were classified as *Novice* on the basis of the common items. To remedy this discrepancy, matrix-item-based performance level cut scores on the ability estimate scale were adjusted to produce percentages of *Novice, Apprentice, Proficient* and *Distinguished* students equivalent to the common item-based percentages. Because the five common items produced

> ability estimates containing less error than the two item, matrix-based estimates, it was reasoned that the common item-based performance level percentages more accurately reflected the underlying ability of Kentucky students. Therefore, the matrix-based cut score point were adjusted rather than the common item-based cut score points. (Kentucky Department of Education, 1995, p. 147)

As the quote indicates, before adjustment, the matrix-based approach resulted in 50% of Kentucky students being classified in the Novice category, whereas the common-items approach resulted in only 33% of the students being classified in the Novice category. Question: Is the decision to adjust the matrix-based cut score points to yield "percentages of Novice, Apprentice, Proficient, and Distinguished students equivalent to the common-item based percentages" (Kentucky Department of Education, 1995, March 1, p. 143) justified? Note that this works in the direction of increasing 1993-94 performance since matrix-based estimates were adjusted from the original 50% of students classified as Novice to the common-items estimates of 33% of students classified as Novice.

In addition, the latter portion of the above quote seems to contradict the notion that the purpose of the matrix items was to measure content not covered by the common items. If one assumes that students do not perform equally across subskills within a content area, adjusting the matrix results to match the common item results seems to negate this purpose. If the matrix items are just a less accurate estimate of the same ability, why not use a single ability estimate based on all items or administer only common items to all students?

5.7.5 Performance Events

Performance events were equated by administering pairs of events from different years to a common sample outside Kentucky. This seems like the best approach that was feasible under the circumstances. However, the sample sizes used for equating performance events were much smaller than desirable (Kentucky Department of Education, 1995, March 1, p. 149). Such small samples lead to large equating errors.

5.7.6 Writing Portfolios

Writing portfolios were not equated. One could include in the outside audit scoring a sample of portfolios from each of the three years and compare the scores to the ones assigned within year, either by within-school raters or by early audit teams. The audit study with its corresponding choice of adjustments for 1992-93 may have affected comparability across years.

5.7.7 Moving the Grade 12 Assessment to Grade 11

Due to motivational problems in Grade 12, a decision has been made to move the Grade 12 assessment to Grade 11. While it is possible to numerically equate the assessments in the two grades, it may not be meaningful to do so. Annual growth at the high school level tends to be small relative to the elementary grades. Thus the effect to be measured in adjusting the Grade 11 results may be less than the equating error. A better procedure would be to establish a new baseline in Grade 11 during the first year the assessment is moved.

5.7.8 Equating Errors

Equating errors can be both random and systematic. Equating errors for KIRIS include the usual random calibration, sampling and linking errors and systematic errors related to compounding of ad hoc effects. Random errors in KIRIS equatings could have been reduced and systematic errors could have been eliminated. The following factors contributed to avoidable KIRIS equating errors:

1. Changing procedures across years. This partially negated the purpose of equating by producing linked scores that were not equivalent across years during Biennium I.
2. Using inefficient designs for linking forms and equating results across years.
3. Ignoring multiple-choice data. Anecdotal evidence suggests that schools and students were not motivated to perform well on the multiple-choice items after the baseline year because they learned these items were not counted in the school accountability index. Thus, the multiple-choice items could not be used for equating in years 2 and 3 as originally planned. This result could easily have been avoided by using the multiple-choice items for forms equating in the baseline year and including them in the computation of the accountability index. It was a waste of valuable assessment time to administer items which were not used but could have contributed to more accurate equating.

4. Repeatedly using ad hoc, judgmental adjustments that may not have adequately addressed the problems they were designed to remedy introduced additional avoidable error. While any one adjustment may increase error only marginally, the accumulation of such errors across multiple ad hoc procedures is significant.
5. Basing multiple student classifications on small numbers of items when a single classification with a smaller error could have been obtained using all the available information.
6. Including performance events which were equated using small sample sizes and therefore had large equating errors.
7. Including writing portfolios and alternative assessments for which no equating was possible. Adjustments to writing portfolios based on scoring discrepancies and the lack of comparability of alternative assessments to regular KIRIS assessments suggest these noncomparable components contributed significant error.
8. Use of a four-point scale for classifying students as Novice, Apprentice, Proficient and Distinguished also increases equating errors.
9. Lack of adequate assessment security to ensure that common item links across years are not based on spuriously inflated scores.

Although it would be desirable to estimate the overall equating error from all sources, data and resource limitations preclude such an estimation. However, because the equating error in the KIRIS system includes avoidable error, it may be judged unacceptable for high-stakes decision-making.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: APA.

Baker, F. B. (1985). The basics of item response theory. Portsmouth, NH: Heinemann.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston: Kluwer Academic Publishers.

Kentucky Department of Education. (1993). KIRIS 1991-92 technical report. Frankfort, KY: Author.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton-Mifflin.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MEAS Press.

CHAPTER 6

Are the Performance Standards Defensible?

6.1 Central Questions and Principal Findings

The Kentucky Instructional Results Information System (KIRIS) serves two central purposes. By law (Kentucky Education Reform Act of 1990), it is a school accountability program that provides information for classifying the state's schools into five categories – Level 1: (Reward) Schools in which teachers receive financial rewards for the performance of the school, Level 2: (Successful) Schools that receive no reward and are not subject to sanction, Level 3: (Improving) Schools required to develop an improvement plan, Level 4: (In Decline) Schools assigned a Kentucky distinguished educator to assist with the development of an improvement plan, and Level 5: (In Crisis) Schools assigned a Kentucky distinguished educator to develop an improvement plan (see Kentucky Department of Education, 1993, p. 3). By a 1994 act of the Kentucky Legislature, no schools were classified as "In Crisis" during the first biennium of the KIRIS program. Instructional improvement in Kentucky's schools is the ultimate goal of the accountability program. In addition, KIRIS serves an assessment purpose, in that it identifies students whose school achievement in a number of subjects warrants their classification into four categories labeled Novice, Apprentice, Proficient, and Distinguished.

Classification of schools is formula driven, and depends on the initial classification of students who attend the schools. As noted earlier, students are classified into one of four

achievement levels in each of the subject areas that contribute to the computation of an Accountability Index for schools (in 1991-92, Reading, Mathematics, Science, Social Studies and Writing). The proportions of students in a school who are placed into each of the four categories for each of the subject areas forms the basis of the school's Accountability Index.

Determination of the levels of achievement on particular test items or exercises (or combinations of them) that warrant classification of students as Novice, Apprentice, Proficient, or Distinguished is termed performance standard-setting. The validity of the KIRIS Accountability Index and the resulting classification of schools depends critically on the validity of the procedures used to establish performance standards for students in each subject area that contributes to a school's Accountability Index.

Changing the definition of performance that warrants classification of students as Apprentice, say, in just one subject area, such as Reading, will affect the particular students who are so classified, the number of students so classified, the Accountability Index value assigned to particular schools, and the number of schools that are assigned to each of the five KIRIS accountability categories. The same is true of definitions attached to each of the categories used to classify students for each of the subjects that contribute to the Accountability Index. It is not an exaggeration to note that the integrity of the entire KIRIS accountability system depends to a

substantial degree on the integrity of the processes and procedures used to set standards of student performance.

Setting performance standards is a judgmental process. It demands the reasoned deliberation of qualified experts. It is, by definition, a subjective process (but one that is often informed by information on the consequences of subjectively-based decisions). Standard-setting is arbitrary[1] but must not be capricious.[2]

The measurement profession has not identified one best method of setting performance standards. However, the profession has identified a number of characteristics and features that must be present in all standard-setting procedures in its Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), hereafter termed the Test Standards. A central requirement that applies to

---

[1]The word *arbitrary* is used here in accordance with the second meaning provided in the *Oxford English Dictionary* (1971): "Relating to, or dependent on, the discretion of the arbiter, arbitrator, or other legally-recognized authority; discretionary, not fixed (p. 107)." Even though judges used to recommend test standards rarely have legal authority to do so, those who set standards of performance typically have statutory authority to establish education policy.

[2]The word *capricious* is also used in accordance with the second meaning provided in the *Oxford English Dictionary* (1971): "Full of, subject to, or characterized by caprice; guided by whim or fancy rather than by judgment or settled purpose; whimsical, humoursome (p. 335)." Although the words *arbitrary* and *capricious* are sometimes used interchangeably, their connotative meanings are quite different. Reflective consideration and a well-defined judgment process often underlies arbitrariness, but never capriciousness. This distinction is central to the contrast made here.

performance standard-setting in addition to all other testing activities is embodied in Standard 3.1 (p. 25):

> Tests and testing programs should be developed on a sound scientific basis. Test developers should compile the evidence bearing on a test, decide which information is needed prior to test publication or distribution and which information can be provided later, and conduct any needed research (Primary). (emphasis added)

Other test Standards apply more specifically to the establishment of performance standards. They include:

> Standard 6.9
>
> When a specific cut score is used to select, classify, or certify test takers, the method and rationale for setting the cut score, including any technical analyses, should be presented in a manual or report. When cut-scores are based primarily on professional judgment, the qualifications of the judges should be documented. (Primary)

and

> Standard 10.9
>
> A clear explanation should be given of any technical basis for any cut score used to make personnel decisions. Cut scores should not be set solely on the basis of recommendations made in the test manual. (Primary)

Beyond these specific criteria, the Test Standards note (p. 50), in a discussion of certification testing in elementary and secondary schools, that:

> Most education certification testing programs use a predetermined cut score to distinguish passing from failing scores. The cut score becomes the linchpin in the decision process. Research has shown that there can be large discrepancies between the cut scores produced by the most common methods of setting cut scores. Therefore, the reliability and validity of decisions and inferences based on cut scores from educational certification tests need to be studied carefully. (emphasis added)

This chapter contains a technical review of the methods used to establish performance standards on the indicators that contribute to the KIRIS Accountability Index (the basis for classifying Kentucky schools). In addition, it contains suggestions for improvement of methods used to establish performance standards. The chapter begins with a review of features of the KIRIS Accountability Index that are essential to an understanding of performance standards. It continues with a description and evaluation of the methods used to establish performance standards for students whose performances contributed to schools' baseline Indicator values in 1991. Modifications of those methods, as they contributed to computation of Achievement Index values during the first biennium of KIRIS operation (1992-93 and 1993-94) are also reviewed and evaluated. The final section contains an analysis of the sensitivity of school classification to the placement of performance standards for students.

The essential questions addressed in this chapter are (1) Are the procedures used to establish performance standards for KIRIS sound and defensible in light of professional measurement standards? (2) Would changing performance standards materially affect the numbers and percentages of schools classified into the five categories that range from "Reward" to "In Crisis?" (3) What must be done in the future to improve performance-standard-setting for KIRIS?

These questions, and the answers proposed here, are quite important from a policy perspective. If the procedures used to

set performance standards for individual students cannot be defended, then every element of the KIRIS accountability and assessment systems is suspect. Classifications of students and of schools would likely be indefensible. Some schools judged to be eligible for rewards would, in that case, be so classified simply as an artifact of indefensible performance standards, as would some schools judged to be in the lowest of the five KIRIS categories. The same would be true of classifications of students.

The review and analyses that follow in this chapter raise serious questions about the validity of procedures used to establish standards of student performance within KIRIS. The judgmental basis used to establish initial performance standards for common open-ended questions in all subjects is tenuous because of the small samples of items used and the largely undocumented qualifications of the judges used in the standard-setting process. The analytic procedures used to produce performance standards for matrix-sampled open-ended items, although innovative and novel, are without judgmental or empirical foundation, depend on untested assumptions, and are sensitive to biasing statistical and measurement artifacts. Their sole justification appears to be the "limited resources" available to set performance standards appropriately, through a scientifically- based, well- documented, judgmental procedure (Kentucky Department of Education, 1993, p. 57). Somewhat simpler statistical procedures were used to set performance standards for performance events. Again, these performance

standards were formed in an unorthodox way and are based on untested assumptions. Their validity, therefore is unsubstantiated.

The analytic procedure used to set performance standards during the 1992-94 biennium differed in detail and in fundamental assumptions from the procedure used with open-ended items in 1991-92. Initial performance standards required specific patterns of performance on common open-ended items and were therefore, at least in part, conjunctive in nature: high performance on some common items was not allowed to compensate for low performance on other common items. At the end of the 1992-94 biennium, performance standards were determined through an item response modeling procedure that is inherently compensatory. With this procedure, high performance on some items <u>can</u> compensate for low performance on others. The degree to which this change in approach resulted in some schools being classified differently than would have been the case had consistent standard-setting models been used cannot be determined. It is likely that some school classifications are a consequence of this change in models.

An important analysis completed by Richard Hill of Advanced Systems in Measurement and Evaluation (and extended here) shows that classification of schools into the five categories ranging from "Reward" to "In Crisis" is quite sensitive to the setting of performance standards for students. It is clear, therefore, that procedures used to set performance standards for KIRIS matter operationally, not just in theory.

The basic judgmental approach used to set initial performance standards for open-ended exercises should, if conducted by demonstrably qualified judges and if well documented, satisfy the requirements of the Test Standards and produce trustworthy and defensible performance standards. Performance standards used to classify students within a given subject area should be determined judgmentally for all items and exercises in that subject area, whether administered in common or through matrix sampling. As the Test Standards require, the procedure used should be fully described and should be applied by experts whose qualifications are well documented.

No purely analytic modifications of the performance standards established in 1991-92 for common, open-ended items would increase their trustworthiness materially. What is needed instead is a well-conceived, well-documented strategy for developing judgmental performance standards, based on adequate samples of exercises and a sufficient number of carefully-trained, well-qualified expert panelists, for each component of the KIRIS accountability index.

The judgmental approach to performance-standard setting used with open-ended exercises in 1991-92 is both reasonable and defensible, but it should be applied to all open-ended exercises in a subject area. There are no obvious advantages, and several disadvantages (discussed earlier) to setting performance standards on common items, and then using unproven analytic procedures that depend on strong statistical assumptions to estimate performance standards for matrix-sampled items.

Judgmental procedures of the type used in 1991-92 should be used to establish individual performance standards for all open-ended items in a subject area, whether commonly administered or matrix sampled. The two-parameter graded-response IRT model could then be used to estimate the IRT ability-scale equivalent of each judged performance standard for each item, and differences in results across items could be used to inform standard-setting panelists of their inconsistencies through a well-structured, iterative procedure. Residual discrepancies could be eliminated by computing average ability-scale values corresponding to final judgments of appropriate scale points that define boundaries between Novice, Apprentice, Proficient, and Distinguished levels of achievement. Averaging across judgments rendered by multiple panelists and items would contribute to the stability of the resulting performance standards.

Separate judgmentally-derived performance standards should be established for each of the components that contributes to the KIRIS Accountability Index. The kinds of untested statistical assumptions that threaten the validity of current KIRIS performance standards for all but common, open-ended items should be avoided in the future.

6.2 Some Essential Background on the KIRIS Accountability Index and its Use

During the first biennium of KIRIS, schools were held accountable for growth in their students' performances on tests in mathematics, reading, science, social studies, and writing, and for their performances on a variety of non-cognitive

indicators that include, among other factors, their contribution to the school's attendance rate, retention rate, and dropout rate. Students' performances on each of the cognitive indicators were classified as Novice, Apprentice, Proficient, or Distinguished. Schools received point values that contributed to their Accountability Index scores in proportion to the numbers of students they enrolled who were classified in each of these four categories. A student who was classified in the Novice category contributed zero points; a student who was classified in the Apprentice category contributed 40 points; a student who was classified in the Proficient category contributed 100 points; and a student who was classified in the Distinguished category contributed 140 points. A school's Accountability Index value was determined by the average of its students' point values, computed across all test-eligible students it enrolled, and across all cognitive and non-cognitive indicators. Separate school-level standards were set for non-cognitive indicators, and collectively, the non-cognitive indicators contributed one-sixth of a school's Accountability Index value.

A baseline Accountability Index value was established for each Kentucky school using data collected during the 1991-92 school year. From this baseline, an Accountability Index threshold value was established for each school for the end of the 1992-1994 biennium. Each school's progress was evaluated against its threshold using data from the 1992-93 and 1993-94 school years. Schools with an Accountability Index value that was at least one point above their threshold value were

classified as Level 1 (Reward); schools with an Accountability Index value that was between their threshold value and one point above their threshold value were classified as Level 2 (Successful); schools with an Accountability Index value that was below their threshold value but above their 1991-92 baseline value were classified as Level 3 (Improving); schools with an Accountability Index value that was between their baseline value and five points below their baseline value were classified as Level 4 (In Decline); schools with an Accountability Index value that was more than five points below their baseline value were classified as Level 5 (In Crisis). Rewards and sanctions have been assigned to schools depending on their classification.

Clearly, the establishment of performance standards that determine the categorization of students as Novice, Apprentice, Proficient, or Distinguished is at the heart of the KIRIS accountability system, since it is students' individual categorizations that determine a school's baseline Accountability Index value, its end-of biennium Accountability Index value, and its classification into a Level 1 through Level 5 category.

In its introduction to performance standard-setting (Chapter 5 of the KIRIS Biennium I Technical Manual, March 1, 1995 draft, p. 65) the Kentucky Department of Education argues that standard-setting is a less critical issue for KIRIS than is generally the case with high-stakes testing. The essence of the argument was advanced in three points. First, the Kentucky Department of Education argued that since the purpose of KIRIS is to "drive instruction" and there is no reason to believe that this purpose

will be affected by the placement of performance standards, where the standards are placed doesn't matter. Second, the Department argued that since schools are not held accountable for the number of students who meet high standards at any point, but for improvement in those numbers, placement of standards matters little. Third, the Department of Education argued that since most schools began with very low distributions of student achievement in the baseline year, different placement of performance standards wouldn't have affected Accountability Index results or school placement.

The first of these arguments is speculative and untested. It assumes that school personnel will respond positively to information about the categorization of their school, regardless of the basis for that categorization. The second and third arguments have been refuted by analyses conducted by Richard Hill of Advanced Systems in Measurement and Evaluation (discussed in some detail below). The placement of schools in accountability categories is quite sensitive to the specification of performance standards for students.

6.3 How Performance Standards were Set for KIRIS

Initially, performance standards were established for the KIRIS writing assessment. This was done by a Writing Advisory Committee composed of "approximately ten teachers per accountability grade level, staff from the Kentucky Writing Project, university faculty and members of the Department of Education" (Kentucky Department of Education, 1993, p. 54).

Tables 44 and 45 of the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1, pp. 69-70) present affiliation and demographic information (but no information on qualifications) of standard-setting panelists. For the writing assessment, they show that panels collectively included 37 teachers, 5 administrators, 3 staff of the Kentucky Department of Education, and 11 "Others." The composition of panels is not differentiated across grade levels, even though separate performance standards were set for each KIRIS grade.

The basic procedure used was to have Committee members classify scored samples of student writing into the four categories, Novice, Apprentice, Proficient, or Distinguished. Committee members did not know the scores assigned to the writing samples at the time they were classified.

Performance standards were next established for the "Transitional Tests" that were used in the 1991-92 and later KIRIS assessments. These tests contained some questions that were administered to all students in a grade (known as "common items") and some questions that were matrix sampled. The questions were "open ended," thereby requiring constructed responses from students rather than selection among prespecified responses, as in multiple-choice items. A judgmental method, similar to that used for the writing assessment, was used to establish performance standards for the common questions and for combinations of them. Performance standards for matrix-sampled questions were determined through an analytical procedure that was based on their relationship to the common items.

A panel of judges examined actual student work in response to a common question and then classified that work on the four-level "Novice" to "Distinguished" scale. The judgment panels used for the reading, mathematics, science, and social studies questions were described in the 1991-92 KIRIS Technical Report (Kentucky Department of Education, 1993, p. 55) as composed of "staff from two divisions (KIRIS and Curriculum Development) within the Department of Education, and members of the content area advisory committees for test development. The content area advisory committees included people with a range of backgrounds, the majority of them classroom teachers." The KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1, Table 44, p. 69) suggests that the panel used to set performance standards for mathematics (presumably all grades) was composed of 15 people, 11 of whom were teachers; that the panel used to set performance standards for reading was composed of 13 people, 7 of whom were teachers; that the panel used to set performance standards for science was composed of 11 people, 8 of whom were teachers; and that the panel used to set performance standards for social studies was composed of 13 people, 7 of whom were teachers.

The description of standard-setting procedures used for common, open-ended items in reading, mathematics, science, and social studies, provided in the 1991-92 KIRIS Technical Report (Kentucky Department of Education, 1993) includes these elements:

1. Judges were first trained to score and understand the score scale used for the writing assessment.

2. Judges were provided with several examples of scored writing samples and their classification on the "Novice" to "Distinguished" scale.
3. Judges were trained on the scoring guides for the open-ended questions in the content area for which they were to set performance standards.
4. Judges were provided with "several papers that been assigned each score point on the [0 to 4] score scale.
5. Judges independently decided which performance level should be assigned to each score level.
6. Judges met as a group and debated their classifications of scores to performance levels.
7. Judges independently reconsidered their original decisions and, perhaps, altered them.

On pages 55 and 56 of the 1991-92 KIRIS Technical Report (Kentucky Department of Education, 1993), the authors note that

> At this point, there was a measure of consensus about the correspondence between score points and performance levels. Committee members felt that scores of 0 and 1 generally indicated Novice work, 2 was [Apprentice], 3 was Proficient, and 4 was Distinguished. There were several exceptions to this general pattern, as well as a feeling that the scoring guides, as created, did not truly require distinguished-level work to obtain even the highest scores.

This description of procedures used to set performance standards on common open-ended items must be regarded as sketchy at best. No concrete information is provided about the training of standard-setting panelists, the specific questions posed to them, or the nature of their deliberations and discussions. Adequate training to set performance standards would be particularly critical for KIRIS, since the definitions of

performance levels adopted by the Kentucky Department of Education are not subject-matter specific and are far from operational. Although the definition of Distinguished might be regarded as reasonably concrete, the definitions of Proficient, Apprentice and Novice are quite brief and vague. Novice is defined as "The student is beginning to show an understanding of new information or skills" (Celebrate the Progress! 1992-94 Kentucky Accountability Results - Summary News Packet, Kentucky Department of Education, 1995, p. 32). This definition would apply to virtually all students as they encounter new material, regardless of their level of achievement. The definition of Apprentice is almost as vague: "The student has gained more understanding and can do some important parts of the task." More understanding than what? Which parts of the task should be considered important? How many parts constitute "some parts?" The definition of Proficient is "The student understands the major concepts, can do almost all of the task, and can communicate concepts clearly." It implies, incorrectly, that all tasks require communication of concepts and provides no guidance on identification of major versus minor concepts or how understanding of major concepts is to be demonstrated through performance.

The definitions also make Apprentice-level and Proficient performances dependent on the particular task to which they are applied. Since tasks vary substantially in their cognitive processing and problem-solving demands, completion of all parts some tasks would likely be trivial, and certainly not worthy of

the label "Proficient. " Conversely, completion of few parts of a very difficult task might well warrant the label.

Only for the Grade 12 reading common items does the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1, Table 47, p. 72) provide information on the degree to which panelists agreed or disagreed on recommended performance standards. That table shows that seven panelists reached consensus only on the placement of items scored 1 or 2 on a five-point scale, and for two of three common items, were almost evenly divided between Proficient and Distinguished in their mapping of score level 5 and between Apprentice and Proficient in their mapping of score level 4.

A criterion often used to justify the size of a standard-setting panel and, indeed, the success of a standard-setting activity, is the degree to which panelists reached consensus on a recommended performance standard. The standard-deviation of recommended performance standards across panelists is often used as an index of consensus. Since neither the 1991-92 Technical Manual (Kentucky Department of Education, 1993) nor subsequent documents provide this information, it is not possible to determine whether, for subjects other than Grade 12 Reading, panelists achieved consensus in their recommendations or were widely divergent.

After mapping scores for three common items per subject onto the Novice through Distinguished scale, panelists were asked to consider the three items collectively. They developed (with unspecified levels of agreement) mappings that were partially

compensatory and partially conjunctive, shown in Tables 48 through 51 (pp. 73-74) of the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1).

It is highly unusual to establish performance standards through review of students' performance on only three test items. Generalization to labels such as "Distinguished," "Proficient," "Apprentice," and "Novice" on the basis of these very small samples of performance are very unlikely to be trustworthy. The problem is further exacerbated by the partially conjunctive nature of the performance standards established in 1991-92 to set baseline Accountability Index values for schools. For example, on the Grade 4 Mathematics assessment, a student would be classified as "Proficient" on the common test items only if he or she earned a total score across the three items that was in the range 9 to 11, with "no score less than 3." Otherwise, a total score in this range with a score less than 3 on any one item would result in the student being classified as "Apprentice." The reliability of a student's classification as "Proficient" or "Apprentice" therefore depends substantially on the reliability of a single Grade 4 Mathematics item! No analyses of the reliability of the KIRIS assessment (reviewed elsewhere in this report) have addressed this limiting feature of the accountability system.

It is one thing to decide that a student's performance on a single item is characteristic of what would be regarded as Distinguished, Proficient, Apprentice, or Novice performance. It is quite another to label the student's overall performance in a

subject area as Distinguished, Proficient, Apprentice, or Novice based on her or his responses to only three items or, in the case of matrix-sampled items (see below), on the basis of her or his response to a single item.

Descriptions of the backgrounds of standard-setting panelists that are provided in the 1991-92 Technical Report (Kentucky Department of Education, 1993) are quite sketchy and no elaboration is provided in the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1). Although the Test Standards require that the qualifications of those who set performance standards be thoroughly documented, the Kentucky Department of Education has not done so. Even though the organizations from which standard-setting panelists were selected appear reasonable, the Department provides no assurance that the individuals selected were qualified to undertake the standard-setting tasks in which they engaged. Since the procedures used required panelists to have expertise in the teaching of writing as well as in another subject area (such as mathematics), the issue of appropriate qualification is particularly critical.

6.4 Setting Performance Standards for Matrix-Sampled Items in 1991-92

The procedure used to set performance standards for matrix-sampled items depended entirely on a statistical analysis and was not based on expert judgment. The validity of this unprecedented procedure is therefore subject to the validity of untested statistical assumptions, and must be regarded as highly suspect. In particular, the procedure assumes that students' performances

on all items (common and matrix-sampled) are normally distributed despite the discrete four- or five-point scales used to score each item and that the relationship between each common item and each matrix-sampled item is linear. The performance standards that result from the procedure depend on the correlations among students' scores on common and matrix-sampled test items, a statistic that has no logical relationship to the appropriate definition of performance standards.

The linking procedure used to determine performance standards for the 12 matrix-sampled items within a subject area was as follows:

1. For each student, scores on the three common items within a subject area were summed.
2. The normalized z-score equivalent of each common-item sum was computed, so that, across examinees, the transformed distribution of common-item total scores within a subject area was normally distributed with mean zero and variance one (N[0,1]).
3. The normal distribution was divided into four regions, where the percent of students within each interval corresponded to the judgmentally-determined percent that were categorized as Distinguished, Proficient, Apprentice and Novice, based on the performance standard-setting results for common items in that subject area.
4. The normalized z-score equivalents of the percentile midpoints of each of the four regions were then found.

For example, if, as noted on p. 76 of the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1), 10 percent of the students in a subject area were classified as Distinguished on the common items, and 15 percent were classified as Proficient, the Proficient region of the distribution would contain students with percentile ranks between 76 and 90, and the midpoint of the region, in terms of percentile-ranks, would be 83. The normalized z-score equivalent of 83 is 0.9542. This z-score was then used as one value of the dependent variable (the value corresponding to Proficient) in a regression of students' common item z-scores on indicator values corresponding to their scores on matrix sampled items (see Step 5).

5. Vectors of indicator variables were created for students, to represent the students' scores on matrix-sampled items. For example, a student who scored at Level 2 on a matrix-sampled item would receive indicator-variable scores of 0, 0, 1, 0, 0; where the first and second scores of zero indicate that the student did not receive scores of 0 or 1, the score of 1 indicates that the student did receive a score of 2, and the remaining indicator scores of zero indicate that the student did not receive scores of 3 or 4 on the matrix-sampled item.

6. A regression analysis was conducted for each matrix-sampled item to compute a normalized common item z-score estimate corresponding to each of the indicator-variable scores on each of the 12 matrix-sampled items in a subject area. (See General Item Scoring Guide, Table 46, p. 71 of the KIRIS Biennium I Technical Manual, March 1, 1995 draft). These regression analyses produced 60 predicted normalized z-scores (one for each of the five score levels on each of the 12 matrix-sampled items).
7. These predicted z-scores were then rank-ordered, and can be denoted z_1, z_2, ..., z_{60}.
8. The number of students n_i who earned predicted score z_i on each matrix-sampled item was then computed. These numbers were accumulated from the largest of the rank-ordered z_i values to the smallest, across all 12 matrix-sampled items in a subject field.
9. The cumulative percent of students that most closely corresponded to the boundary of the Distinguished category was then computed, and scores on matrix-sampled items with predicted z_i values above this point were then labeled Distinguished. This mapping procedure was replicated for each category, so that every score on each matrix-sampled item was mapped into a category (Novice to Distinguished).

Since each student completed a single matrix-sampled item in a subject area, that student's categorization as Novice, Apprentice, Proficient, or Distinguished on a matrix-sampled item was then based on her or his response to the single item and the prediction of her or his performance on common items through the regression procedure described above.

It should be noted that the method used to set performance standards for matrix-sampled items will be quite sensitive to the correlations between students' scores on matrix-sampled items and students' total scores on common items. For matrix-sampled items that do not correlate highly with total scores on common items, extreme scores on the matrix-sampled item will map into z-scores in the middle of the distribution (i.e., the regression will be relatively flat), so that a very high level of performance on the matrix-sampled item is unlikely to be recognized as Distinguished and a very low level of performance on the matrix-sampled item is unlikely to be recognized as at the Novice level. In contrast, matrix-sampled items that correlate more highly with total scores on common items will produce steeper regression functions and extreme scores on these items will more likely map into extreme categories of achievement. This will happen regardless of the relative difficulty of the matrix-sampled and common items. Since the steepness of the regression functions does not depend on factors that should influence judged quality of performance (i. e., whether the performance would, if subject to expert judgment, be regarded as Distinguished, Proficient, Apprentice, or Novice), but rather on the reliability of individual item

scores and on the uniqueness of the specific subject matter assessed by a matrix-sampled item, compared to common items, there is no empirical or logical justification for use of the regression method described above to categorize students as Distinguished, Proficient, Apprentice, or Novice, based on their performance on matrix-sampled items. Indeed, the results are likely to be artifactual.

No information in the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1) or in earlier technical reports permits an assessment of the strength of regression relationships between matrix-sampled and common items or of the variation in regression slopes across matrix-sampled items. Tables of biserial correlation coefficients on pages 90 through 93, 97 through 100, and 104 through 107 in the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995, March 1) do not distinguish between common and matrix-sampled items in mathematics, reading, science, and social studies. In sum, it appears to be impossible to ascertain the seriousness of this problem from published information on KIRIS.

When computing the point-value to be attached to a student's performance in a subject area based on data collected during 1991-92, the student's performance on the three common items and his or her performance on a single matrix-sampled item were weighted equally. Thus a student whose performance on the three common items in reading was judged to be at the Apprentice level would be counted as contributing 40 points to a school's accountability index, based on her or his performance on those

items. But if the same student's performance on the single matrix-sampled item was mapped into the Novice level by the regression procedure, that student would be counted as contributing 0 points to a school's accountability index, based on her or his performance on that reading item. Thus the student would contribute an average of 20 points to her or his school's accountability index through her or his performance on the common and matrix-sampled reading items. Viewed in this way, the procedure used to determine the categorization of a student's performance on matrix-sampled items assumes substantial importance from the perspective of a school's accountability classification and from an assessment perspective as well.

6.5 Setting Performance Standards for Performance Events in 1991-92

No judgmental procedures were used to set performance standards for Performance Events. Instead, it was assumed that, within a subject area, the percent of students who should be classified as Distinguished on a performance event should be the same as the percent of students so classified on the common open-ended items. The same assumption was made about the percent of students who should be classified as Proficient, Apprentice, and Novice. An empirical procedure that divided the distributions of scores on Performance Events so that they most closely approximated the desired distributions was applied.

The only rationale provided for these assumptions and the procedures employed was the "availability of experienced people"

... "limited familiarity with students' mastery of performance events" and the need to develop performance standards "within the resources of the program" (Kentucky Department of Education, 1995, March 1, p. 79).

The resulting performance standards must be regarded as ad hoc, absent a defensible rationale, and lacking a sound scientific basis, as required by the Test Standards. Very tenuous performance standards developed on the basis of judgmental review of only three common open-ended test items by a small panel of persons of indeterminate qualification have been extrapolated to an entirely different type of performance assessment with no evidence whatsoever of the validity of the extrapolation.

6.6 Setting Performance Standards for Open-Ended Items in 1992-93 and 1993-94

No judgmental standard-setting activities made use of data on tests used in the first biennium of KIRIS during the 1992-93 and 1993-94 school years. Instead, Samejima's two-parameter graded-response IRT model was used to estimate item parameters and students' abilities in each subject area that was tested with open-ended items. Then, compensatory performance standards were established on the IRT ability scale within each subject area, based on the assumption that the overall percent of students that fell into each of the four achievement categories – Novice, Apprentice, Proficient, and Distinguished – statewide, within each grade, in 1991-92, would be the same

under the judgmentally-based conjunctive standard-setting and under the IRT-based compensatory model. However, some deviations from this assumed outcome were incorporated in practice, so as to have some students fall into the Distinguished category, and so as to avoid reversals between score-scale values and categorizations. That is, a higher score could not map into a lower category under the conjunctive model than did a lower score under the compensatory model.

The quality and defensibility of KIRIS equating procedures are discussed in another chapter of this report. The only point to be made with respect to performance-standard-setting here is that, once again, the tenuous performance standards established using just three open-ended items per subject area in 1991-92 were once again extrapolated using analytic rather than judgmental procedures, this time to successive years of assessment. The quality of these performance standards can be no greater than the quality of the original standards, established during the 1991-92 school year, and are likely to be markedly lower because of errors in equating the standards across years. The Panel's judgments concerning the quality of the 1991-92 performance standards were noted above.

6.7 Does the Location of Performance Standards Matter?

As mentioned earlier, the Kentucky Department of Education has advanced the argument that the specific location of performance standards is of little consequence since the determination of schools' classification as to level depends not

on their students' performance in a single year (or pair of years), but on the difference in students' performances across a biennium. Richard Hill of Advanced Systems in Measurement and Evaluation completed a useful analysis that illuminates this question. The nature and results of the analysis are described below. Hill's findings have been re-summarized so as to more clearly indicate the effect on the distribution of school classifications of modifying the performance standard for students.

Hill investigated the effect on 351 Grade 8 schools of changing the performance standards that defined the Distinguished and Proficient categories, by lowering the performance standards on the item-response-theory (IRT) scale by half the distance between their original positions and the next-lower original performance standard. Originally, the performance standards were located at -.692, 1.154, and 2.361 logits to separate Novice from Apprentice, Apprentice from Proficient, and Proficient from Distinguished, respectively. These performance standards were changed to -.692, 0.550, and 1.757 logits.

The results of these modifications of performance standards on the percent of schools that would be categorized as Level 1 (Reward) through Level 5 (In Crisis) are shown in Table 6.1, below. The rows of the Table define school levels under the original performance standards, and the columns of the Table define school levels under the modified performance standards. The body of the Table contains numbers of schools, by level,

under the original and modified performance standards. The numbers in parentheses are "row percents." That is, they denote the percent of schools that had been classified at the level indicated in the row heading under the original performance standards, that would be classified at the level indicated by the column heading under the modified performance standards.

The row and column headed "Marginal Distribution" show the marginal distributions of schools assigned to each of the five categories under the original performance standards (see the rightmost column) and under the modified performance standards (see the bottom row of Table 6.1).

Table 6.1. Distributions of schools across five accountability categories under original and modified performance standards.

Classification of Schools Under Original Performance Standards	Classification of Schools Under Modified Performance Standards					Marginal Distribution
	Level 1 - Reward	Level 2 - Successful	Level 3 - Improving	Level 4 - In Decline	Level 5 - In Crisis	
Level 1 - Reward	106 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	106 (30%)
Level 2 - Successful	32 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	32 (9%)
Level 3 - Improving	37 (21%)	22 (13%)	109 (63%)	5 (3%)	0 (0%)	173 (49%)
Level 4 - In Decline	0 (0%)	0 (0%)	8 (21%)	25 (64%)	6 (15%)	39 (11%)
Level 5 - In Crisis	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)	1 (0%)
Marginal Distribution	175 (50%)	22 (6%)	117 (33%)	30 (9%)	7 (2%)	351 (100%)

Inspection of the marginal distributions of school classifications shows clearly, the substantial effect of modifying just two performance standards – the standard that

separates Apprentice and Proficient, and the standard that separates Proficient and Distinguished. Under the original performance standards, 30 percent of the 351 schools were classified as Level 1 (Reward), 9 percent as Level 2 (Successful), 49 percent as Level 3 (Improving), 11 percent as Level 4 (In Decline) and less than 1 percent as Level 5 (In crisis). After the two performance standards had been modified, schools classified as Level 1 (Reward) increased markedly from 30 percent to 50 percent, schools classified as Level 2 (Successful) decreased slightly from 9 percent to 6 percent, schools classified as Level 3 (Improving) decreased markedly from 49 percent to 33 percent, schools classified as Level 4 (In Decline) decreased slightly from 11 percent to 9 percent, and schools classified as Level 5 increased from less than 1 percent to 2 percent. The most notable changes are the 20 percent increase in schools eligible for reward and the 16 percent decrease in schools classified as "Improving." These changes as a function of the performance standards for students are not trivial, and would have a substantial effect on KIRIS accountability results and on the overall cost of KIRIS operation, were they actually to be made.

One must conclude, therefore, that the placement of performance standards for students matters a great deal, despite the use of change scores in determining the categorization of individual schools. If performance standards are lowered, the percent of schools eligible for rewards could be affected materially. The impact of changing performance standards

depends, of course, on which standards are altered, and the degree to which they are altered. Raising the performance standard that divides Novice and Apprentice performance could materially increase the number of schools classified as "In Crisis," just as lowering the performance standard that divides Proficient and Distinguished materially increased the number of schools classified as eligible for a reward.

The sensitivity of school classifications to the placement of performance standards is not surprising, but it is of substantial concern for KIRIS because of the tenuous scaffold on which KIRIS performance standards rest. For virtually all components of the KIRIS school accountability index, performance standards were linked to judgmentally-based standards for the open-ended items administered during the 1991-92 school year. And these judgments are grounded in review of only three open-ended test items per subject area. As noted earlier, the links are statistically based, are subject to the influence of a number of statistical artifacts, and depend on a host of untested assumptions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: APA.

Kentucky Department of Education. (1993). KIRIS 1991-92 Technical Report. Frankfort, KY: Author.

Kentucky Department of Education. (1995, February). Celebrate the Progress! 1992-94 Kentucky accountability results - summary news packet. Frankfort, KY: author.

Kentucky Department of Education. (1995, March 1). KIRIS biennium I technical manual. Frankfort, KY: author.

Kentucky Education Reform Act (KERA). (1990). House Bill No. 940; School accountability provisions: Kentucky Revised Statutes 158.645 et seq. Reproduced in Kentucky Department of Education, 1994 Kentucky Schools Laws Annotated, Banks-Baldwin Law Publishing Company, 1994.

CHAPTER 7

Are the Reports of Assessment and Accountability Results Clear and Understandable?

7.1 Central Questions and Principal Findings

Whatever the merits of the Kentucky educational accountability and assessment system may be for improving instruction and learning, some of these merits will not accrue if the reports containing assessment results are not clear and understandable to the intended audiences. The task of producing clear and understandable reports is surprisingly difficult for state departments of education. This is because the intended audiences of policy-makers, administrators, educators, students, parents, and the media are not usually familiar with assessment jargon (e.g., reliability, validity, scaled scores, percentile ranks) and assessment practices, and reporting scales are often confusing to them (Hambleton & Slater, 1994).

One only has to look at the confusion currently over the interpretation of ACT scores with a mean or average score of 21 and SAT scores with a mean or average score of 500. It is hard for many persons to understand how a difference of 4 points on the ACT score scale could correspond to about the same difference in achievement as 100 points on the SAT scale. Even such totally different concepts as percentages and percentiles are routinely confused by educators and this confusion was one of the reasons for the College Board revising the SAT scale earlier this year.

To the extent that the intended audiences are misinformed about the meaning of the results for whatever reason (e.g., overly complicated report forms and/or confusing report scales),

the potential value of the assessment information is reduced and may even lead to misdirected activities on the part of those persons misinformed. Teachers who confuse percentiles with percentages, or policy-makers who fail to understand the cognitive differences between students at the Proficient level and the Distinguished level will not be able to use the test results in meaningful ways or simply not use the information at all. The media, to the extent that they are misinformed or misinterpret KIRIS results, will send the wrong messages about schools to the public; administrators will not be able to correctly implement the educational reforms which are needed; teachers can make major mistakes in their instructional plans; and students (and their parents) could have totally wrong impressions of student performance levels and academic strengths and weaknesses.

The main question addressed in this chapter is "Are the reports of accountability and assessment results prepared for policy-makers, administrators, educators, students, and parents clear and understandable?" It was not possible to collect any direct information on this topic. That is, it was not possible to interview persons from the audiences for whom reports are prepared to determine their opinions about the reports and determine the reports' levels of understandability. Nor was there any evidence provided that reported on the Department of Education or Advanced Systems' efforts to field test their report forms or to compile evidence after the fact that the report forms were meeting the needs of users. Therefore, the only option left

was to review a sample of reports and make some judgments ourselves about the clarity and understandability of the reports the Panel reviewed.

The Panel's review of a sample of reports indicated that, in the main, the reports were generally acceptable. At the same time, there appears to be considerable room for improvement in score reporting. For example, two statistical concepts which are important are absent from nearly all of the reports. Readers need to know that the numbers they are looking at contain measurement error and they need to have an idea of the amount of that error in the scores they are using. Second, the concept of variability is absent from most of the reports where it would be useful to include. For example, when reporting the average of the accountability indices across schools, it would also be useful to know about the extent to which schools differ from the average. There are several ways in which variability might be reported, including the use of graphs to display variability.

Based upon the Panel's review of the KIRIS 1993-94 score reports, three main recommendations seem justified:

1. All score reports should be field-tested on the intended audiences to identify potential problems with the report forms. Field-testing provides a basis for the early detection of potentially faulty report forms. Also, documentation of the results of these field-tests should be reported in the Technical Manual.
2. Any new report forms should conform to the four Test Standards given in Section 7.2. Information about the

size of the error estimates in various test scores is needed, for example, to comply with Standard 2.1. Reporting of error bands to minimize overinterpretations of the results is one important addition to the current reports.

3. The Interpretive Guide needs to be considerably expanded to focus on detailed explanations of the score reports it purports to describe, and the various scores and scales which are contained in these reports. The concepts of errors of measurement and reliability need to be introduced along with, perhaps, a glossary to define such terms as normal curve equivalents and percentile ranks. All in all, readers need to be given the details necessary to interpret the score reports correctly. Examples of correct and incorrect interpretations of results might also be useful. Common questions which have been asked over the years together with clear answers could also enhance the usefulness of this important report.

Several minor recommendations can also be offered:

1. The Individual Student Report is especially in need of major revisions. Scores need to be represented on an equal-interval scale to minimize confusion and considerably clearer explanations of score meanings are needed. The Panel agrees with the Department of Education that the task of reporting both normative as well as criterion-referenced score information is a

challenge but it is essential that this challenge be met to minimize confusion among users.

2. Appropriate levels of precision should be reported. For example, reporting two decimal places on some reports seems unnecessarily and inappropriately precise and more likely to confuse readers than inform them. If some readers want to check their calculations, then perhaps two reports might be provided. The main report would be focused on clearly presented results.

3. All assessment information which is collected about students should be used in the reporting to enhance both the reliability and validity of the reported information. This is especially important when bands are used, because the bands are likely to be wide, and enhanced score reliability and validity because of the use of the multiple-choice items will reduce the width of the bands.

7.2 Relevant Test Standards

As in the areas of assessment development, reliability, validity, and equating, the Test Standards (AERA, APA, NCME, 1985) also include standards (2.1, 3.16, 4.1, 6.8) which pertain to score reporting and which can be used as a basis for reviewing the current score reports in KIRIS:

Standard 2.1

For each total score, subscore, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in

adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test.

Standard 3.16

The score report forms and instructional materials for a test, including computerized reports and materials, should facilitate appropriate interpretations.

Standard 4.1

Scales used for reporting scores and the rationale for choosing them should be described clearly in test publications to facilitate accurate interpretation of scores by both the test user and the test taker. A publication should specify how scaled scores are derived from raw scores.

Standard 6.8

When test results are released to the news media, those responsible for releasing the results should provide information to help minimize the possibility of the misinterpretation of the test results.

In the remainder of this chapter, the four standards will be applied to various score reports used to transmit KIRIS assessment results in 1993-94 to determine the extent to which current reporting forms meet the requirements of the Test Standards. The list of reports produced in 1993-94, organized by the date of publication, is given below.

September Reports

Summaries of Cognitive Results

Curriculum Reports*

A Summary of Student Questionnaire Information*

Subgroup Analyses*

Item-Level Multiple-Choice Reports

Item-Level Reports for Open-Response Tasks

Student Listings

Individual Student Report*

January Reports

School Accountability Reports*

March Reports

Baseline and Threshold Scores for Second Biennium

Reports marked with a "*" will be commented on below.

7.3 KIRIS 1993-94 Assessment Curriculum Report

This is a 10-page report intended for schools. The report is divided into four parts: (1) details on the cognitive results, (2) student questionnaire results, (3) some relations between questionnaire results and test results, and (4) summary data for the district, state, and nation. This report appears to be well organized, generally clear, and informative. The first page includes a letter from the Commissioner, an identification of the school, and a table of contents. Page 2 contains an informative review of the 1993-94 assessment including subjects covered, how the data were handled, and the exclusionary policy. This material is important and clearly presented. What follows are some specific comments on the reports themselves:

1. Using the symbol "#" for number of students without explanation will be confusing to readers. Also, by reporting the number and percent of students in each achievement category side by side, the general trend in the results is more difficult to determine. The number of students in the analysis is important, to be sure, but it may not need to be presented alongside each percent. That the number of students is sometimes reported as a fraction because of the way the statistics are combined will be a further source of confusion.
2. The breakouts of percent of students in each achievement category by year and subject is informative. Still, some indication of the instability of the percents would be useful. At what point should school personnel interpret the changes as real, and when should the changes be attributed to error or instability in the results? Guidance is needed to avoid overinterpretations of the results. Again, the number of students column is unclear and appears to detract from the readability of the information. Some indication of error seems especially important when the breakouts are subsets of the subject area material (e.g., literary, informational, practical/workplace, and persuasive in Reading, or number, procedures, and structure, space, and dimensionality, and measurement, and change and data in Mathematics).

3. The Part II: Student Questionnaire Results are potentially quite valuable. This is a useful addition to aid in data reporting that is rarely available to administrators and teachers. (NAEP has a similar feature.) Information on questions about time spent on homework, hours spent per week on a part-time job, the use of a calculator on the assessment, grades in school, etc., can be helpful information when interpreting the school results. Perhaps some averages for the state on many of the same questions might help schools better interpret the results. For example, if most students in a school spent less than one hour per day on homework, whereas in the state most students reported spending two hours, the school might want to consider adding more homework especially if their assessment results were below state averages. Again, this is the sort of reporting which is used by NAEP.

4. The subgroup breakdowns (e.g., male, female; White, non-White; Chapter I, non-Chapter I) are potentially useful. Comments made in points 1 and 2 apply here also. The breakout of assessment results by response groups to selected questions on the student survey is especially valuable. For example, the pattern of results for students sorted by the amount of homework they report doing, or the hours spent on a part-time job can be valuable information in interpreting the school results. Breakouts on parents' educational

level seems far less useful since there is little the schools can do with this information.

5. Part IV: District, State, and National Summary Data (see page 10) conveys considerable information. It is hard to know whether the focus should be on comparisons across units (School, District, State, and Nation) or across subjects. The most likely guess is that it is the former and therefore this chart needs to be reorganized for easier reading. Also, the school results need to be added to this chart to permit easier comparisons. Presently, the school must go back to page 3 to pick up its results before comparisons can be made.

In summary, this particular report appears to be quite useful, and, with modest revisions in formatting, could be even more clear and useful to school administrators and teachers.

7.4 KIRIS Biennium I: School Accountability Report

This particular report, in two pages, provides a summary of the accountability results. The title says "School Accountability Report," but it appears to be a report for the district. Some brief comments on this report follow:

1. Would readers find it easier if the columns from left to right were Novice, Apprentice, . .. The current presentation appears backwards. Don't we normally move from low to high?

2. The distributions of students across the four achievement categories are reported for each subject area, for each year, and for four subgroups: common transition, matrix transition, performance events, and combined. No indications of sample sizes are given, no error estimates are given, and percents are reported to two decimal places making the statistics more difficult to interpret and appearing more precise than warranted. It may be, for example, that it is not useful to report the breakout for common and matrix items. These breakouts may represent information overload for users and/or be insufficiently reliable to serve any purposes.

3. Perhaps it is necessary to define terms such as dropout, retention, and successful transition somewhere in the report. Otherwise, their meanings are likely to be unclear to users of this report.

Clearly, there is room for improvement with this important report. The suggestions above should be helpful in reformatting and reorganizing the report.

7.5 Student/Parent Reports

The Individual Student Report is in the form of a letter to the student and his/her parent. The student is placed in a performance category in the four subject areas, his/her approximate percentile is marked in each subject area, and the

percent of students in Kentucky and in the nation in each performance category is also given.

Testing students for eight or more hours is certainly a long enough time to permit more extensive student reporting than was done in 1991-92 and 1992-93. The performance-oriented portion of the assessment may have been only moderately reliable, but, at the same time, it appears that data from the (multiple-choice) transitional tests which can be scored quickly and contain valid assessment information is not being fully utilized in the score reporting. This seems inappropriate.

There is a related concern and that is the matter of reporting scales: a continuous score scale was not used in reporting student performance at Grades 4, 8, and 12. Students are sorted into proficiency levels, and the percentages of students in each level are reported. In 1991-92 and 1992-93, students were informed about their achievement levels (Distinguished, Proficient, Apprentice, Novice). But these levels of achievement are broad, and it seems possible to distinguish "low Novice" from "middle Novice" and "high Novice" for example, if reliability levels are high enough. More informative score reporting at the individual (and perhaps group) level is possible, reliability levels permitting, and we were pleased to see these changes implemented in the 1993-94 student score reports. However, error bands must be provided along with scores to avoid overinterpretations.

The Individual Student Report reviewed by the Panel was dated September 1994 and was printed in the form of a letter to

the student and his/her guardian. The amount and nature of field testing of this (and other) reports is not included in the Technical Manual nor are evaluative results from the field. Our Panel review is as follows:

1. The type-size seemed rather small, making the report unnecessarily difficult to read.
2. The use of a graphic for displaying information seems like a good idea. On the other hand, one of the important pieces of information in this graphic, the student's score, does not stand out. Also, readers are left to determine that there is a score scale being reported, and the arrow marks the location of their score. No numbers at all are used in the score reporting and one wonders whether this unusual way to report student performance will be clear.
3. Expanded descriptions of the four achievement levels would seem to be appropriate. What is the meaning of being at the "Apprentice level" in reading, for example? What can the student do and what is expected at the next level? It would be useful to convey such information to students and their parents.
4. About the only interpretations that are offered are normative and these seem less than clear. For example, the Grade 12 student at the Apprentice level in reading is told, "55% of Kentucky students scored at this level; another 29% scored below this level. Nationally, an estimated 49% scored at the Apprentice

level; another 37% scored below this level." The reader is left to put many things together including sorting out normative information for two samples of students.

5. An example of the sort of confusion parents and students will encounter occurs when the student scores at the Proficient level. The report contains this statement, "Congratulations! You have achieved the goal established by the Kentucky Legislature, a goal achieved in 93-94 by only 19% of Kentucky students and an estimated 17% of students nationwide." The confusion occurs because when the student looks at the graphic, it is clear that only 13% were Proficient. To obtain the 19% figure, a student must combine the 13% at the proficient level and the 6% at the Distinguished level.

6. A student's approximate score in each subject area is marked. Unfortunately, there is no indication of the amount of error in each score. Such information is highly desirable in an effective score reporting system.

7. The scale itself will be confusing. It has ordinal scale properties but is likely to be interpreted by students' parents as if it were an equal-interval scale. With the current reporting scale, students are being poorly informed about the closeness of their performance to the cutoff scores separating the

proficiency levels. This is the sort of information which will be of interest to users.

In summary, there appears to be considerable room for improvement in the reporting of assessment results to parents and students. Criterion-referenced information should be added; the displays of data can be improved; the normative presentation of data might be clarified (possibly presenting a percentile score for the student in each norm group); and an indication of the error in the student's score is needed. Finally, an equal-interval scale should be used in displaying student performance because of the likely interest in the relationship between a student's level of performance and the cutoff scores separating the four proficiency categories.

7.6 Briefing Packet - 1993-94 KIRIS Assessment Results

This packet is intended for the media (print and television). It contains many useful documents including a press release, description of Kentucky's assessment program, description of the performance levels, glossary of assessment terms, packet of sample items, and various charts and graphs highlighting the KIRIS results. A few comments follow:

1. The press release describes the apparent successes in Kentucky. The expression "Substantial and Dramatic Growth" is used. There is no attention in the press release to possible shortcomings in the assessment system or to irregularities, or to doubts or concerns about the reliability and validity of the assessment

results. In fact, at no point in any of the materials are cautions extended as to instability in the statistics or the difficulties in assessing what it is that students are learning, even with performance assessments, for detecting school growth. The only message conveyed to the press is that the educational system is considerably improved.

2. The descriptions of the achievement levels for the media are very sparse. For example, the definition offered for Novices is "The student is beginning to show an understanding of new information or skills." For Apprentices, the definition offered is, "The student has gained more understanding and can do some important parts of the task." These descriptions convey very little information about the cognitive skills of students in these achievement categories. If the achievement levels are to be the prime focus of score reporting, then these categories need to be described in considerable detail, perhaps with clear examples of how performance differs for students placed in the four categories.

3. The glossary is a useful idea but words such as error, reliability, standard error of measurement, consequential validity, construct validity, content validity, multiple-choice item, etc., need to be added.

4. The inclusion of sample assessment material in the package is a good feature.

5. The reporting of the statewide performance summary is flawed to the extent that readers will not understand the accountability index scale (or specifically, the cognitive index scale). There are many numbers such as baseline, threshold, and average cognitive index which will not have meaning to the intended audience. More descriptive information on the reporting scale is needed to conform with Standard 4.1. Also, there is no indication that there are errors associated with these statistics (albeit small because of the large sample sizes). The breakdowns by subject seem equally flawed because of the failure to fully describe the reporting scale. The footnote at the bottom of several tables seems insufficient.
6. The graphical displays of cognitive index scale scores by subject area are useful. For one, they highlight the results over three years and place on the same graph, the desired goal. There is the possibility of persons interpreting the scale as if it were a percent score scale. On the other hand, the footnote describing the cognitive index scale seems quite clear.

The reports for the media have value but they do not appear to present a balanced or clear presentation of the results. With the exception of a few graphs which point to the discrepancy between the current status and targets, no attention is drawn to the gap between what is and the goal, and no attention is given to alternative explanations of the findings. All findings are

described as if they were real and without error, dispute, and alternative interpretations. The suggestions above should be helpful for improving the media packages.

7.7 1993-94 Interpretive Guide

This report is circulated to all schools at the same time as the score reports. It is intended to help schools make sense of their results. Given that this is the third edition of the report (one is produced for each release of results), it is disappointing to see that it is still very brief with few details about the score reports themselves. A few comments follow:

1. The Individual Student Report, the Student Listing, the Item Level Reports, and the KIRIS Assessment Curriculum Report are all briefly described and certain cautions are extended (e.g., performance levels are based solely on students' responses to the common open-response questions). For example, this would be an excellent report in which to address the issue of measurement error and unreliability, and to explain the score scale. NCEs are introduced without explanation. Percentile rank scores are also mentioned but not defined.
2. In view of the complexities of the assessment and accountability system, and the number of reports, and the various scales and scores which are used, this report would better serve KIRIS if it was rewritten to focus on each report, with a detailed explanation of

each piece of information that appears. Even popular questions which have been asked previously might be addressed. For example, what are NCE scores and why are they used? Why do these NCE scores contain the multiple-choice items and, at the same time, the multiple-choice items are not included in the test battery scores used in compiling percentile ranks? How reliable are the total battery scores that are reported? Why are the score categories in Appendix A overlapping? (There is an explanation offered for this last question but it seems unclear.)

3. In view of the likelihood of overinterpretation of results because of the failure to address the concept of error anywhere in the reports, this Interpretive Guide would be an excellent place to introduce the concept of error.

In summary, the Interpretive Guide appears to be an especially important document. However, considerable expansion is needed for this document to contribute in a positive way to KIRIS score interpretations.

7.8 Celebrate the Progress - News Packet

This is a second news packet that we looked at. This one contained the 1992-94 Kentucky Accountability Results. Nearly all of the comments offered in Section 7.6 apply to this packet as well. In this section, we will comment on some additional points which arose with the new reporting formats:

1. The display of statewide accountability scores seems clear. Labelling is appropriate and the graph does not distort the amount of growth which took place relative to the ultimate goal.
2. The reporting of state averages by total (i.e., across grades), grade, and region is clear, but the reporting does not display the substantial variability which exists in baselines, growth gains, etc., across schools. The variability across schools is valuable information which is not conveyed by a detailed reporting of means.
3. Graphics were used to display the percentage of schools in each rewards and assistance category. These graphs are confusing because a reader must look at the shading of a portion of the circle, then try to match up the shading to codes and descriptions presented at the bottom of the graph, and then go back to the graph and read the percentage. The graphics would be clearer if the shading was dropped, the percentages were placed in the circles, and brief descriptions with arrows were used to highlight each portion of the circle (e.g., eligible for rewards). Then, instead of trying to sort out the codes, shading, and percentages, a reader's focus could be solely on the percentages in each reward category.
4. Evidence is offered to the press on score reliability, quality of scoring, and validity. All of this

information is offered without qualifications or concerns. This becomes problematic when correlations between NAEP and KIRIS are reported as high, but then NAEP fails to show the same level of growth as the KIRIS results.

In summary, some of the news packet material appears useful, but improved graphics and more balanced presentations of KIRIS strengths and weaknesses would be desirable when reporting information to the media.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Hambleton, R. K., & Slater, S. C. (1994, October). Improving the communication of NAEP results to policymakers and educators. Paper presented at the NCES-NAGB Standards-Based Reporting Conference, Washington, DC.

Kentucky Department of Education. (1995, February). Celebrate the progress: 1992-94 Kentucky accountability results summary--news packet. Frankfort, KY: Author.

CHAPTER 8

Is There Evidence That Education Has Improved in Kentucky?

8.1 Central Questions and Principal Findings

In Celebrate the Progress (Kentucky Department of Education, 1995a), the Kentucky Department of Education reported large statewide gains in KIRIS scores between the baseline year (1991-92) and the 1993-94 school year. For example, the Kentucky Department of Education reports that the statewide accountability index across all tested grades was 36.4 in the baseline year, on the basis of which the Kentucky Department of Education set an improvement goal of 42.9 for 1993-94. Statewide, Kentucky schools exceeded this goal in 1993-94, reaching an index of 44.8. The index used in the accountability system was somewhat lower (42.1), because that index required averaging the 1992-93 and 1993-94 scores, and much of the improvement (particularly in grades 8 and 12) occurred between those years (see Table 8.1). Nonetheless, even this two-year average showed sizable gains.

Table 8.1 Statewide Average KIRIS Scores[1]

	1991-92	1992-93	1993-94	Raw Change[2]
Grade 4	33.9	39.3	43.7	9.8
Grade 8	37.4	39.9	44.9	7.5
Grade 12	39.3	39.4	46.4	7.1

[1]Change score is difference between 1991-92 and 1993-94, for comparability to results reported below. In the source table, Kentucky Department of Education reported changes between 1991-92 and the average of 1992-93 and 1993-94.

[2]Celebrate the Progress, (Kentucky Department of Education, 1995a, p. 3).

But to what extent do these gains in KIRIS scores represent real improvements in student learning? The public is (or should be) little interested in test scores per se. Rather, test scores have value when they indicate mastery of some important domain of knowledge or skills, or when they indicate the development of skills or capacities that have value in the student's later experience. Unfortunately, test-score gains need not represent commensurate improvements in learning. Particularly in the case of high-stakes testing programs, which provide strong incentives to raise scores, gains in scores can exaggerate actual improvements in student achievement.

The Panel believes that 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. It is important to note, however, that this external evidence is limited; it includes only a limited number of grades and subject areas, and it does not include direct assessments of writing, which is an area in which a positive effect of KIRIS might seem particularly likely. The Panel is also unable to determine which of many factors might have caused exaggerated gains on KIRIS.

Although these findings are not comprehensive, they strongly suggest that a large share of the KIRIS gains in some subjects and grades – fourth-grade reading and twelfth-grade reading and mathematics – are substantially exaggerated. That is, in these

cases, the gains in KIRIS scores appear to have been far larger than actual gains in achievement.

Of the data discussed here, the NAEP data may be the most important. The severity of the disparity between NAEP and KIRIS is remarkable. It might be reasonable to expect real gains in KIRIS to be reflected only in smaller gains on NAEP, because the tests are somewhat different. However, it seems hard to explain massive gains on KIRIS, coupled with essentially no change on NAEP, solely on the basis of differences between the tests. Both are tests of reading, and KIRIS was in fact built in part to reflect the NAEP framework. As the 1991-92 KIRIS Technical Report noted, "Kentucky's reading framework [is] similar to that of the 1992 NAEP framework" (Kentucky Department of Education, 1993, p. 14). Indeed, some of the items used in KIRIS in the baseline year were actually taken directly from NAEP, with or without modification. (Lisa Ehrlich, personal communication to Neal Kingston, April 30, 1995; forwarded to the Panel by Neal Kingston on May 3, 1995. The Panel does not have information indicating the percentage of the KIRIS reading items derived from NAEP.)

The ACT comparison may be less important than the NAEP comparison, because the ACT framework was not considered in constructing KIRIS, and the discrepancy between ACT and KIRIS is smaller than the corresponding discrepancy between NAEP and KIRIS. It might therefore be unrealistic to expect changes in ACT scores to parallel as closely as NAEP changes in KIRIS scores. Nonetheless, the difference between ACT and KIRIS is

still striking, and the Panel does not consider it reasonable to maintain that the tests are so different that there should be no reflection of KIRIS gains at all in ACT scores of the same students. Moreover, the NAEP results above suggest that far more is at play than differences between the ACT and KIRIS assessments. Particularly in the light of the NAEP/KIRIS comparisons above, these results call into question the extent to which KIRIS gains represent real improvements in student achievement.

The Panel believes that taken together, the ACT and NAEP findings are sufficient to suggest that gains in KIRIS scores are substantially inflated and provide the public with a misleading view of improvements in student performance. (Although the CTBS/CAT comparisons noted in Section 8.5 are also consistent with this conclusion, the Panel considers the limitations of those comparisons to be serious and reached its conclusion about likely inflation of KIRIS scores without consideration of those data.) The available data do not indicate, however, what role specific factors such as teaching to the test, motivational changes specific to KIRIS, or inaccurately low baseline scores might have played in producing the discrepancies between KIRIS and NAEP and ACT.

8.2 Evaluating Gains in KIRIS Scores

The measurement field has recognized for some time that gains in test scores can be inflated, as compared to improvements in the aspects of student learning they are designed to measure,

and that the risk of inflated scores is increased when high stakes are instituted for gains in scores. Inflation of scores can arise for a number of reasons, a few of which are noted below.

Thus, in a program such as KIRIS, one of the most essential tasks for evaluation is gauging the extent to which observed gains in scores indicate improvements in student learning. Among the strongest evidence is students' performance on test questions that measure related aspects of learning but for which coaching has been less intense. The high-stakes assessment itself could be built to include comparisons of this sort; in addition, one can use external comparisons (from other testing programs) to the extent that such data are available and appropriate.

One obvious external comparison is the National Assessment of Educational Progress (NAEP), to which KIRIS is by statute supposed to be linked. Through its Trial State Assessment (TSA), NAEP offers performance information that is representative at the state level for a limited number of subjects and grades -- in 1994, only fourth-grade reading. The comparison between KIRIS and NAEP illustrates the mechanisms that can lead to divergent results between tests.

Before 1994 NAEP results were available, the Kentucky Department of Education offered the following potential explanations if KIRIS gains turned out not to be reflected in NAEP or other nationally normative data:

1. Increased student learning to which NAEP is not sensitive because it may not measure higher-order skills or ability to present information as well as KIRIS;
2. Increased motivation to do well on KIRIS tests in recent years compared to 1991-92 (Kentucky Department of Education, Celebrate the Progress, 1995a, p. 23).

The first of these explanations would imply that the KIRIS gains are real but occur largely in domains not assessed by NAEP. The second, the motivational argument, on the other hand, would imply that the gains on KIRIS are at least in part misleading. That is, gains attributable to motivational changes specific to the test would reduce the validity of KIRIS gains as a measure of improved achievement.

There are at least two other explanations of KIRIS gains that are not reflected in scores on tests such as NAEP, and both would imply that KIRIS gains are exaggerated. First, increases in KIRIS scores not reflected in NAEP could also represent inflation of test scores from undesirable forms of teaching to the test -- specifically, methods of teaching or coaching that are so closely aligned with the test that they yield gains in scores on that specific test without commensurate gains in real achievement that would be apparent in other, related tests. Other research has shown that inflation of scores in test-based accountability systems can be very large (Koretz, Linn, Dunbar, & Shepard, 1991), and innovative performance assessments such as KIRIS may be as vulnerable as traditional tests to inflated scores. (Indeed, some measurement experts believe that because

of the small number of tasks used in many performance assessments and the weak correlations among them, many performance assessments may be more vulnerable than many traditional tests to inflation of scores.) Second, KIRIS gains could have been inflated if some schools aimed for low scores in the baseline year. (Low scores in the baseline year carried no penalty and would have made increases in subsequent years easier to obtain.) The Panel is aware of anecdotes about this but does not have systematic evidence indicating whether it in fact occurred frequently enough to have had an appreciable impact on gains.

Unfortunately, the KIRIS system was not constructed to include a clear method for auditing the meaningfulness of gains in KIRIS scores. Therefore, the Panel must turn to external comparisons. Apart from NAEP results in fourth-grade reading from the 1992 and 1994 TSAs, there is one statewide comparison available: scores on the American College Testing (ACT) college admissions tests (for those Kentucky high school students who took both ACT and KIRIS). Both the NAEP and ACT comparisons suggest that KIRIS score gains in the subjects that they examine -- reading in grades 4 and 12 and mathematics in grade 12 -- are badly inflated and that true improvements in student achievement in those three cases are either negligible or much smaller than KIRIS would suggest. However, it is important to stress that these comparisons are confined only to those three subject-by-grade combinations and suffer from important limitations. In addition, data are available from a small number of Kentucky school districts that have administered the Comprehensive Tests

of Basic Skills (CTBS) and the California Achievement Test (CAT). The Panel considers these data to be much less valuable, for reasons noted below, but it is worth noting that they too failed to show a consistent pattern of gains reflecting KIRIS.

8.3 NAEP Trial State Assessment

In 1994, NAEP TSA was administered only in a single subject (reading) and a single grade (fourth). The limited scope of the 1994 TSA was the result of lower-than-anticipated federal appropriations for NAEP. State-level results of the 1994 TSA were released this spring. The TSA results here are taken from Williams, et al. (1995) and reflect only public schools.

KIRIS reports very large gains in fourth grade reading. (It should be noted that the gains in fourth-grade reading were among the largest reported for KIRIS.) The gains were striking in terms of both average scores and the percentages reaching each of the KIRIS standards. In contrast, NAEP showed essentially no change.

First, the statewide average fourth grade KIRIS reading score roughly doubled between 1992 and 1994 (see Table 8.2). The KIRIS gain corresponds to about three-fourths of a standard deviation, which is a very large impact compared to those commonly found in large-scale educational interventions. (Approximately 69 percent of the 1994 fourth-graders scored above the 1992 average score.) In contrast, the Kentucky statewide fourth-grade average on NAEP remained virtually unchanged. (The

1-point decline on NAEP was both substantively trivial and statistically unreliable.)

Table 8.2 KIRIS and Kentucky NAEP Fourth-Grade Average Reading Scores, 1992 and 1994[1]

Assessment	1991-92	1993-94	Raw change	Standardized Change[2]
KIRIS	21.0	39.8	18.8	0.76
NAEP	214	213	-1	-0.03

[1]KIRIS scores are from the KIRIS Biennium I Technical Manual, March 1 draft, 1995. Panel tabulations showed a slightly smaller gain in KIRIS scores, but the discussion here would not be materially affected by the difference. NAEP scores are from Williams, et al. (1995).

[2]The standardized change expresses the change as a fraction of a standard deviation and allows comparison of changes on tests with different scales. The KIRIS standardized change was based on the 1994 reading standard deviation, calculated by the panel, and probably would have been substantially larger if the 1992 standard deviation had been used. The NAEP standardized change was based on the national standard deviation from 1992 because neither the 1992 national nor the Kentucky standard deviation was available; it would have been slightly larger if the Kentucky standard deviation had been used.

The change in KIRIS was particularly striking in terms of the percentage of students reaching the lowest of the three standards, Apprentice. In 1992, the KIRIS Novice and NAEP Basic standards appeared to represent roughly comparable levels of difficulty. About half (51 percent) of Kentucky students failed to reach the Apprentice standard and were therefore labeled "Novice." Similarly, 45 percent of Kentucky fourth-graders failed to reach NAEP's Basic level (see Table 8.3). By 1994, however, the percentage scored as Novice on KIRIS fell by over half, from 51 to 22 percent (Table 8.3). In contrast, the percentage of Kentucky fourth-graders scoring Below Basic on NAEP

remained essentially unchanged (increasing by a trivial and unreliable 2 percentage points). These changes are necessarily in another commonly reported statistic, the percentage of

Table 8.3 Percent of Fourth-Grade Students Below Lowest Standard in Reading, KIRIS and Kentucky NAEP, 1992 and 1994[1]

Assessment	1991-92	1993-94	Raw change
KIRIS: Novice	51	22	-29
NAEP: Below Basic	45	47	+2

[1]KIRIS scores are from the KIRIS Biennium I Technical Manual, March 1 draft, 1995. NAEP scores are from Williams, et al. (1995).

students reaching or exceeding the lowest standard (Basic or Apprentice, Table 8.4).

Table 8.4 Percent of Fourth-Grade Students At or Above Lowest Standard in Reading, KIRIS and Kentucky NAEP, 1992 and 1994[1]

Assessment	1991-92	1993-94	Raw change
KIRIS: Apprentice or above	48	79	+31
NAEP: basic or above	55	53	-2

[1]KIRIS scores are from the KIRIS Biennium I Technical Manual, March 1 draft, 1995. NAEP scores are from Williams, et al. (1995).

The disparity between KIRIS and NAEP trends was apparent at the high end of the distribution as well. KIRIS showed the percentage scoring at or above Proficient (i.e., Proficient or Distinguished) increasing roughly fourfold, from 3 to 13 percent (see Table 8.5). NAEP, by contrast, showed only a small increase scoring at or above Proficient (from 19 to 22 percent), and this increase is mostly due to a small and statistically unreliable

increase in the percentage scoring at the Advanced level (see Table 8.6).

Table 8.5 Percent of Fourth-Grade Students At or Above Middle Standard in Reading, KIRIS and Kentucky NAEP, 1992 and 1994[1]

Assessment	1991-92	1993-94	Raw change
KIRIS: Proficient or above	3	13	+10
NAEP: Proficient or above	19	22	+3

[1]KIRIS scores are from the KIRIS Biennium I Technical Manual, March 1 draft, 1995. NAEP scores are from Williams, et al. (1995).

Table 8.6 Percent of Fourth-Grade Students At or Above High Standard in Reading, KIRIS and Kentucky NAEP, 1992 and 1994[1]

Assessment	1991-92	1993-94	Raw change
KIRIS: Distinguished	0	1	+1
NAEP: Advanced	2	4	+2

[1]KIRIS scores are from the KIRIS Biennium I Technical Manual, March 1 draft, 1995. NAEP scores are from Williams, et al. (1995).

8.4 ACT Scores for Students Taking ACT and KIRIS

ACT and KIRIS scores are available for a large number of students who took both tests: over 18,000 students in 1992 about 20,600 students in 1994. Trends in ACT scores provide another basis for assessing the extent to which KIRIS gains represent real gains in achievement.

The Panel compared ACT changes to KIRIS changes in both reading and mathematics. (The Panel expresses its gratitude to Kentucky Department of Education for providing the merged ACT/KIRIS data used here and to American College Testing for

carrying out the merge.) One cannot directly compare ACT scores to statewide KIRIS scores because the ACT test-takers are a self-selected group that is not representative of the state as a whole and may change over time. Therefore, we compared changes in ACT scores to changes in KIRIS scores *only for students who took both tests*. Our estimates of KIRIS scores reflect only common items on the transitional tests because the scaled responses to the matrix items are unavailable on that data set. However, the comparisons below to scores reported in the KIRIS Biennium I Technical Manual (Kentucky Department of Education, 1995b, March 1 draft) suggest that excluding matrix items should have had relatively little effect on mean scores.

In grade 12 reading, KIRIS scores showed a fairly large improvement, while ACT scores showed essentially no change. The KIRIS scores of students who took both the ACT and KIRIS increased by 9.2 points between 1992 and 1994 (see Table 8.7), or about 28 percent of a standard deviation. This increase, while much smaller than the increase reported above for fourth-graders, is nonetheless sizable for a short-term, large-scale intervention. (This change is only slightly smaller than the 10.4 point gain reported for all twelfth graders based on all items other than performance events [Biennium I Technical Manual, March 1 draft].) In contrast, scores of those same students on the ACT remained virtually unchanged between 1992 and 1994 (see Table 8.7).

Table 8.7 KIRIS and ACT Average Reading Scores, 1992 and 1994, only for Students Who Took Both Tests

Assessment	1991-92	1993-94	Raw change	Standardized Change[2]
KIRIS	40.3	49.5	+9.2	+0.26
ACT	20.5	20.6	+0.11	+0.02

[1]Panel tabulations of merged ACT/KIRIS database.

[2]Change scores may include rounding error. The standardized change expresses the change as a fraction of a standard deviation. Both ACT and KIRIS scores were standardized using the standard deviations of the sample who took both tests in 1994.

The discrepancy between KIRIS and ACT is even more striking in the case of mathematics. In mathematics, the KIRIS scores of students who took both the ACT and KIRIS increased by 15 points between 1992 and 1994 (see Table 8.8), or about 36 percent of a standard deviation. (This change is a little larger than the 12.6 point gain reported for all twelfth graders based on all items other than performance events [KIRIS Biennium I Technical Manual, 1995, March 1 draft].) Here again, the mathematics ACT scores of those students showed essentially no change.

Table 8.8 KIRIS and ACT Average Mathematics Scores, 1992 and 1994, only for Students Who Took Both Tests

Assessment	1991-92	1993-94	Raw change	Standardized Change[2]
KIRIS	38.2	53.2	+14.9	+0.36
ACT	19.1	19.0	-0.07	-0.02

[1]Panel tabulations of merged ACT/KIRIS database.

[2]Change scores may include rounding error. The standardized change expresses the change as a fraction of a standard deviation. Both ACT and KIRIS scores were standardized using the standard deviations of the sample who took both tests in 1994.

To explore these findings further, the Panel examined the school-level correlations between mean ACT scores and mean KIRIS scores, considering only students who took both tests and only schools in which 10 or more students took the ACT. That is, we examined the degree to which a high average score on KIRIS predicted that a school would have a high average score on the corresponding ACT subtest.

In 1992 and 1993, schools with high average KIRIS scores in mathematics tended to have high average ACT scores in mathematics; the correlations were .68 and .75, respectively (Table 8.9). (Higher correlations between the two tests would generally be seen as supporting their validity as measures of overlapping domains of mathematics.) In 1994, however, when KIRIS mathematics scores increased markedly, the correlation between average ACT and KIRIS scores dropped to .53. The same pattern appears in reading, although the correlations in reading were in every year lower than those in mathematics. In 1992 and 1993, the correlations between average KIRIS and ACT reading scores were about .60; in 1994, when KIRIS scores increased markedly, the correlation with ACT scores dropped to less than .40.

In other words, as KIRIS scores increased, their consistency with ACT scores dropped. This indicates that the increase in KIRIS scores was reflected in ACT scores less in some schools than in others. This pattern could have a number of causes; for example, changes in the characteristics of the sample taking the

Table 8.9 School-Level Correlations Between Mean KIRIS and Mean ACT Scores, 1992, 1993, and 1994, only for Students Who Took Both Tests

Subject	1991-92	1992-93	1993-94
Reading	.59	.59	.38
Mathematics	.68	.75	.53

[1]Panel tabulations of merged ACT/KIRIS database. N=222 schools in which 10 or more students took both tests.

ACT might have contributed. However, one possible explanation is that a sizable share of the gains on KIRIS resulted from test preparation activities in some schools that did not produce gains in achievement that generalized to the ACT.

8.5 CTBS/4 and CAT 5 Scores From District Testing Programs

The Office of Education Accountability obtained standardized test scores from eight districts that administered the Comprehensive Tests of Basic Skills (CTBS/4) in 1989-90 and the California Achievement Test (CAT 5) in 1994-95. OEA staff, with assistance from Kentucky staff of CTB/McGraw Hill (the publisher of both tests) used equating tables from national norming samples to convert the 1990 CTBS scores to the norms of the CAT 5.

OEA tabulations of these equated data do not show a consistent pattern of gains. In grade 5, for example, five districts provided data, and the change in their total battery scores ranged from +4.0 to -6.3 NCEs (normal curve equivalents). In grade 5, the (unweighted) average total battery change among the 5 districts providing data was essentially zero (-0.2 NCE).

In grade 7, the average change was -4.1 NCEs, and in grade 10, -1.0 NCEs.

These data are in broad outline consistent with the NAEP and ACT data noted above, but in the Panel's view, they should be given little weight in evaluating KIRIS gains. The Panel does not have KIRIS scores separately for the students represented in these data, and the districts providing data may not have shown increases comparable to the state as a whole. Moreover, the lack of consistent gains on these tests may in part reflect a de-emphasis on them – including perhaps a decrease in coaching – as a result of the introduction of KIRIS as a higher-stakes assessment. The equating of CTBS and CAT may also add error to them.

References

Kentucky Department of Education. (1993). KIRIS 1991-92 technical report. Frankfort, KY: author.

Kentucky Department of Education. (1995a, February). Celebrate the progress: 1992-94 Kentucky accountability results summary news packet. Frankfort, KY: Author.

Kentucky Department of Education. (1995b, March 1). KIRIS biennium I technical manual. Frankfort, KY: Author.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). The effects of high-stakes testing: preliminary evidence about generalization across conventional tests. A symposium presented at the meetings of AERA and NCME, Chicago.

Williams, P. L, Reese, C. M., Campbell, J. R., Mazzeo, J., & Phillips, G. W. (1995). 1994 NAEP reading: a first look. Washington, DC: National Center for Education Statistics.

CHAPTER 9

Conclusions and Recommendations

9.1 Purposes of the Review and General Conclusion

In September of 1994 a national panel of six measurement specialists was appointed by the Office of Education Accountability of the Kentucky General Assembly to address the following questions:

> Is the measurement quality of KIRIS sufficient to support the intended uses of the KIRIS results and the actions taken by the Kentucky Department of Education and the legislature? And, to the extent that shortcomings in KIRIS are identified, what changes would need to be made to improve the accountability and assessment system?

After reviewing large numbers of curriculum and technical documents and assessments, conducting a number of relevant analyses, and carefully considering our findings, the Panel is in unanimous agreement that KIRIS is seriously flawed and needs to be substantially revised. The Panel is not suggesting that the educational reform movement in Kentucky is a failure or that educators are not working hard to implement the goals of KERA. We are not suggesting either that the educational reforms taking place in the areas of curriculum design and instruction and teacher in-service training are not worthwhile. The Panel is saying that the accountability and assessment system has major flaws which need to be corrected as Kentucky moves into the second accountability cycle. This broad conclusion does not arise from any single piece of evidence but rather is based on the Panel's review and analysis of (1) assessment development procedures, (2) the accountability index and its reliability, (3) portfolio scoring, (4) the linking or equating of assessments

from one year to the next, (5) the procedures used to set performance standards, and (6) evidence addressing the impact of KIRIS on student learning. Appropriate corrections or modifications along the lines of the recommendations below should allow KERA to come closer to achieving its full potential.

9.2 Major Findings

As KIRIS is currently designed and implemented, it is flawed to the extent that, in the Panel's judgment, the present form of KIRIS cannot support its accountability and assessment goals and objectives. Also, as a result of the serious measurement flaws, (1) the public is being misinformed about the extent to which student achievement has improved statewide; and (2) the public is being misled by being given information about the accomplishments of individual students that may be inaccurate.

Our major findings which support the main conclusion above are as follows:

1. The misclassification rates of schools in some reward categories are high and therefore the rewards and sanctions may be difficult to defend.
2. Although limited, evidence from other assessments (National Assessment of Educational Progress and the American College Tests) fails to show any reflection of the large gains observed on KIRIS. This suggests that at least in the grades and subject areas to which these data pertain, KIRIS markedly overstates actual gains in student achievement. The Panel did not have enough

information to ascertain the causes of the exaggerated gains on KIRIS, but many factors, such as teaching students to be better test-takers, faulty baseline scores, or breaches of assessment security, could have contributed.

3. Repeated use of ad-hoc, judgmental procedures for linking or equating assessments in 1992, 1993, and 1994 result in an accumulation of equating errors which make year to year comparisons of KIRIS results of questionable validity. Changes in equating methods over the years further reduce the validity of the equating results and challenge the desired interpretations of KIRIS results.

4. The setting of performance standards was seriously flawed -- for example, by the decision to set standards only on the basis of three items per subject and grade. In addition, critical aspects of the standard-setting process were not well enough documented to determine their adequacy. Therefore, the Panel considers the performance standards used to classify students as Novice, Apprentice, Proficient, and Advanced to be untrustworthy. This makes the information provided to parents and others about student performance potentially misleading. It also compounds the errors in the classification of schools that result from other factors, such as sampling and the limited number of test items.

5. For numerous reasons, the Panel concluded that the KIRIS portfolio assessments are currently inappropriate for use in the KIRIS accountability system. Despite improvements in rater consistency over the first biennium, the scoring of portfolios remains too flawed for use in a high-stakes system. In particular, scores provided by teachers in students' own schools -- the scores used in the accountability system -- remain biased upward by a substantial and variable amount. Numerous aspects of the operation of the program cast doubt on the validity of portfolio scores, and the evidence pertaining to validity is both very limited and unpersuasive. On the other hand, the Panel remains guardedly optimistic about the potential instructional effects of the portfolio program and would endorse its continuation in a lower-stakes form.

Several other findings which support our main conclusion but which are of lesser importance than the five major findings above are as follows:

1. There is inadequate documentation of the assessment development process (e.g. assessment review process), equating, and standard-setting. This inadequacy of documentation undermines the trustworthiness of the key conclusions the assessment is intended to support because it makes it impossible to evaluate numerous threats to validity.

2. Although the Panel recognizes that the Department has terminated the multiple-choice component of KIRIS primarily for the presumed effects on instruction rather than for measurement reasons, the negative ramifications of this decision for the quality of measurement are numerous and severe. The Panel strongly recommends re-incorporating multiple-choice items into the assessment. The elimination of multiple-choice items by the Department of Education from all of the important analyses unnecessarily restricts content coverage, lowers the reliability of the school accountability index, reduces the stability of the equating or linking of assessments from one year to the next, reduces the stability of the performance standards, and creates less reliable and valid scores for school and individual score reporting.

3. Collaboration between students and others (other students, teachers, and others) is substantial in KIRIS and poses a potentially serious threat to the validity of both student-level and school-level results. Although formally identified group products play a very limited role in the KIRIS assessment, collaboration between students and others is substantial in both the portfolios and the performance events. The difficulties inherent in collaborative assessment are clear but the Panel is not aware of any evidence having been collected to support the validity of KIRIS's use

of collaborative work. The Panel considers it essential that evidence of validity be presented for any collaborative work included in KIRIS for high-stakes purposes.

Other conclusions based upon our Panel's review are important to highlight because they are encouraging for the future of KIRIS:

1. The Panel applauds the Department of Education for encouraging research and evaluation pertaining to KIRIS and for permitting and facilitating independent research by third parties unconnected to the Department or its contractor. This research will be essential for improving KIRIS, and the Panel strongly recommends that the Department continue its support of both internal and external research efforts.
2. The Panel's impression is that the Department of Education is willing to improve KIRIS and there is evidence that improvements have been made over the last four years when problems have been identified (e.g. improvements in portfolio scoring, and assessment development). Even some of the preliminary suggestions from our Panel have already been accepted and incorporated into KIRIS.
3. The Department of Education and Advanced Systems have demonstrated the capability of developing essential material and training documents in the areas of KIRIS

implementation, curriculum development, assessment development, and portfolio implementation and scoring.

4. The Department of Education has recognized the central role of teacher training and has devoted significant effort to the delivery of essential teacher training.

9.3 Recommendations

The Panel's major recommendation is that KIRIS needs to be restructured and redesigned if it is to accomplish its objectives. The Panel's main recommendations follow below:

1. The portfolios should not be used at this time in the accountability index.
2. Any assessment data used for accountability purposes should be scored externally to the schools in which the data are collected.
3. An ethical code of conduct for assessment activities should be developed and implemented. The Department of Education should also develop an in-house capability for auditing school procedures and sanctioning confirmed breaches of assessment security.
4. The amount of validation work on the assessments should be expanded. Additional construct validation evidence is needed to support the various uses and interpretations of the performance assessment data.
5. The design for equating assessments should be strengthened and the ad hoc procedures eliminated. The Department should use a scientifically sound and

rigorous approach to assessment equating because this activity is absolutely critical to the integrity of the total KIRIS system.

6. Performance standards should be re-established and full documentation of the process should be provided. The Department is strongly advised to avoid the use of ad hoc, unjustified statistical linkages in establishing standards.

7. The various score report forms should be field tested prior to their use to insure that communication of important information is clear and understandable. Also, effective communication of results must include caveats about the results (e.g., alternative interpretations and cautions) and indications of the measurement error in the data to reduce the likelihood of overinterpretation and misinterpretation.

8. Material provided to the press and the public should provide alternative interpretations of the results when such alternative interpretations are plausible. Results released in the media should contain important caveats such as the lack of generalizability of findings to new sets of tasks.

9. There is a great need to establish routine auditing procedures on all aspects of KIRIS including assessment development, standard-setting, equating, etc. Because of the high-stakes nature of KIRIS and the resulting potential for inflated gains in scores, it is essential

that mechanisms be established for ongoing auditing of observed gains on KIRIS.

10. Item formats should be used which contribute to the validity of the educational assessments. This means that multiple-choice items should have a role to play and will be valuable in enhancing content validity, the reliability of school and student scores, score equating, and score reporting.

11. The documentation in technical areas of equating, standard-setting, and score reporting, needs to be substantially improved to facilitate review and replication.

12. There has been a shift toward process at the expense of content in the curricula and this shift needs to be reconsidered. Our Panel does not have a view that the current situation is wrong. We simply feel that this situation needs to be reviewed to be sure that the impact on instruction, while presumed by the Department of Education to be positive, is, in fact, positive. In addition, the implications of this shift away from content for the adequacy of measurement--for example, for the accuracy of the estimates of change upon which KIRIS focuses--should be more fully evaluated.

In summary, the Panel feels that the Kentucky Department of Education set unrealistic expectations for itself and its contractor to deliver an accountability and assessment system

that would meet all of the essential Test Standards within a short period of several years. The result is a system which has many technical shortcomings and, therefore, is not meeting the accountability and assessment needs of the Commonwealth as represented in KIRIS. Progress has been made in the last four years and the Department, its contractor, and everyone who has worked on KIRIS and its implementation deserve credit for the efforts and accomplishments which have been made. On the other hand, considerably more progress is needed to establish KIRIS as a technically sound accountability and assessment system. The Panel hopes that our conclusions and recommendations will be valuable to the Department of Education as it sets out on the second accountability cycle.

APPENDIX A

LEGAL ISSUES

Preface

A statewide assessment and accountability program, such as KIRIS, can be reviewed and evaluated usefully from a number of perspectives. Since the focus of any review, regardless of perspective, is the assessment and accountability program itself, measurement issues are unavoidable and some degree of redundancy is inevitable in all reviews.

In this Appendix, the KIRIS program has been reviewed from a legal perspective, in contrast with the technical measurement perspective assumed in the body of this report. The review in this Appendix can be distinguished from the review in the body of the report in several ways. First, it focuses principally on measurement issues that have been at the heart of legal challenges to testing, assessment, and accountability programs. Second, the arguments it advances are grounded principally in the law, a perspective that is absent from the body of this report. Third, the rules of evidence that undergird the conclusions reached in this Appendix draw upon the law more than on the traditions of educational measurement despite the inevitable measurement content of the issues raised. It is the primacy of focus and the structure of argument and evidence that serve best to distinguish this Appendix from the body of this report.

This Appendix is the work of Dr. S. E. Phillips of Michigan State University. Dr. Phillips is the only member of the Panel who is qualified to provide a review and evaluation of the KIRIS program from a legal perspective. It is for that reason that other Panel members have not participated in the development of this Appendix nor taken a position on the arguments presented or the conclusions reached in this Appendix despite our belief that a legally oriented review of KIRIS will be of substantial value to the State of Kentucky.

Dr. Phillips' formal education includes a Ph.D. degree in educational measurement and statistics from the University of Iowa and a JD degree from the Thomas M. Cooley School of Law. Dr. Phillips has extensive assessment experience and has conducted extensive research on legal issues in assessment. By virtue of education and experience that other faculty members lack, Dr. Phillips is uniquely qualified to prepare this Appendix.

Appendix A Contents

A.1 ISSUES AND MAIN CONCLUSIONS A-1
A.2 INTRODUCTION A-1
A.3 ROLE OF PROFESSIONAL STANDARDS A-3
A.4 THE DEBRA P. CASE A-3
A.5 DUE PROCESS A-4
A.5.1 Notice A-6
Assessment Specifications & Content/Skill Weighting A-7
Complex Score Weighting A-9
Nominal Versus Actual Weights A-10
Notice of Assessment System Modifications A-10
A.5.2 Curricular Validity for School Accountability Systems A-10
Defining the Purpose for Assessment A-13
Classroom Versus High-Stakes School Assessment . . A-14
Confounding Social & Academic Skills in Writing Assessment A-15
Confounding of Measurement in Other Subjects . . . A-16
A.5.3 School Accountability Standards A-17
Articulating Defensible Standards A-17
Legislative/Administrative Accountability Standards A-18
A.5.4 Use of Accountability Index For Distrib Rewards to Schls A-18
Errors of Measurement A-18
Evaluating Error -- How Much is Too Much? A-19
Gain Scores A-19
Cohort Variability A-20
Framework For Evaluating the Seriousness of Errors A-21
Consideration of Overall Error A-22
A.5.5 Apportionment of Rewards A-23
A.5.6 Appeals A-24
A.5.7 Other Due Process Requirements A-24
A.6 ADVERSE IMPACT A-25
A.6.1 Changes in Adverse Impact A-25
A.6.2 KIRIS Data A-25
A.7 OPPORTUNITY FOR SUCCESS A-28
A.7.1 Application to KIRIS A-30
A.8 ASSESSMENT SECURITY A-31
A.8.1 Application to KIRIS A-33
A.9 ASSESSMENT DISCLOSURE--FREE EXERCISE OF RELIGION/ FREEDOM OF SPEECH A-35
A.9.1 Legal Background For Parental Challenges A-35
A.9.2 Policy Issues Related to Public Test Disclosure . A-37
Equal Opportunity For Success -- Fairness A-37
Professional Standards A-37
Threats to Assessment Validity A-37
Security Risk of Post-Administration Disclosure . A-38
A.9.3 Application to KIRIS A-39
A.9.4 KIRIS Sensitivity Guidelines A-43
A.9.5 Independent Review A-44

A.10 ACCOMMODATIONS FOR PERSONS WITH DISABILITIES A-44
A.10.1 Physical Versus Cognitive Disabilities A-44
A.10.2 Valid and Invalid Accommodations A-45
A.10.3 Excluding Scores From Nonstandard Administrations A-45
A.10.4 Explicating Assumptions A-46
A.10.5 Accommodation Alternatives A-46
A.10.6 Accommodation Issues Unique to Performance Assessments A-46
A.10.7 KIRIS Accommodations A-47
A.11 SUMMARY AND CONCLUSIONS A-49
A.12 RECOMMENDATIONS FOR KIRIS A-50
A.13 REFERENCES A-52
A.14 ENDNOTES A-55

APPENDIX A: LEGAL ISSUES[1]

A.1 Issues and Main Conclusions

This appendix addresses major legal issues of importance to high-stakes assessment programs. It includes discussions of due process (notice, curricular validity, appeals), adverse impact, opportunity for success, assessment security, assessment disclosure and accommodations for persons with disabilities as related to KIRIS. The purpose is to alert OEA to legal issues that may be raised in litigation related to KIRIS. States that attend to these issues in the design of their high-stakes assessment programs decrease the likelihood of court imposed remedies.

Based on prior legal precedents and professional standards, several potentially serious concerns are raised in each of the major areas listed above. If KIRIS remains a high-stakes school accountability system, the following concerns should be addressed:

1. alignment of assessment activities with a single KIRIS purpose;
2. communication of detailed assessment specifications and content weighting to schools;
3. assessment security, ethical guidelines and investigatory resources;
4. validity of accommodations;
5. use of controversial topics;
6. appropriateness of content for all students;
7. confounding of social, process, reading and writing skills with intended measurement;
8. magnitude of error associated with school accountability index change scores;
9. use of ad hoc procedures in equating;
10. defensibility of classification standards, scoring rubrics and portfolio scores;
11. unintended consequences and the implications for content no longer being taught; and
12. adequacy of notification of program changes and appeal procedures.

When assessments are used for high-stakes decisions, good intentions and anecdotal evidence are not sufficient; all relevant professional and legal standards must be met.

A.2 Introduction

In 1990, the Kentucky legislature passed the Kentucky Education Reform Act (KERA; KRS 158.645 et seq.). In response to KERA, the Kentucky Instructional Results Information System (KIRIS) was developed by the Kentucky Department of Education (KDE) under contract with Advanced Systems in Measurement and Evaluation (AS).

KIRIS is a performance assessment system for school accountability designed to identify schools deserving rewards or sanctions based on gains (or losses) over a 2-year period (biennium). For the first biennium (1992-94), KIRIS assessments measured reading, mathematics, science, social studies, writing and noncognitive indicators.

The Kentucky reform effort is an outgrowth of a lawsuit brought by a group of Kentucky districts challenging the constitutionality of the Kentucky education system (*Rose v. Council for Better Education, Inc.*, 1989). The case was based on allegations that the Kentucky public education system violated a provision of the Kentucky constitution requiring the General Assembly to "provide an efficient system of common schools throughout the state" (Section 183).

The emphasis in the *Rose* case was on the inequity of educational opportunity between poor and wealthy districts in Kentucky. There was also evidence from which the court concluded that overall Kentucky was spending too little on education and did not compare favorably with adjacent states or national norms. The implication of the unfavorable national comparisons was that Kentucky schools as a whole were not teaching well the knowledge and skills measured by standardized tests.

Nowhere in the opinion does the *Rose* court indicate that a different kind of assessment system is required. The opinion implies that the goal is to raise the achievement of Kentucky students to a level commensurate with that of surrounding states while at the same time providing equal educational opportunity for all students in poor and rich districts. The majority opinion stated:

> The overall effect of [the] evidence is a virtual concession that Kentucky's system of common schools is underfunded and inadequate; is fraught with inequalities and inequities throughout the 177 local school districts; is ranked nationally in the lower 20-25% in virtually every category that is used to evaluate educational performance; and is not uniform among the districts in educational opportunities. . . .
> . . . A substantial difference in the curricula offered in the poorer districts contrasts with that of the richer districts, particularly in the areas of foreign language, science, mathematics, music and art.
> . . . The achievement test scores in the poorer districts are lower than those in the richer districts and expert opinion clearly established that there is a correlation between those scores and the wealth of the district. . . .
> . . .
> . . . Testimony indicated that not only do the so-called poorer districts provide inadequate education to fulfill the needs of the students but the more affluent districts' efforts are inadequate as well, as judged by accepted national standards. (*Rose* at 197-98)

Therefore, it is not equal outcomes that the court has mandated. According to the *Rose* court, the goal to be achieved was not to find an assessment that will make poor schools and rich schools achieve equal outcomes but to change the allocation of resources to provide equal opportunity and higher achievement for all students. The message was clear but general: spend more money on education and allocate resources fairly across districts. Thus, the *Rose* decision did not mandate that the state adopt unproven reforms or an assessment system which is different from that of all other states.

The court left it solely to the General Assembly to determine how to make the Kentucky education system conform to the uniformity and equality of an "efficient system of common schools" mandated by Section 183 of the Kentucky constitution and consistent with the seven broad goals stated in the opinion (*Rose* at 212). Subsequently, the legislature in enacting KERA and KDE in implementing KIRIS substantially expanded and elaborated the remedy prescribed by the *Rose* court.

Legislators and administrators of high-stakes, large-scale assessment programs such as KIRIS would like a direct answer to the question, "What is required to make the program legally defensible?" Because there are a multitude of ways in which such programs may be challenged, there is no single answer to this question. Answering this question is further complicated for educational performance assessments because as yet there are no specific cases in which courts have enunciated required legal standards. However, there are precedent cases for large-scale multiple-choice tests in education and for performance assessments in employment and higher education contexts. Thus, it is possible to provide guidelines for what may be legally required.

The overarching consideration in all legal challenges is fairness. Legal cases to date have addressed fairness to individual students for assessments used in

high-stakes decisions such as awarding a high school diploma. Although no case has directly addressed fairness for schools when assessments are used for school accountability decisions, what we have learned about fairness for individuals may be applicable by analogy. It would be hard to imagine an assessment which is judged unfair to students to meet standards of fairness for schools. Thus, the following sections of this appendix will examine past legal precedents and their applicability to school accountability decisions.

The following list summarizes the legal issues related to KIRIS which are considered in this appendix:

A. The requirements for notice, curricular validity, and adherence to professional standards as applied to a school-based accountability system;
B. Other due process fairness issues related to school accountability standards, use of an accountability index to distribute rewards/sanctions, appeals, and other issues related to professional assessment standards;
C. Adverse impact on historically disadvantaged groups;
D. Opportunity for success;
E. Assessment security;
F. Free exercise of religion and freedom of speech considerations in assessment disclosure; and
G. Accommodations for persons with disabilities.

The detailed discussions of these legal issues are followed by a summary and conclusions section and a section providing specific recommendations for KIRIS.

A.3 Role of Professional Standards

In reviewing prior cases involving challenges to assessments, it is evident that courts apply relevant professional standards as articulated by expert witnesses. Thus, the AERA/APA/NCME *Standards For Educational and Psychological Testing* (*Test Standards*; 1985) will likely be cited in any legal challenge to an educational assessment program. However, the *Test Standards* were developed when multiple-choice testing was dominant and may not be detailed enough to represent professional consensus for the emerging technologies associated with performance assessments. The revision of the *Test Standards* currently in progress may more fully address such issues. The *Code of Fair Testing Practices* (*Code*; 1988) also provides guidance for assessment developers and users and may be cited by measurement experts. The *Test Standards* and *Code* are cited where relevant in the following discussions.

According to the *Test Standards*, "***Primary standards*** are those that should be met by all tests before their operational use and in all test uses, unless a sound professional reason is available to show why it is not necessary, or technically feasible, to do so in a particular case. Test developers and users and, where appropriate, sponsors, are expected to be able to explain why any primary standards have not been met" (p. 2). The *Test Standards* define **conditional standards** as those which vary in importance by application and may be primary or secondary depending on the context. "In deciding whether to take an individual conditional standard as primary or secondary, one should consider carefully the feasibility of meeting that standard in relation to the potential consequences to all parties involved in the testing process. . . . [I]f the use of a test is likely to have serious consequences for test takers, especially if a large number of people may be affected, conditional standards assume increased importance" (p. 3).

A.4 The *Debra P.* Case

To provide a framework for evaluating the legal defensibility of assessments used for school accountability, it is instructive to examine past cases to determine the ways in which courts tend to approach assessment-related challenges. The landmark case for large-scale educational assessment is *Debra P. v. Turlington* (1979; 1981; 1983; 1984). The *Debra P.* case involved the denial of high school

diplomas to historically disadvantaged students who had failed a minimum competency examination in Florida.

In 1976, the Florida legislature established the Functional Literacy Examination (FLE) as the state's graduation test, effective for the 1979 graduating class. The FLE was a multiple-choice test of basic communication and mathematics skills applied to real life situations. Graduating seniors who had not passed the test after multiple retakes were awarded a certificate of completion. After three administrations of the FLE, approximately 2% of the Caucasian seniors had not passed while approximately 20% of the African-American seniors had not passed. A more extensive discussion of the *Debra P.* case can be found in Phillips (1990).

A.5 Due Process

The court's response to *Debra P.*, a case of first impression on mandatory statewide graduation tests, was to craft new legal requirements based on professional testing standards and prior legal cases on related issues. These new due process legal requirements included notice and curricular validity. Florida's graduation test was judged legally defensible only after the state presented the court with evidence that the graduation test satisfied the new legal standards imposed by the court. The new legal requirements test imposed by the *Debra P.* court are consistent with constitutional requirements in all due process or equal protection issues and will be pertinent to any such concerns or challenges to KIRIS.

The challenge in the *Debra P.* case occurred because some people were upset with the new student accountability testing. Similarly, those who believe they have been treated unfairly by the new KIRIS school accountability assessment program may ask the courts to craft new legal requirements for school accountability assessments. It is hard to imagine that a court presented with the deficiencies described in this panel report would decide that KIRIS is exempt from professional standards because no legal standards currently exist for school accountability assessment. If sufficient evidence is presented to convince a court of unfairness, the court may choose to establish legal requirements for school accountability assessments which are consistent with professional standards as articulated by national psychometric experts and which incorporate legal principles adapted from related cases. As in the *Debra P.* case, the court could remand such a case to give Kentucky an opportunity to present the requisite evidence.

Kentucky has two choices for addressing a potential legal challenge to KIRIS. Kentucky may decide to defer action until a court mandates new legal requirements. Alternatively, Kentucky could be proactive in revising KIRIS and collecting evidence to demonstrate that the school accountability assessment program is fair and consistent with professional standards. Such good faith efforts may deter future legal challenges and/or avoid the kind of protracted litigation that occurred in the *Debra P.* case.

The due process requirements involved in the *Debra P.* case derived from the fourteenth amendment of the U.S. Constitution. Infringement of a property right is a threshold condition for applicability of fourteenth amendment protections.

Several groups of individuals, including students, teachers and administrators, may have property rights with respect to KIRIS assessments. First, if high school graduation is contingent upon KIRIS performance, students who do not graduate because of KIRIS scores may have a property interest in the diploma which triggers due process scrutiny of KIRIS. For example, the Oldham County Board of Education Policy on Graduation Requirements states, "Beginning with the Class of 1996, in order to receive a diploma from the Oldham County Schools, a student would need to: . . . C . . . 2. Complete the 12th grade writing AND mathematics portfolio to *at least the apprentice level* . . ." (p. 3, emphasis added).

Second, teachers may have a property interest because of employment status changes or employment bonuses which are triggered by KIRIS results. KRS 158.6455(5) provides: "When a school is declared to be a *'school in crisis,'* the following actions shall be required: (a) the full-time and part-time *certified staff of that school shall be placed on probation;* (b) the principal shall immediately notify the students' parents of the students' right to transfer to a successful school . . . (c) . . . If a decline in student enrollment causes overstaffing at the 'school in crisis,' personnel shall be reduced or transferred . . ." (emphasis added). Part (a) mandates a change in employment status and part (b) may precipitate a reduction in force and consequent loss of employment for some staff. Such reductions may fall more heavily on minority staff if they have less seniority. Moreover, teachers on probation due to the "in crisis" status of their school are subject to intensive evaluation and personnel recommendations by a distinguished educator. Such scrutiny and potential for dismissal would not occur in the absence of unacceptable KIRIS scores. Further, KRS 161.790(2) provides: "Charges under [subsections] of this section shall be supported by a written record of teacher performance by the superintendent, principal or other supervisory personnel of the district, *except when the charges are brought as a result of a recommendation made by a Kentucky distinguished educator*. . . ." (emphasis added).

At the other end of the scale, rewards may also trigger due process issues due to employment implications. Even though rewards accrue to schools, staff determine the distribution of funds. Because the majority of staff have voted to award the bulk of reward monies to teachers, the teachers in reward schools are effectively receiving employment bonuses. Such employment bonuses may be covered by the due process protections in the EEOC *Uniform Guidelines*. The *Uniform Guidelines* include professional assessment standards on validity and reliability for assessments used for employment decisions. The assessment need not be the sole factor in an employment decision to trigger due process protections (see *Wards Cove*, 1989).

Third, property interests may be implicated for superintendents. KRS 158.6455(7) provides: "[A distinguished educator assigned to a school in crisis shall make a recommendation on the superintendent's status.] *If the recommendation is to terminate the superintendent, the board shall terminate the contract*. . . ." Although KIRIS results do not directly remove a superintendent, they trigger intensive scrutiny by a distinguished educator which may result in dismissal. Such scrutiny and increased probability of termination are triggered only when KIRIS results fall below a specified level. Even though these in crisis provisions were not fully implemented in Biennium I, they are scheduled to be fully operational for the next biennium.

Notwithstanding any interests an individual student or teacher may have, there may also be the possibility of a class action by students or teachers in a school which believes it has been treated unfairly. While the state may have the legal authority to establish unique school accountability rules for benefits and sanctions, once established, rules for receiving the benefits or sanctions must be applied consistently and fairly. Consequently, in administering the program, the state may not treat similarly situated schools differently.

For example, it is unlikely that a court would uphold the distribution of welfare benefits in a manner in which the total benefit received depended on the particular clerk who processed the application. If two similarly situated applicants received different welfare benefits because one clerk counted alimony as income while the other did not, the administration of the benefit would violate due process requirements.

Assessments that do not meet professional standards significantly increase the likelihood that similarly situated schools may be treated differently. Thus, given appropriate evidence of unfairness, a court could find that KERA is legally defensible on its face but defective in its application using KIRIS results.

There are two types of due process which have been recognized by federal courts: procedural and substantive (Nowak, et al. 1986; *Debra P.*, 1979-83). Procedural due process focuses on assessment administration and sets an expectation that the procedures and processes implemented by an assessment program will be fair and equitable. Substantive due process focuses on the assessment instrument itself and sets an expectation that it follow professional standards, be valid and reliable, and be fair to all students.

Substantive due process can be violated when the knowledge and skills being assessed are judged to be invalid, arbitrary, capricious or unfair. Although in other contexts the cases seem to suggest that substantive due process is violated only by "behavior so outrageous as to shock to conscience," courts have seemed more willing in assessment cases to rule on the appropriateness of actual test items vis-a-vis professional standards (see e.g., *PASE v. Hannon*, 1980 [judge reviewed ability test items to determine if culturally biased]; *Allen v. Alabama State Board of Education*, 1985 [items reviewed for content necessary for teachers]; *United States v. LULAC*, 1986 [court considered the validity and fairness of a test for measuring the skills necessary for teacher education training]).

A.5.1 Notice

Notice is a procedural due process requirement. The *Debra P.* court held that students who might be deprived of their property rights in a diploma must receive adequate prior notification of any required assessment. Similarly, for schools being held accountable for the assessment performance of their students, fairness dictates that they have adequate notice of what they will be held accountable for. Notice means not only knowing in advance but being given clear information about what is to be taught and the specific outcomes to be expected.

The courts have not specified the length of the required notice period but it probably depends, in part, on the success of assessment administrators in disseminating relevant information about content, format and scoring to school personnel. Although it is probably not necessary to communicate specific passing standards ahead of time, school personnel should be provided with clear indications of the specific knowledge and skills for which they will be held accountable and guidelines on what constitutes acceptable performance.

To illustrate the notice requirement for school accountability, consider the following example. Suppose schools in State Q were told in the spring of 2001 that assessment to determine the percent of good citizens would begin the following year. To determine if each student was a good citizen, a written citizenship essay exam would be given and a classroom observer would rate each student based on one classroom visit. No further information would be available about the written exam or classroom observation. The written test and classroom observation score would each count 50%. Schools with more than 90% good citizens would receive monetary rewards; schools with less than 25% good citizens would be taken over by the state education department. School Q1 is facing a deficit and wants to earn the reward money. Has School Q1 been given adequate notice for school accountability? Or put another way, has School Q1 been given a fair and equitable opportunity to earn the reward money?

Most educators and parents would probably agree that they want students to become good citizens. But what is a "good citizen"? Citizenship has many facets such as knowledge of government and how it works, adherence to moral and legal standards of behavior, informed voting, obeying local, state and federal laws, etc. Thus, a major problem with the above example is that the state has not clearly defined what constitutes good citizenship and what specific student outcomes are desired. Reasonable educators would likely differ on both the definition of "good citizenship" and on what observable behaviors indicate its presence. This is an example of an undefined assessment domain. School Q1 has no idea what to teach students and therefore has not received proper notice of what is expected. If School Q1 receives a reward, it will result from luck rather than targeted instruction.

In Kentucky, the same question about notice could be asked about KIRIS: were schools given adequate notice of the domains of knowledge and skills that were to be assessed? Was sufficient detail provided so that schools could determine what student outcomes were necessary to be successful? To examine these questions, one must consider the domain definitions provided to schools and the information schools were given about how the various assessment scores would be combined to determine school accountability indices.

Assessment Specifications and Content/Skill Weighting

With respect to test development and domain definition, the *Test Standards* (1985) provide:

> **Standard 1.6**
> When content-related evidence serves as a significant demonstration of validity for a particular test use, a clear definition of the universe represented, its relevance to the proposed test use, and the procedures followed in generating test content to represent that universe should be described. When the content sampling is intended to reflect criticality rather than representativeness, the rationale for the relative emphasis given to critical factors in the universe should also be described carefully. (p. 14) **(Primary)**
>
> **Standard 3.2**
> The specifications used in constructing items or selecting observations and in designing the test instruments as a whole should be stated clearly. The definition of a universe or domain that is used for constructing or selecting items should be described. When, for reasons of security, sample copies of a test are unavailable for inspection, the descriptive information should include a representative item identified with each major cell in the classification or domain definition. When item difficulty is a facet of such a system, items representative of the difficulty levels should be provided. (p. 25) **(Conditional)**
>
> **Comment**
> . . .
> The adequacy and usefulness of criterion-referenced interpretations depend on the rigor with which the behavioral domain represented by the test has been defined. Such interpretations are intended to describe the status of individuals or groups with respect to one or more behavioral domains, and it is the domain definition that provides the primary reference for interpretation of scores and for judging the adequacy of the test. The domain definition should be sufficiently detailed and delimiting to show clearly what facets of behavior are included and what facets are excluded in the domain. Within the domain, the classification system adopted should show clearly what and how many facets of behavior the domain comprises. (p. 26)

Initially, schools were on notice of the outcomes specified in the Kentucky Educational Reform Act (KERA; 1990). KERA contained general statements of broad educational goals including communication skills, making choices, understanding government, maintaining mental and physical health, arts appreciation, vocational choice, and favorable competition with students in other states. In December 1991, the State Board adopted a list of 75 valued outcomes which would form the basis for the KIRIS assessments. In 1992, the list was revised and reduced to 57 academic expectations. These were more specific but still included undefined terms such as critical thinking, creative thinking, and connecting knowledge, and these expectations lacked specificity regarding the context/content in which skills such as problem solving would be assessed. Because process skills must be assessed in a specific context and performance on one task often does not

generalize to others, it is likely that students with more background knowledge about the content of a problem-solving task will perform better.

Meanwhile, during the same 1991-92 school year, KDE and AS developed and administered baseline assessments in reading, mathematics, science, social studies and writing. Although the assessments contained both multiple-choice and open-ended questions, only the new format open-ended questions counted in determining school accountability. Two other novel formats, performance events and portfolios were also included in the school accountability assessments.

"The first task for test developers is to specify adequately the universe of content that a test is intended to represent, given the proposed uses of the test " (*Test Standards*, p. 10). Assessment developers typically begin by producing a table of specifications for each assessment (e.g. reading) which lists the content/skills to be covered and the proportion of the assessment devoted to each. By comparing assessment tasks/items to the table of specifications, the assessment developer can ensure that only the intended content is assessed in the proportions specified by the assessment user (in this case, KDE). For the 1991-92 KIRIS assessments, general matrix specifications were given (see Technical Manual p. 14-24) for the open-ended items but no cell weights were specified. Thus, for example in science, a school would not know what proportion of the assessment would include astronomy nor would the school know what knowledge and skills in astronomy students would be expected to know.

Because AS had been awarded the KIRIS contract in late 1991 with short timelines for producing the first assessments in 1992, AS produced the initial KIRIS assessments by using items from NAEP, Maine and Massachusetts with a few newly-developed items. "These items were used as they were or modified to reflect a Kentucky context" (memo from Lisa Ehrlich, AS, to Neal Kingston, KDE, April 30, 1995). These items from other states and from NAEP were selected without developing a detailed table of specifications. Thus, what was assessed in the first year of KIRIS was a function of KDE and AS judgment of appropriateness of or revision of existing items/tasks from programs outside Kentucky. After the first administration of KIRIS, AS and KDE reviewed and classified the assessment items according to the relevant subject framework but still did not specify cell weights for each assessment. There appeared to have been no attempt made to describe or justify the particular weighting of content/skills that appeared on the initial KIRIS assessments.

General descriptions and sample items were printed in an issue of a statewide school bulletin (All About Assessment, *EdNews*, KDE, Jan/Feb 1992). The 1991-92 assessments were administered that spring.

The initial 1991-92 KIRIS assessments became the model for KIRIS assessments in subsequent years, 1992-93 and 1993-94. Substantial numbers of items from the first assessments were reused in subsequent KIRIS assessments. No explanation was given for the particular skills selected for common items and those selected for matrix items nor was there a systematic plan for reuse and replacement of items. Thus, it is doubtful schools were able to anticipate precisely what would be assessed, how the various content/skills would be weighted, or where to target their instructional efforts.

In 1994, a statewide Sensitivity Review Committee identified a lack of content balance across forms of the science assessment and stated: "These finding once again emphasize the need for a 'blueprint' prior to developing any new test item. . . . Include more content and less opinion" (KIRIS Review Committee Report, June 29, 1994, p. 3).

Even if schools had had some idea of assessment content, there was little preparation time between adoption of the revised outcomes (1992) and the two years of assessment which counted in Biennium I (1992-93 & 1993-94). KDE and AS have since developed more detailed content specifications but they were disseminated too late for use in Biennium I.

After the 1991-92 assessments had been administered, the common items were released to schools. Although such release probably helped schools understand the format of assessment questions, such items were only a sample of what could be tested. Schools thus were on notice of the processes and format of assessment but not what specific content to teach. Further, the 1991-92 common items may not have been released soon enough to impact instructional planning for the first year of Biennium I (1992-93).

Complex Score Weighting

With respect to use of gain scores and score weighting, the *Test Standards* (1985) provide:

> **Standard 1.3**
> Whenever interpretation of . . . score differences . . . is suggested, the evidence justifying such interpretation should be made explicit. Where composite scores are developed, the basis and rationale for weighting the subscores should be given. (p. 14) **(Primary)**
>
> **Standard 1.15**
> If more than one criterion measure is obtained, but, for purposes of a particular study, a single composite criterion score is used, the rules for criterion combination should be described. (p. 16) **(Primary)**
>
> **Standard 12.4**
> The methods used to aggregate test results from an individual level to a group level should be described clearly. (p. 68) **(Primary)**
> **Comment**
> The aggregation method that is used will influence the description of needs and outcomes in an evaluation, the size of estimated program effects, and the inferences and legitimacy of statistical tests. (p. 68)

In addition to lack of clear specification of the domain of knowledge/skills to be assessed, the complexity of the computation of the accountability index made it difficult for schools to determine what they should do to score well and receive reward money. Six major steps were necessary to produce a school accountability index from individual student responses. The steps were as follows:

1. score student responses;
2. convert scores to performance levels;
3. combine/weight common & matrix data;
4. combine/weight common/matrix & performance event data;
5. combine/weight cognitive data across five subjects;
6. combine/weight cognitive and noncognitive data;
7. compute weighted average across grades; and
8. compute weighted average across the two years of the biennium.[2]

The complexity of sequential mathematical computations, repeated weighting of adjusted and combined data, and final weighted averaging across grades and years made it extremely difficult for schools to figure out what was necessary to score well or to evaluate whether the accountability index was capturing the desired curricular/instructional change. Thus, the transition from student performance to accountability index may have appeared to have been like a black box in which schools were asked to trust that the indices appropriately reflected student performance on important goals. However, teachers did have professional development opportunities (e.g., serving on committees; scoring portfolios) which provided opportunities to learn more about the content of the KIRIS assessments.

It is ironic that the richness of data and responses beyond mere recognition of answers cited as advantages of adopting performance assessments are obscured when one reduces such performances to a single number for school accountability purposes.

Nominal Versus Actual Weights

Nominal weights are those set by the Board of Education in the KIRIS Regulations. The actual weight of a component represents its real contribution to the total score.

Because KIRIS scores differed in variability across components, the actual weights were only approximately equal to the nominal weights set by the Board. For example, due to lack of variability in the noncognitive indicators, the noncognitive component functioned like a constant and had an actual weight much less than its 1/6 nominal weight.

Notice of Assessment System Modifications

Adequate notice to schools is also required when the state decides to change the assessment system rules, add new assessments, or change the specifications for existing assessments. Such changes should not be made in midstream. Changes should be announced far enough in advance to give schools adequate time to adjust and whatever assessment system rules are in effect at the beginning of a biennium should remain fixed for the duration of that biennium.
Unfortunately, changes in KIRIS were made during the first biennium. For example, schools learned that the multiple-choice items did not count and that they had three choices for adjusting their initial writing portfolio scores.

A.5.2 Curricular Validity for School Accountability Systems

Curricular validity is a substantive due process requirement. The *Debra P.* court derived the curricular validity requirement from the substantive due process requirement for fundamental fairness in a high-stakes environment.

Curricular validity for a graduation assessment requires assessment administrators to demonstrate that students have had an opportunity to learn the knowledge and skills included on a graduation test. The analogous requirement for school accountability assessments is teachability. That is, to be fair to schools being held accountable, the state must assess knowledge and skills for which schools have a reasonable expectation of teaching.

> **Standard 8.11**
> Test users should not imply that empirical evidence exists for a relationship among particular test results, prescribed educational plans, and desired student outcomes unless such evidence is available. (p. 54) **(Primary)**

For example, suppose the state decided to use an IQ test to assess schools for rewards/sanctions. Educators, psychometricians, legislators and policymakers would all agree that such an action would be inappropriate because IQ tests do not measure the content schools are expected to teach.

The *Debra P.* court held that the relevant evidence for establishing curricular validity was that "the skills be included in the official curriculum and that the majority of the teachers recognize them as being something they should teach." (564 F.Supp. at 186). The collection of a variety of evidence including teacher surveys, student surveys and reviews of textbooks and curricular guides is desirable in establishing curricular validity. The analogous requirement for school accountability is that schools be given clear guidance about what they are accountable for and that what is assessed be clearly linked to teachable student outcomes. To the extent that the skills required for success on performance tasks can not be specifically taught in a short time frame, those assessed skills

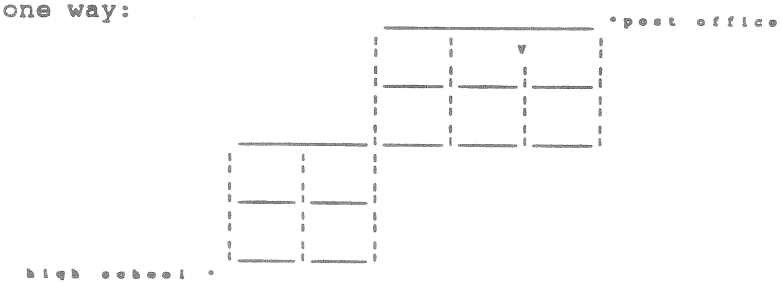
may be closer to abilities (for which focused instruction has only limited success) than to achievements (which can be directly taught).

> **Standard 3.14**
> A test that is designed to measure learning from practice, coaching, or instruction should be shown to do so . . . . (p. 28) **(Secondary)**

To illustrate, consider the teachability of the mathematics portion of the KIRIS assessments. Some of the items on the 1993-94 12th grade assessments might be characterized as brain teaser or math puzzle type exercises (see Beck, in press). While intended to measure problem solving skills, they appear not to be consistently solvable by the systematic strategies typically taught in mathematics. Instead, success on these items may be more a reflection of trial and error, persistence and luck. Examples:[3]

a) 1993-94 common item asking for the number of zeros at the end of 100! In addition to finding a systematic way to count combinations of factors which end in zero (e.g. 25 x 4), the student must remember to account for 2 zeros for the "100".

b) 1993-94 common item asking how many routes from the high school to the post office, in the diagram below, if travel must always be to right and up and the block labeled "v" is one way:



c) 1991-92 common item asking students to set up a schedule for 3 riders to each ride 4 horses for 30 minutes from 2 to 4 o'clock in the afternoon. The following are the sample distinguished and proficient responses:[4]

Distinguished

Horses A, B, C, D; Riders a, b, c

Time	Horse A's Rider	Horse B's Rider	Horse C's Rider	Horse D's Rider
2:00-2:30	a	b	c	
2:30-3:00	b	c		a
3:00-3:30	c		a	b
3:30-4:00		a	b	c

Proficient

	horse 1	horse 2	horse 3	horse 4
2:00-2:30	rider #1	rider #2	rider #3	
2:30-3:00		rider #1	rider #2	rider #3
3:00-3:30	rider #3		rider #1	rider #2
3:30-4:00	rider #2	rider #3		rider #1

The criteria for a distinguished response included: student understands all aspects of task; schedule shows good notation and organization, making it easy to read and check; representation of riders and horses is well defined; strategy for placement of items within table is accurate and efficient.

Apparently, the first solution is better because letters were used for riders and horses. Since the question is specific about 3 riders and 4 horses, it is not clear why letter designations are better than number designations. How would the second solution have scored if H1 . . . H4 and R1 . . . R4 had been used? Further, the scheme used in the second solution appears more efficient, systematic and easily checked with rider #1 riding each horse in turn. Thus, it appears that the second solution satisfies all the criteria for a distinguished response.

Other questions from the Grade 12 KIRIS Mathematics Assessments measured skills such as solving quadratic inequalities, graphing and solving exponential functions, probability and trigonometric functions. These are advanced mathematics topics typically studied by college bound students.

In addition, one question included a reading passage about mathematics. Does this represent an appropriate classroom activity? To what extent does it confound the measurement of math skills with the measurement of reading skills in another section of the assessment?

The 1993-94 math assessment also contained a question involving musical notes. Students were given all the information necessary to figure out the answer, but students in band and choir probably would find this item easier than those with no musical background. Again, is musical notation something math teachers ought to teach?

The following table provides a summary of the content of the 12th grade KIRIS mathematics assessments across forms for 1991-92 and 1993-94. The numbers of multiple-choice items in each content category are given; numbers in parentheses are for open-ended items.

1991-92	Common	1	2	3	4	5	6	7	8	9	10	11	12
computation	8	m	1	1	4	5	3	6	3	3	5	3	4
algebra	13	i	5	4	2	4	2	2	2	(1)2	1	4	2
adv. algebra	2	s	2	2	1	1	3	0	0	1	1	2	1
geometry	8	s	4	5	2	2	(1)4	4	(1)6	4	4	(1)4	5
trigonometry	0	i	0	0	0	0	0	0	1	0	0	0	0
puzzles	(2)4	n	1	2	(1)3	(1)1	1	1	1	1	(1)2	0	(1)0
graph read	(1)3	g	0	1	0	0	0	(1)1	0	0	0	1	1
statistics	2	!	(1)2	(1)2	1	2	2	1	2	4	2	1	2
	(3)40		(1)15	.	.	.	.	.	.	.	.	.	.

1993-94	Common	1	2	3	4	5	6	7	8	9	10	11	12
computation	3	1	(1)2	1	2	(1)1	0	2	1	1	2	0	2
algebra	2	0	0	0	(1)1	3	2	1	2	0	1	3	0
adv. algebra	(1)3	0	0	2	0	0	(2)2	0	2	2	2	1	(1)1
geometry	(2)3	4	3	2	3	(1)2	2	(2)2	(1)2	2	(1)1	(1)1	2
trigonometry	0	0	0	0	0	0	0	0	1	0	0	0	0
puzzles	(2)6	(1)1	(1)1	(2)3	(1)1	1	0	0	0	(1)2	(1)1	2	(1)2
graph read	0	1	1	0	0	0	0	1	0	0	1	(1)0	0
statistics	3	(1)1	1	0	1	1	2	2	(1)0	(1)1	0	1	1
	(5)20	(2)8	.	.	.	.	.	.	.	.	.	.	.

Notice that the distribution of content for the open-ended items differs significantly across forms. For example, in 1993-94, the content of both matrix items was advanced algebra for form 6 but geometry for form 7. Students who had never taken advanced algebra but were unlucky enough to be administered form 6 would have little chance of demonstrating competence. In addition, 20% of the common items score would also be based on advanced algebra content. For students administered forms 1, 2, 3, 4, 9, 10, or 12, 40% of the common items score and half of the matrix score was based on puzzles.

The table below provides an overall comparison of the content assessed in 1991-92 and 1993-94 on the common and matrix open-ended items.

Content	1991-92 Common	1991-92 Matrix*	1993-94 Common	1993-94 Matrix
computation		0		2
algebra		1		1
adv. algebra		0	1	3
geometry		3	2	6
trigonometry		0		0
puzzles	2	4	2	8
graph read	1	1		1
statistics		2		3

* Form 1 data (1 item) missing

The only overlap in content for the common items is puzzles which constituted 67% of the 1991-92 assessment and 40% of the 1993-94 assessment. To score well on the 1991-92 common items, students needed competence in graph reading but in 1993-94 needed advanced algebra and geometry skills. Thus, the content and weighting of the common items in the two years differed and the two common item assessments would not be considered parallel forms.

Similarly, the overall content of the matrix items in the two years differed. There were no computation or advanced algebra open-ended matrix items in 1991-92 but 20% of the 1993-94 open-ended matrix items covered these two content areas.

The examples and tables presented above might lead policy makers to ask the following questions:

1. Do the math puzzles in the KIRIS assessments measure teachable skills or generalized ability?

2. If these tasks do represent teachable skills, do they represent the kinds of activities appropriate for math instructional time in Kentucky public schools?

3. Do these tasks use the kinds of systematic problem solving skills taught in mathematics or are they solved by trial and error, persistence and luck?

4. Do these tasks reflect the "real world" skills the Kentucky Supreme Court and Council of School performance Standards envisioned?

5. Should *all* students in Kentucky be expected to learn advanced algebra and trigonometry? Is such content necessary for daily living?

6. Given the different mix of content assessed in 1991-92 and 1993-94, do the scores provide meaningful comparisons for assessing school improvement?

Defining the Purpose For Assessment

The cardinal rule in assessment is that one must begin by clearly specifying the purpose of the assessment. The purpose of the assessment must be clearly articulated before assessment construction begins. Once the purpose has been established, a table of specifications should be developed which enumerates the content and processes to be measured and the proportional weight given to each. Once the specifications have been developed and approved, assessment construction can begin. As assessment construction proceeds, all decisions should be consistent with the stated purpose and written specifications.

Using a single assessment system to serve two purposes can be at best inefficient and at worst indefensible. Such assessment systems may serve neither purpose well.

The Kentucky Instructional Results Information System (KIRIS) has two major purposes: (1) to alter curricula and instructional practices in Kentucky; and (2) to identify schools for rewards and sanctions based on their success in achieving the educational goals mandated by the Kentucky legislature in KERA. In addition to articulating general educational goals, the Kentucky legislature also mandated that the assessment system be primarily performance based.

Purpose (1) includes an underlying assumption that implementation of popular instructional/curricular reforms and the use of performance assessment will result in improved educational outcomes for Kentucky students. To determine whether purpose (1) has been achieved, evidence of actual instructional/curricular changes together with their positive and negative consequences is required. Efforts to achieve purpose (1) may not have direct, high-stakes implications for students or schools, but probably have significant consequences for the educational experiences and outcomes for the next generation.

Purpose (2) involves a high-stakes use of assessment results and therefore presents the most serious legal concerns. Although there may also be legal concerns surrounding purpose (1), they are less clear and may involve issues such as local control and state funding which are beyond the scope of this review. The remainder of this appendix will concentrate on the legal concerns related to the high-stakes use of KIRIS for school accountability.

Good teachers used performance assessments in the classroom for years prior to the advent of statewide accountability assessments. The move toward performance assessment for high-stakes accountability appears to have been driven by teachers' inappropriate use of multiple-choice tests as models for instruction (see DePascale, 1992). There are two major reasons why teachers may have done this: (1) they were looking for a shortcut to obtain rewards or avoid sanctions associated with student performance, or (2) they were ignorant of appropriate classroom assessment practices. Rather than change their instructional practices, teachers for whom the former explanation applies might be expected to look for new shortcuts when performance assessments are implemented. However, it is likely that the latter explanation applies to many teachers because most states and teacher preparation programs require no formal training in assessment.

If the second explanation is correct, rather than using classroom performance assessment for a high-stakes accountability purpose for which it is not well-suited, there may be more cost effective and direct methods for changing teachers' classroom practices. For example, the state could provide professional development on classroom assessment and instruction for all teachers. If ignorance is the problem, teachers should welcome such assistance. Without concurrent accountability pressures, teachers may more readily embrace the hard work of doing performance assessment well and have an incentive for improving long term instructional outcomes rather than looking for short-term shortcuts for raising student scores.

Classroom versus High-stakes School Assessment

Assessments used to determine whether desired curricular/instructional changes have been successfully implemented (purpose 1) are most appropriately focused on the classroom and typically would include the collection of data documenting classroom practices and processes. This data might take the form of systematic observations, checklists, anecdotal records, teacher surveys, etc. For example, based on current reform efforts in the area of writing, officials might seek evidence of prewriting activities, teacher collaboration, interdisciplinary topics, simulation of real world tasks, students working together in groups, feedback from student to student on editing and polishing a piece of writing, and more able students helping the less able.

The valued outcomes listed in the example above reflect an assumption that the specified process is superior to all others and will result in improved student outcomes. Regarding purpose (1), there are two important assessment questions -- Is the specified process being used? and Does the use of this specified process create improved outcomes for students? The former question can be addressed using the types of evidence suggested earlier. The latter question requires further delineation of what outcomes are important, establishment of the link between process and product, and examination of possible unintended outcomes.

Purpose (2) has a different focus. Rather than focusing on instructional processes in the classroom, purpose (2) seeks to measure large-scale, high-stakes school outcomes and to reward schools for improving student products. The language of the statute indicates that schools should be rewarded when student skills (e.g. writing products) have improved. The statute does not specify any particular process that schools must use to improve writing. However, the assessment system which has been designed to measure the improvement of writing includes components mandating that schools implement a specific writing process. Thus the curricular/instructional reform purpose has been confounded with the school reward for improved writing product purpose. This results in an assessment that is not directly targeted toward measuring achievement of either objective. The implications of this confounding for the high-stakes use of the writing assessment are further illustrated in the following extension of the collaborative writing example discussed earlier.

Confounding Social & Academic Skills in Writing Assessment

> **Standard 3.13**
> Probable sources of variance that would confound the . . . domain definitions underlying the test should be investigated by the test developer, and the implications of the results for test design, interpretation, and use should be presented in the technical manual or in supplementary reports. In general, evidence from research should be provided to justify the use of novel item or test formats. (p. 28) **(Conditional)**

Collaborative student writing emphasizes the importance of following the correct process when writing. It contains a social as well as an academic agenda -- students are expected to help each other in the writing process. Good writers may be penalized if they refuse to help their less able peers and less able writers may be rewarded for being adept at finding persons to help them. For example, a student with good social skills who can get the teacher to help suggest a topic and how to approach it, a fellow student to provide ideas for the piece, another peer student to edit the piece for grammar mistakes, and a parent to print the final copy on a laser printer with fancy graphics, may receive a passing score even though the student could have accomplished none of these steps alone. Thus, when the final product is judged, it is not clear to what extent a passing grade reflects the student's, the teacher's, the two peer students' and the parent's efforts. The score does not reflect what the student can do academically but is influenced by what the student can do socially. Thus, the assessment confounds the measurement of academic and social skills and measures neither accurately.

The point of the above example is that an assessment system which prescribes a process (collaborative student writing) but judges the product (the quality of the piece of writing produced) confounds the assessment of the process (social skills) with the assessment of the quality of the product produced (academic skills). As a result, one cannot tell whether a high score results from superior social skills, superior academic skills, or a combination of both. This is especially important when the two purposes of the assessment system are considered. Purpose (1) is probably more concerned with process outcomes while purpose (2) is more concerned with product outcomes. That is, in rewarding schools (purpose (2)), legislators may be most concerned with improved writing products and willing to accept multiple methods for doing so. Thus, they would

desire an assessment measure which does not confound the measurement of writing product quality with process requirements.

Similarly, legislators interested primarily in reform may want assessment measures which focus on the writing process. For this purpose, rather than assuming the specified process was implemented correctly, assuming that good products result only from the specified process, and then judging the process indirectly through the writing products produced, a more direct assessment of the process itself is needed. Once the quality of the specified process and the link between the process and quality outcomes have been established, unintended outcomes have been examined, and the efficacy of alternative processes considered, legislators will be in a position to judge the success of specific reforms. One should not assume without evidence that a particular process will produce the desired outcomes and one must carefully examine any unintended consequences of mandating specific processes for all students. Educators may support individualized instruction to meet each student's needs but then mandate the latest reform for everyone without clear evidence that all students will benefit.

> **Standard 6.5**
> Test users should be alert to probable unintended consequences of test use and should attempt to avoid actions that have unintended negative consequences. (p. 42) **(Primary)**
> **Comment**
> . . . Obviously, test users cannot anticipate every unintended consequence. What is required is an attempt that is reasonable and made in good faith to avoid unintended consequences that might be anticipated.
>
> **Standard 12.8**
> When test results are used wholly or in part to allocate funds to [school districts], the positive and negative anticipated consequences of such use should be described to policy makers by those test professionals who are closest to the policy before the policy is implemented. (p. 69) **(Secondary)**

Confounding of Measurement in Other Subjects

The confounding in the Kentucky assessment system is not confined to writing. Performance events and math portfolios are also confounded by the interaction of social and academic skills. Open-ended questions in subject areas such as science and social studies confound reading and writing skills with subject matter knowledge. Thus, when a student scores poorly, one can not tell whether the problem is reading skill, writing skill, content knowledge, or a combination of all three. Writing is also confounded with the measurement of other subjects because all other standards were set based on the writing standards.

In October 1994, an ad hoc committee reviewed the appropriateness of KIRIS for deaf and hearing impaired students (Biennium I Technical Manual, 1985). The committee concluded that "although the assessment of reading and writing skills was critically important, having all subject areas so dependent on this form of communication posed a particular challenge and perhaps a barrier to success for deaf and severely hearing impaired students" (p. 63).

Until very recently, schools have been given very little guidance regarding content coverage of the assessments; schools have been encouraged to focus their attention on processes such as problem solving. But problems are not solved in a vacuum and background knowledge of the context can make a significant difference in student success. We know this is true because research has shown that assessment tasks are not generalizable. That is, students who score well on one math problem solving task may score poorly on another, even when the specific skills required are the same. Thus, content does make a difference and schools whose students knew something about the content of a task probably scored higher than those whose who did not. But because little information was

available about the domain of content which might appear on the assessment or about the weighting of content and process areas on the assessment, schools had little information for preparing students or for determining their content accountability.

More specific content guidelines were issues by KDE in 1994, too late to be of use for the first biennium. The document states: "The Content Guidelines do not indicate particular topics to be taught and learned at specific grade levels nor the sequence of subject matter" (p. ii). Thus, this document is primarily process-orientated and is fairly general with respect to content to be covered on the various KIRIS assessments.

Finally, there are special concerns to be addressed as programs move from multiple-choice tests to performance assessments. This occurs for a variety of reasons including the lack of experience with performance assessments in subjects other than writing, the lag in the development of defensible technologies and methodologies for scaling and equating performance assessments, and the increase in errors due to lack of generalizability of task samples and the fallibility of human judgments. These issues are discussed in greater detail in Chapters 3 - 5. Further discussion of the legal issues related to performance assessment can be found in Phillips (March 11, 1993; 1993).

In Kentucky, the multiple-choice items, most psychometrically sound and least likely to have confounded the measurement of academic skills with social, writing or reading skills, were not used at all in determining school outcomes. When schools found out that the multiple-choice items did not count, students became unmotivated or failed to respond at all to the multiple-choice items. Thus, a potentially strong equating link between years and forms was also lost.

A.5.3 School Accountability Standards

Due process also requires that school accountability standards be fair. There are two interpretations of the term "standards" which may play a role in the legal defensibility of a high-stakes, large-scale assessment program. These interpretations include: (1) goal statements describing what students should know and be able to do in specific content subjects, and (2) the specification of the score or level of performance corresponding to a category of achievement such as "novice" or "proficient." Both interpretations of the term "standards" are included in the discussions of legal defensibility which follow.

Challenges may be targeted at specific standards statements or at the performance levels required for a particular category of achievement. A standard may be acceptable on its face but the specific way in which it is assessed or implemented may render the overall process unfair to some individuals or groups. When challengers perceive that direct challenges to objectionable standards may be infeasible or unsuccessful, other aspects of the assessment program may be targeted. Thus, the assessment itself may become a lightning rod for criticisms which originate in disagreement with state standards and policies.

Articulating Defensible Standards

A major issue specifically related to the articulation of goals or standards for high-stakes, large-scale assessments involves the wording of such statements. If students are to be held accountable, goals or standards must specify clearly observable behaviors. One must be careful not to set standards for which it is impossible to reliably determine whether students have met them. For example, an attitude such as appreciation of art or belief in diversity may be fakable if students know what they are expected to say. That is, students may give the "desired answer" while actually believing something entirely different. In such cases, it may be impossible to set up a performance situation which will capture the students' real attitudes. The Kentucky goal 3 (self-sufficiency) and goal 4 (responsible group membership), which are yet to be added to the assessment system, may suffer from the problem described above.

With respect to standards related to performance categories, Chapter 4 of this report has detailed the failure of Kentucky to follow professional standards in setting performance standards. In addition, the equating section describes how ad hoc procedures were used to convert an initial conjunctive standard to compensatory and to modify the correspondence between common item scores and performance categories (see Chapter 5). These ad hoc procedures also do not conform to accepted professional practice.

Legislative/Administrative Accountability Standards

As in Kentucky, it is common for final decisions on passing standards for statewide assessment programs to be made by legislators, the commissioner of education, and/or the board of education. This is desirable from a legal point of view because the decision is being made by those with the power to pass laws or administrative regulations and to appropriate the necessary funds to administer the assessment program. It is usually the case that the setting of passing standards by such policymakers is informed by standard setting studies which implemented methodology recommended by measurement professionals. While policymakers may consider additional data when establishing the final standard (e.g., estimated failure rates), this does not render their decisions arbitrary. Rather, as the appropriate authority for such decisions, these policymakers are fulfilling their responsibility to consider all the facts and circumstances, including recommendations from standard setting panels, in formulating the final accountability standard. However, basing such decisions on flawed procedures which fail to meet professional standards may not be defensible.

A.5.4 Use of Accountability Index For Distributing Rewards to Schools

Kentucky policymakers face a difficult choice. KIRIS rewards for the first biennium have been announced and distributed. The panel has found significant deviations from legal and professional standards in the KIRIS assessments and recommends significant changes in the operation of the assessment system. Legislators must now decide whether KIRIS should continue to be used for high-stakes school accountability decisions. Legislators and policymakers must consider the issues carefully and weigh the pros and cons of the various options open to them.

The goal of the KIRIS assessment system is to have all schools with 100% of students proficient within 20 years. However, this goal is unobtainable under the present system. If a school exactly achieves its goal each biennium for 10 biennia, a school which begins with an accountability index of 60 will end at 92.6; a school which begins at 30 will end at 82.1. No school reaches the goal unless it exceeds its target. Even for schools that are successful, the complexity of the accountability index and the magnitude of associated errors make it difficult to interpret what has actually been accomplished.

Errors of Measurement

"Differences between scores from . . . one occasion to another may be attributable to what is commonly called errors of measurement. . . . The magnitude of the error notwithstanding, the importance of a particular source of error depends on the specific use of a test" (*Test Standards*, 1985, p. 19).

Performance assessments and their associated standards may be vulnerable to challenge because their errors of measurement tend to be larger than for multiple-choice tests. There are two sources of error of particular concern for performance assessments: generalizability of the sampled tasks and rater effects. Even raters who are well-trained to apply previously agreed-upon standards may introduce errors due to irrelevant variables such as handwriting, appearance, speech, and culture. Generalizability of sampled tasks is of concern because a particular small set of tasks may not be a reliable indicator of student achievement of the domain from which they were sampled. These issues are considered in greater detail in Chapters 3 and 4. Courts are likely to

scrutinize procedures, the performance tasks, and scored student responses in determining what is good enough to withstand challenge.

Evaluating Error -- How much is too much?

"Fundamental to the proper evaluation of a test are the identification of major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected between pairs of scores under particular circumstances, and the generalizability of results across items, forms, raters, administrations, and other measurement facets. . . . It is essential, therefore, that the method used to estimate [errors] takes into account those sources of error of greatest concern for a particular use and interpretation of a test" (*Test Standards*, 1985, p. 19).

All assessment systems will contain some error. There is no perfect system which can completely eliminate all the potential sources of assessment error. In the Kentucky high-stakes accountability system where rewards accrue to schools, the error of primary concern is the misclassification of schools. Misclassification occurs when a school has achieved its targeted goal but fails to be rewarded OR when a school fails to achieve its targeted goal but receives a reward anyway. Put simply, misclassification occurs when errors in the system cause a school's accountability index to be higher or lower than it should be. Both of these possible misclassifications may place a school in the wrong category and result in state action which is contrary to the stated purpose of the accountability system.

As the panel has articulated in Chapter 3, the relevant question is whether the rate of misclassification is too high for policymakers to defend. In other words, are too many schools being given the wrong message (rewards they have not earned or lack of rewards they have earned)? Psychometricians may be able to estimate the magnitude of such misclassifications but determining where to draw the line between acceptable and unacceptable numbers of misclassifications is a policy decision.

Caution: While it may be possible to determine approximately how many schools have been misclassified, neither psychometricians nor policymakers can identify with certainty which schools have been misclassified. Therefore, policymakers can only decide in the abstract whether the total number of misclassifications is too high; it is impossible to identify and correct such errors for individual schools.

Gain Scores

> **Standard 2.1 Comment**
> [S]cores representing differences between scores obtained from two tests or from repeated administrations of the same test (called gain scores) are generally less reliable than either of [the tests alone]. (p. 20) **(Primary)**

One of the sources of error unique to the KIRIS accountability system derives from the use of school level gain scores. Gain scores are less reliable (accurate) than the scores used to calculate the gain.

The *Test Standards* (1985) state:

> **Standard 12.2**
> When change or growth scores are used in an evaluation, the definition of change or growth and the derived score that is used to measure it should be made explicit. These definitions should be explained in terms of how the definition of growth chosen, the particular question asked, and the underlying scales used to assess growth match. (p. 68) **(Primary)**
>
> **Standard 12.3**

> Gain scores should not be calculated when using tests that have been modified between administrations, unless scores on the modified test have been equated to scores on the original test. (p.68) **(Primary)**

Paraphrasing the Comment to Standard 1.3 in the *Test Standards* (1985, p. 14) to reflect the analogy to school accountability, "Gains in a school's accountability index from 1991-92 to 1993-94 should be found substantially more often in schools known to have instructional success in producing desired student outcomes than in other schools before such gains can be taken as indicative of having earned monetary rewards."

KDE has provided data to demonstrate that the total battery student reliability of KIRIS is similar to ACT and CTBS (*Why is the Accountability Index a Reasonable Basis for Making Decisions About Schools?*, handout from meeting with legislators, Lexington, KY, March 20, 1995). However, this comparison is misleading for two reasons. First, high-stakes decisions in Kentucky are based on gains in school accountability scores, not student total battery scores. Thus, the relevant reliability measure is school classification accuracy.

Second, assuming that individual student reliabilities were relevant, reliabilities should be reported at the level at which decisions are made. In KIRIS, no student is classified based on a total battery score; rather, students are classified as novice, apprentice, proficient or distinguished in each subject area separately. Thus, the relevant student level reliability estimate is at the subject level.

Separate subject reliabilities are given in the *Biennium I Technical Manual* (p. 224). These values range from .59 to .87 and are consistently lower than the .94-.95 reported in the KDE handout (p.2). The school level battery score reliabilities reported in the handout are also misleading because they are based on sampling that did not occur in KIRIS and because they exclude most of the significant sources of error identified in the next sections.

Cohort Variability

The issue of student cohort variability is considered in Chapter 3. The *Test Standards* provide the following guidance for school accountability programs:

> **Standard 12.7**
> Evaluations of [teachers] and administrators should not rest exclusively on the test scores of those people that they serve. (p. 69) **(Primary)**
>
> **Comment**
> Test scores of individuals served (e.g., students), will be affected by a great many factors not directly related to the quality of service they receive.

A concurring opinion in the *Rose* case observed that:

> We must keep in mind that the primary responsibility for the education of children is with the parents. The rights and responsibilities of the parent must always be recognized. The fractured fabric of the family is one of the prime causes for educational failure. Obviously money alone cannot heal such a break. The lack of scholastic success is not just the fault of the system. Education is a joint venture in which the parents, students and school must be committed to cooperation rather than conflict. (*Rose* at 218-19)

The use of gain scores to measure different cohorts of students across years includes the following effects:

1. effects due to differences in background variables between cohorts of students (e.g., ability, SES) and exogenous context variables (e.g., parent involvement, community support of education, unemployment rate) outside the control of schools;
2. effects due to school practice (e.g., administrative leadership, curricula, classroom instruction, use of resources) which represent true differences in school effectiveness and are within the control of schools; and
3. unstable effects due to unreliability of measurement and sampling errors.

Effects 1. and 3. may be large relative to 2., the effect the accountability index is intended to measure. Procedures called hierarchical linear models (HLM) can estimate effect 2. adjusting for effects 1. and 3. (see Willms & Raudenbush, 1989; Raudenbush & Willms, in press). In such models, adjusted pre-post outcome measures are highly correlated and the wild fluctuations of gain scores are markedly attenuated in the more stable estimated school practice effects. If such analyses were run on the Kentucky data, the estimated true school effects would be less variable and smaller than those reported and would provide a more accurate basis for judging school effects. Such data allow for recognition of quality teaching of disadvantaged students and forbearance of recognition to high-scoring schools with little value added.

Framework For Evaluating the Seriousness of Errors

Given the task of deciding how much error in the system is too much, how might policymakers proceed? One way to begin is by considering two types of errors: unavoidable and avoidable, together with the responsibilities of assessment developers with respect to each type of error.

Unavoidable errors are those errors which assessment developers can not completely eliminate. These include such areas as measurement error for individual students, sampling of tasks, and equating. Assessment developers have a duty to minimize such errors within existing budget and resource constraints. For example, task sampling errors decrease when the number of tasks increases. Assessment developers can minimize this error by administering the largest number of tasks feasible under constraints of testing time, development costs, scoring costs, and implementation timelines. An assessment developer would breach this duty only when resources were available to substantially increase the number of tasks and the developer failed to do so.

Avoidable errors are those that a prudent developer could eliminate with reasonable diligence and without undue resource burdens. In other words, given available resources, time constraints and legislative mandates, were reasonable alternatives available which, if implemented, would have substantially decreased errors? General examples of avoidable errors include use of ad hoc procedures when cost-effective psychometric procedures are available, failure to use available data optimally, and lack of adequate assessment security in a high-stakes environment. Specific areas in which avoidable errors might have occurred in the Kentucky accountability system are detailed in Chapters 2 - 8 and other sections of this appendix. When assessments are used for high-stakes purposes, assessment developers have a duty to eliminate avoidable errors.

To the extent that unavoidable errors were not reasonably minimized and/or avoidable errors were not eliminated, policymakers must decide if the potential unfairness created by these larger than necessary errors renders the decisions made by the accountability system unconscionable and therefore unenforceable on public policy grounds. But in making this decision, policymakers must also determine whether the beneficial effects of the assessment system outweigh any potential unfairness attributable to errors in determining rewards and sanctions. Examples of beneficial effects in the Kentucky accountability system include the professional development of large numbers of teachers who participated in training for portfolio scoring and the implementation of desired curricular changes in anticipation of receiving rewards.

The areas of concern identified by the panel focus on the minimization of unavoidable errors and the elimination of potentially avoidable errors. However, identification of the possible existence of such errors alone does not render the accountability system unconscionably unfair. Further evidence of the magnitude of such errors is needed to determine whether they are substantial enough to matter. To the extent that such errors were trivial, they would not suggest substantial unfairness. In such a case, the beneficial effects of the accountability system may outweigh the minimal unfairness created by these sources of error.

Consideration of Overall Error

Policymakers must still decide if the overall error rate (avoidable plus unavoidable errors) in the accountability system is acceptable. The components of the overall error rate which the panel considered include the following:

a. standard error of measurement for individual students;
b. sampling error due to cross-sectional data, selection of students to participate in performance events, and small schools;
c. scoring error due to rater effects;
d. errors due to lack of task generalizability;
e. equating errors;
f. errors related to assessment security, nonstandard administration conditions, or inappropriate assessment preparation activities;
g. errors due to the use of gain scores; and
h. classification errors.

Each of these potential sources of error can be examined independently. Some can be combined into a single estimate. For example, generalizability coefficients can estimate the combined effects of tasks, raters, and student sampling. However, given present data availability, time constraints, and the state-of-the-art, there is no single psychometric index which can estimate the overall error from all sources combined. This is unfortunate because policymakers need an overall estimate to determine whether the misclassification rate for the existing accountability index is acceptable.

In summary, as they consider whether to continue distributing rewards under the existing accountability system, policymakers may ask the following questions:

(1) Does the accountability system contain avoidable errors which are substantial enough to render the system unconscionably unfair?

(2) Is the overall error rate (estimated number of misclassified schools) acceptable?

According to a recent article in the *Herald Leader* (May 1995):

* 16 of the districts involved in the *Rose* lawsuit earned rewards in the first biennium;

* he top scoring school district was among the 15 poorest in the state; and

* half of the 42 districts receiving rewards rank in the bottom half of all districts in property wealth.

Low property wealth is generally correlated with low socioeconomic status, including minimal formal education of parents, a sizable percent of students on free/reduced lunch, family income below the poverty level, etc. Such variables are generally associated with less knowledge/skills gained outside school, less motivation for homework outside school, skipping breakfast, inadequate sleep and the like. A school with all this to overcome makes fantastic progress when it achieves high reward status in just two years. Before a successful school could

teach the new higher order skills, it would have to alleviate current deficiencies in the prerequisite knowledge and skills of its students. This can be achieved but it usually takes a long time and a lot of sustained work and effort by both teachers and students.

Such dramatic short-term gains by schools with the most "catching up" to do make the Kentucky results less believable and make it more probable that errors had a significant effect on the results. To the extent the performance of poor schools reflects improved curricula and teaching, these same poor schools should be able to earn rewards again in Biennium II. To the extent that these results reflect sandbagging during the baseline year or positive errors of measurement, one might expect different schools to meet the reward criteria in Biennium II.

The *Rose* decision declaring the Kentucky education system unconstitutional addressed the issue of equal opportunity for all students; it did not mandate equal outcomes. To the extent that the apparent equality of poor and rich schools in earning rewards is a function of error in the accountability system, what appears to be equal outcomes may mask real inequalities in educational opportunity.

A report by The Assessment Team of Consultants to the Kentucky Board of Elementary and Secondary Education which reviewed KIRIS contractor bids and made recommendations to the Board stated:

> The KERA should be amended such that judgments about high stakes rewards and sanctions will not be based on 1% and 5% deviations from thresholds. Basing accurate judgments upon such deviations exceeds the limits of technical possibility. Even if the measurement procedures are perfectly valid and comparison measures used two years apart are perfectly equivalent, changes based on such small percentages will cause schools to be incorrectly classified. We recommend that the Board work with the contractor to develop a technical amendment to the KERA, making changes in school status more robust against changes due only to population shifts and measurement errors.
>
> . . .
>
> Advanced Systems . . . addresses the issue of detecting "true" score change and points out the statistical limitations of detecting such changes. . . . Advanced Systems provides a well-developed discussion on longitudinal and cross-sectional analyses for school evaluation. (Redfield, Kifer, Millman, Rindone, & Wiggins, *Recommendations on Proposals to Implement an Interim and Full-Scale Assessment Program for the Commonwealth of Kentucky*, June 27, 1991, p. 1, 5)

The response to the Consultants' Report and the information provided by Advanced Systems appears to have been to increase the threshold to a 10% gain. If 10% is close to the overall error rate in the system, this action was not successful in addressing the fundamental issue identified by these sources prior to the implementation of KIRIS.

A.5.5 Apportionment of Rewards

A teacher who left a school prior to the distribution of rewards has sued for her share. The remaining teachers in the school voted to distribute the reward money as a bonus to each teacher but chose not to include teachers who had left. The extra money was used for bonuses for other school staff and for a scholarship fund (see Gibson, 1995). There are bad feelings at other schools where the teachers have voted to give all the reward money to themselves and to not include other school staff such as cafeteria workers or custodians.

The above examples demonstrate that the guidelines for distribution of reward money contain ambiguities and have created added tensions among past and present school staff. Unfortunately, sorting this out in the legal arena will probably consume significant state and district resources.

A.5.6 Appeals

The Kentucky Revised Statutes provide for an appeals process for the KIRIS accountability system. Section 158.6455 (8) states:

> The [Board] shall adopt administrative regulations to establish a process whereby a school shall be allowed to appeal a performance judgment which it considers grossly unfair. The state board may adjust a performance judgment on appeal when evidence of highly unusual circumstances warrants the conclusion that the performance judgment is based on fraud or a mistake in computations, is arbitrary, is lacking any reasonable basis, or when there are significant new circumstances occurring during the biennial assessment period which are beyond the control of the school.

KDE outlined an appeals process for the first biennium in a handout distributed at the District Assessment Coordinators meeting in March 1995. The steps included:

1. Letter of appeal submitted to Commissioner;
2. KDE submits request with recommendation to EIAC;
3. EIAC submits recommendation to appealing party and Commissioner;
4. Within 30 working days, Commissioner submits recommendation to KY State Board of Education and appealing party;
5. Appealing party has 10 working days to submit written responses to state board;
6. The KY State Board of Education makes final determination regarding the appeal.

Twenty-two schools appealed their results for Biennium I; 14% were granted and 86% denied (Summary Appeals of Accountability Cycle I Performance Judgments, March 23, 1995). Of the 19 denials, 5 schools were given an "uncategorized" performance judgment. Appeals that were granted involved a new school with no 1991-92 baseline, a school which changed from 4th & 8th to 4th only, and an error by AS regarding provision of equipment for social studies performance events.

Out of the 22 appeals for Biennium I, 8 involved changes in student population believed to have adversely impacted the results; some schools wanted to retain their designated educators while others requested adjustments that would have raised their scores. Eight other appeals involved arguments that the portfolio or performance events scores were not representative of the school and should be rescored or discarded. Some schools also cited an inconsistency between transitional test classifications and those from performance events or portfolio scores.

The appeals policy does not provide for a due process hearing which may be legally required in some cases. The issues are whether schools' appeal letters and responses provide an adequate opportunity to be heard and whether the Board is making its decisions fairly and impartially based on adequate evidence.

A.5.7 Other Due Process Requirements

Previous sections have highlighted the application of the notice and curricular validity due process requirements from the *Debra P.* case as modified for high-stakes school accountability. But the *Debra P.* court also indicated that the substantive due process requirement that assessments not be arbitrary, capricious or unfair was met only when available evidence demonstrated that the assessments met all applicable professional standards.

In prior chapters, the panel has described the ways in which KIRIS has failed to meet applicable professional standards in the areas of validity, reliability, standard setting, portfolios, equating, and reporting assessment results. The

shortcomings of KIRIS detailed in these prior chapters contribute to potential due process challenges to the use of KIRIS for school accountability.

A.6 Adverse Impact

The equal protection clause of the fourteenth amendment to the U.S. Constitution provides that "[n]o state shall . . . deny to any person within its jurisdiction the equal protection of the laws." To demonstrate an equal protection violation, federal courts require evidence that one of two similarly situated groups was treated differentially by a state or local government which intended to discriminate (Nowak et al., 1986).

Adverse impact occurs when a government policy results in more frequent denial of a property right to one group than another. That is, there is adverse impact when substantially more historically disadvantaged students fail a graduation test than nondisadvantaged students. Adverse impact is a necessary but not sufficient condition for establishing intent to discriminate. To evaluate intent to discriminate, federal courts also consider all the facts and circumstances surrounding the challenged state or local government action (*Village of Arlington Heights v. Metropolitan Housing Dev. Corp.*, 1977; *Personnel Administrator v. Feeney*, 1979).[5]

The *Debra P.* case focused specifically on the initial adverse impact of the Florida test on African-American students. However, after four years of litigation, the gap between the performance of African-American and Caucasian students had narrowed from a difference of approximately 18% to a difference of less than 10% with remediation and retesting. In general, in the last decade, the gap between nondisadvantaged and historically disadvantaged performance on multiple-choice tests has narrowed. After remediation and multiple retakes, historically disadvantaged passing rates on graduation tests have been relatively high and much closer to nondisadvantaged passing rates than when such tests were first adopted (*Debra P. v. Turlington*, 1979-83; Phillips, March 11, 1993).

A.6.1 Changes in Adverse Impact

With new standards and performance assessments, differential passing rates will again receive careful scrutiny. Although differential performance by itself is not sufficient to invalidate an assessment program, defending such a program against a legal challenge based on alleged discrimination can be costly, time-consuming, and detrimental to public relations. It also can exacerbate test security concerns as challengers seek access to the disputed assessment tasks.

Because performance tasks tend to measure complex sets of skills which combine reading and writing with subject-specific knowledge, students who are disadvantaged may perform less well. Initial data suggest increased adverse impact of new standards and performance assessments on historically disadvantaged students (Beck, in press; Mehrens, 1992). If so, opponents of such assessment programs may argue intent to discriminate by replacing a less discriminatory alternative (multiple-choice tests) with a system that results in greater disadvantage. Employment testing cases have held that cost effective alternatives with less adverse impact must be considered (*Wards Cove Packing Co. v. Antonio*, 1989). Critics might argue by analogy that such a requirement should also apply to educational assessment. Further discussion of legal and psychometric issues surrounding procedures for minimizing differential item and assessment performance can be found in Phillips (Dec. 20, 1990; 1993).

A.6.2 KIRIS Data

Legally, adverse impact occurs when scores for majority students are higher than for minority students. However, the presence of adverse impact is not synonymous with "bias." Adverse impact indicates that further investigation for potential item or assessment bias is warranted. Professional judgment based on additional evidence may determine that no discrimination has occurred despite the presence of adverse impact.

In Kentucky, there are two major disadvantaged groups of concern to educators: Appalachian Caucasians and African-Americans. Legal challenges based on adverse impact are typically successful only when the strictest standard of judicial review applies. Based on prior Supreme Court decisions, African-Americans qualify for the strictest standard of judicial review; Appalachian Caucasians do not. Kentucky may be interested in tracking the performance of both groups although legal challenges based on African-American performance are most likely to occur and have the greatest potential for success.

Chapter 6 on Item Analysis in the Biennium I Technical Manual does not include any data or descriptions of differential item performance, although lists of tasks presented in the test development section indicate an intent to calculate such statistics for both multiple-choice and open-ended items (p. 57-59). At a minimum, such statistics should be calculated and reported for Caucasians and African-Americans.

The manual does indicate that committees conducted content and bias reviews for all items. The specific criteria for such reviews is unknown. Minority membership on content advisory committees ranged from 2 to 4 members out of 18 to 24 across subject areas (Biennium I Technical Manual, 1995, Tables 37-38, p. 53-54, numbers handwritten in on copies provided by KDE). No information is available regarding which racial/ethnic groups are represented. The demographic composition of the bias review committee is not presented in the Manual.

The table below summarizes racial/ethnic comparisons for the 1993-94 KIRIS common items.[6]

	Open-ended (OE)			Multiple-choice (MC)			TOTAL SD
Grade/Subject	Caucasian	African-American	Diff	Caucasian	African-American	Diff	OE/MC
12 Reading	8.8	7.2	1.6	13.4	11.2	2.2	3.2/3.8
Mathematics	5.8	3.9	1.9	9.2	7.2	2.0	3.3/3.8
Science	7.4	5.1	2.3	10.2	7.9	2.3	3.5/3.9
Social Studies	9.0	7.2	1.8	11.1	9.5	1.6	3.8/3.8
8 Reading	11.0	9.1	1.9	13.4	11.3	2.1	3.7/4.3
Mathematics	7.6	4.9	2.7	9.8	7.2	2.6	4.5/4.2
Science	6.1	4.4	1.7	9.9	8.0	1.9	3.3/3.2
Social Studies	8.5	7.0	1.5	11.7	9.8	1.9	3.5/4.0
4 Reading	9.6	8.2	1.4	14.3	11.6	2.7	3.3/4.3
Mathematics	8.8	6.2	2.6	12.2	9.9	2.3	4.5/4.0
Science	10.0	7.9	2.1	11.5	9.2	2.3	3.7/3.6
Social Studies	8.9	7.5	1.4	12.2	9.9	2.3	3.4/3.8

For 1993-94, Caucasians consistently outperformed African-Americans in all grades and all subjects for both the open-ended and multiple-choice common items. Two-thirds of the time, the multiple-choice difference was greater than that for open-ended items, primarily in reading, science and social studies. Caucasian/African-American differences ranged from 1.4 to 2.7 on scales with overall standard deviations of about 4.0. For grades 12, 8, and 4, respectively, the largest differences occurred in science, mathematics and mathematics for the open-ended items and in science, mathematics and reading for the multiple-choice items.

Clearly there was adverse impact for African-Americans on the open-ended common items which were included in the calculation of the accountability index. However, this does not establish discrimination unless it can also be shown that there was intent to discriminate by providing an inadequate opportunity for African-Americans to learn the assessed content relative to that provided for Caucasians, or that the assessment content was chosen to deliberately disadvantage schools serving minority students. There is no evidence of either of these contingencies.

The 1991-92 data showed similar patterns (see 1991-92 Technical Manual, p. 68-9). Caucasian students outscored African-American students on both the open-ended and multiple-choice items for all grades and subjects. In 67% of the comparisons, the difference between the two groups on the multiple-choice items was greater than for the open-ended items. This pattern occurred at grade 4 and for reading in both higher grades; results were mixed in the other subjects at the two higher grades.

Overall, Caucasians outscored African-Americans by a smaller amount on open-ended than on multiple-choice items although a third of the time the difference went the other way. For the 1993-94 data, except for reading, the diminution in the Caucasian/African-American difference for the open-ended items was relatively small and Caucasians still outscored African-Americans by 1.5 to 2.5 points on average. To the extent these differences in performance occurred near cut points, they could have adversely affected African-American performance classifications.

No data is available for comparing the effects of these differences on classification rates for novice, apprentice, proficient and distinguished. But if one had to be proficient to pass the assessment and if the differential resulted in significantly fewer African-Americans classified as proficient, there would be adverse impact. For school accountability, the tendency for African-Americans to be classified less often as proficient would also be evidence of adverse impact on schools serving minority students.

Recall, however, that the presence of adverse impact alone is not sufficient to establish discrimination. There is no evidence of differences in opportunity to learn or assessment practices deliberately designed to discriminate. These data merely demonstrate that adverse impact should be monitored consistently across years. In addition, further analyses should be conducted on classification rates and on adverse impact for matrix items, performance events, and portfolio measures.

For high-stakes school accountability systems, the major question of interest is whether schools that serve disadvantaged students are adversely affected. In Kentucky, the major racial/ethnic groups are Caucasian and African-American. KDE data indicates that distress as measured by the percent of students on free/reduced priced lunches was positively correlated with baseline accountability scores (.68) but unrelated to achievement of reward status. (.02; Biennium I Technical Manual, 1995, p. 267-69). KDE interprets this finding as evidence of attainment of the equity mandated by the *Rose* decision but also states it is "impressive given that the schools started off unequally" (p. 267). An alternative explanation is that the overall error in the gain scores used to measure success created random variability which resulted in the low correlation.

Correlations between the percent of minority students (a variable overlapping with free/reduced lunches in some districts) and reward status was weak (-.12; Biennium I Technical Manual, 1995, p. 271). However, over half of Kentucky schools have fewer than 10% minority students.

To provide some evidence of differences at the school level, the accountability indices across years for 6 districts among the top in the state in minority enrollments are presented in the table below. Corresponding values for the state are given in the last row of the table.[7]

District	% Min	Goal	'91	'92	'93	'92-'94	Status	#Schl	#Rew
Christian Co	37	(40.2)	33.5	34.9	40.9	37.8	Success Year 2	16	4
Fayette Co	25	(44.9)	38.8	42.7	49.4	46.1	Reward	48	19
Fulton Ind	37	(41.4)	34.9	40.1	40.2	40.2	Improving	2	0
Jefferson Co	33	(42.9)	36.1	38.1	43.6	40.9	Success Year 2	133	39
Owensboro Ind	18	(43.6)	37.3	40.2	45.3	42.8	Success Year 2	8	3
Paducah Ind	44	(44.6)	38.4	40.5	42.8	41.7	Improving	6	1
STATE			36.4	39.4	44.8	42.1	Gain = 5.7		

Three of the six districts with large minority enrollments had 1991-92 baseline accountability indices above the state average. Two of these three districts also exceeded the state average for each of the two subsequent years of Biennium I while two of the remaining three districts remained consistently below the state average. Overall, for Biennium I, one district became eligible for reward while the other five failed to meet their thresholds. Thus, districts with large minority enrollments appeared to have difficulty qualifying for rewards in Biennium I.

The table also indicates the number of schools in each district and the number receiving rewards. Zero to 40% of the schools in these districts received rewards. Nine of the 55 in decline schools came from these districts. The numbers of rewarded schools with minority enrollments less than the district average were: Christian Co 1/4; Fayette Co 14/19; Fulton Ind 0/0; Jefferson Co 25/39; Owensboro Ind 1/3; and Paducah Ind 1/1. Overall, 64% of rewarded schools from these districts had minority enrollments less than their district percentages. The extent to which rewarded schools within these districts with large minority enrollments differed demographically from the nonrewarded schools has not been documented. Similarly, the ways in which the in decline schools identified in these districts may differ demographically from the others has not been documented.

A.7 Opportunity For Success

The *Test Standards* state:

> **Standard 3.5**
> When selecting the type and content of items for tests . . ., test developers should consider the content and type in relation to cultural backgrounds and prior experiences of the variety of ethnic, cultural, age, and gender groups represented in the intended population of test takers. (p. 26) **(Conditional)**
>
> **Standard 3.23**
> When structured behavior samples are collected within a standardized testing format, the specific type of behavior expected should be defined clearly. Directions to a test taker that are intended to produce a particular behavior sample (often called a "prompt") should be standardized, just as the directions are standardized for any other test. . . . (p. 30) **(Conditional)**
>
> **Standard 8.1**
> Those responsible for school testing programs should ensure that the individuals who administer the tests are properly instructed in the appropriate test administration procedures and that they understand the importance of adhering to the directions for administration that are provided by the test developer. (p. 52) **(Primary)**

In addition to protecting the rights of historically disadvantaged groups, the courts have also recognized a more general fundamental fairness requirement that applies to all students. This requirement is part of the substantive due process

clause of the fourteenth amendment which was described earlier. According to this requirement, assessments must not be arbitrary or capricious and must provide all students with conditions fostering an equal chance for success. Note that this is not a guarantee of equal outcomes but rather of standardized conditions which ensure that no student receives an unfair advantage or penalty. This concept has been referred to as ***opportunity for success*** (Phillips, 1994).

The *Code of Fair Testing Practices* (1988) states:

> Test Developers or Test Users Should: . . . 18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. ***Strive to make such information equally available to all test takers.*** (p. 4, emphasis added).

The *Test Standards* recommend:

> [T]est administrators should follow carefully the standardized procedures for administration and scoring specified by the test publisher. Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. (p. 83).

The potential for violation of the opportunity for success requirement is of particular concern in the administration of high-stakes performance assessments. The reason is that it is more difficult to maintain standardized assessment administration conditions when students work in groups, manipulate equipment or engage in related activities designed to precede the actual assessment.

The following examples illustrate some of these issues. Each example illustrates the potential for unfair advantage or unfair penalties to some students but not others.

Suppose the assessment requires students to measure their classroom and calculate the cost of wall-to-wall carpeting. Suppose further that in most classrooms the students are given yardsticks for measuring but that one classroom does not have yardsticks available so the teacher gives each student a ruler. Although it is possible to complete the task with a ruler, the task is more difficult and measurement with a ruler may take more time. Measurements made with inefficient measuring instruments are also more prone to error. Thus, failing students who were forced to use a ruler rather than a yardstick could claim unfair treatment because they were required to complete a harder task than the other students. Similar arguments can be made in any situation in which equipment used for a high-stakes assessment task differs from classroom to classroom or school to school.

A related issue involving assessment task equipment is the training students have received in using the equipment. For example, a student once had a calculator that featured reverse polish notation. By entering a series of operations in reverse order, certain parentheses could be eliminated. Though cumbersome at first, once learned this feature permitted faster calculation and eliminated errors due to forgetting to enter a last parenthesis. However, facility on this type of calculator made it difficult to remember and revert back to the procedures necessary to use a conventional calculator. If the student had needed to use a conventional calculator on a high-stakes assessment, the student might have become frustrated without some prior practice. This frustration might have negatively impacted performance. Similarly, any equipment which students may be required to use on an assessment and which is different from what they normally use may pose unfair difficulties due to unfamiliarity and lack of practice. Some schools which furnish equipment for an assessment may lack the resources to purchase products of higher quality or having more optional features.

A possible solution to the above issue would be to allow students to use their own equipment. But this alternative also has standardization problems which

might lead to unfairness. For example, suppose the students in some schools can afford to purchase programmable calculators while others can only afford a simple, four-function calculator. Alternatively, a large-scale assessment program might address this issue by furnishing equipment and monitoring training. But in doing so, the program will incur substantial expense. And even then there may be unfairness if some students regularly use calculators they own but others have only one week of practice on calculators furnished specifically for the assessment.

Another opportunity for success issue of concern with some performance assessments is the use of cooperative learning. Although cooperative learning may be an effective teaching technique, it may interact in unpredictable ways with opportunity for success by introducing potential unfairness. For example, suppose a group of four students jointly produces a written essay or completes a science experiment. Suppose further that one academically talented student dominates the process while one weak student contributes little. Should all four students receive the same score? If not, what criteria should be used to judge students' relative contributions to the final product? What if the best writer in the school went off in a corner and produced an excellent portfolio without interacting with anyone else? Would that student fail because the process specified by the assessment had not been followed?

Unfairness can also occur when portions of a high-stakes assessment are completed outside of class. For example, a student might be asked to submit a piece of outside writing for evaluation. Will poor schools be penalized because there are no parents available to help students with their writing projects? Will students in more advantaged schools submit writing products heavily influenced by input from siblings, friends or tutors? Will some schools benefit because their students' parents provide them access to desktop publishing programs and laser printers which can improve the appearance of the final product and provide illustrative graphics?

Finally, opportunity for success can be denied by differences in a variety of factors under the control of the assessment administrator. Students in different schools might be under different time pressures to complete the task due to differences in scheduling, teacher preferences or other factors. Some administrators who facilitate pre-task activities or discussions may cover different information or provide different kinds of assistance in completing the task. Some students may ask questions which result in additional assistance not available to other students. When group work is required, some students may be paired with other students with whom they have difficulty working. Administration of performance events by outside evaluators may be inconsistent across schools.

These and a host of other factors may result in assessment experiences which are significantly different for two schools whose students have similar knowledge and skills. If one school receives a reward and the other does not, the differential assessment conditions may be cited as a significant contributing factor. Although one might believe the latter school would have been unsuccessful anyway, it may be difficult to convince a judge that all schools had an equal opportunity for success on the assessment.

To the extent that specific inequities can be demonstrated, such high-stakes assessments may be judged unfair under the substantive due process standard. To avoid differential opportunities for success in a high-stakes assessment, developers must carefully consider alternative ways to structure the task so that the students in all schools respond under equivalent conditions and so that potential inequities are eliminated to the fullest extent possible.

A.7.1 Application to KIRIS

An assessment irregularity incident investigated by the Board of Review (Owen memo, Jan. 19, 1995) confirmed that a student had used a reference book to answer a science item. In another incident, the Board confirmed that a student had

might lead to unfairness. For example, suppose the students in some schools can afford to purchase programmable calculators while others can only afford a simple, four-function calculator. Alternatively, a large-scale assessment program might address this issue by furnishing equipment and monitoring training. But in doing so, the program will incur substantial expense. And even then there may be unfairness if some students regularly use calculators they own but others have only one week of practice on calculators furnished specifically for the assessment.

Another opportunity for success issue of concern with some performance assessments is the use of cooperative learning. Although cooperative learning may be an effective teaching technique, it may interact in unpredictable ways with opportunity for success by introducing potential unfairness. For example, suppose a group of four students jointly produces a written essay or completes a science experiment. Suppose further that one academically talented student dominates the process while one weak student contributes little. Should all four students receive the same score? If not, what criteria should be used to judge students' relative contributions to the final product? What if the best writer in the school went off in a corner and produced an excellent portfolio without interacting with anyone else? Would that student fail because the process specified by the assessment had not been followed?

Unfairness can also occur when portions of a high-stakes assessment are completed outside of class. For example, a student might be asked to submit a piece of outside writing for evaluation. Will poor schools be penalized because there are no parents available to help students with their writing projects? Will students in more advantaged schools submit writing products heavily influenced by input from siblings, friends or tutors? Will some schools benefit because their students' parents provide them access to desktop publishing programs and laser printers which can improve the appearance of the final product and provide illustrative graphics?

Finally, opportunity for success can be denied by differences in a variety of factors under the control of the assessment administrator. Students in different schools might be under different time pressures to complete the task due to differences in scheduling, teacher preferences or other factors. Some administrators who facilitate pre-task activities or discussions may cover different information or provide different kinds of assistance in completing the task. Some students may ask questions which result in additional assistance not available to other students. When group work is required, some students may be paired with other students with whom they have difficulty working. Administration of performance events by outside evaluators may be inconsistent across schools.

These and a host of other factors may result in assessment experiences which are significantly different for two schools whose students have similar knowledge and skills. If one school receives a reward and the other does not, the differential assessment conditions may be cited as a significant contributing factor. Although one might believe the latter school would have been unsuccessful anyway, it may be difficult to convince a judge that all schools had an equal opportunity for success on the assessment.

To the extent that specific inequities can be demonstrated, such high-stakes assessments may be judged unfair under the substantive due process standard. To avoid differential opportunities for success in a high-stakes assessment, developers must carefully consider alternative ways to structure the task so that the students in all schools respond under equivalent conditions and so that potential inequities are eliminated to the fullest extent possible.

A.7.1 Application to KIRIS

An assessment irregularity incident investigated by the Board of Review (Owen memo, Jan. 19, 1995) confirmed that a student had used a reference book to answer a science item. In another incident, the Board confirmed that a student had

clause of the fourteenth amendment which was described earlier. According to this requirement, assessments must not be arbitrary or capricious and must provide all students with conditions fostering an equal chance for success. Note that this is not a guarantee of equal outcomes but rather of standardized conditions which ensure that no student receives an unfair advantage or penalty. This concept has been referred to as ***opportunity for success*** (Phillips, 1994).

The *Code of Fair Testing Practices* (1988) states:

> Test Developers or Test Users Should: . . . 18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. ***Strive to make such information equally available to all test takers.*** (p. 4, emphasis added).

The *Test Standards* recommend:

> [T]est administrators should follow carefully the standardized procedures for administration and scoring specified by the test publisher. Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. (p. 83).

The potential for violation of the opportunity for success requirement is of particular concern in the administration of high-stakes performance assessments. The reason is that it is more difficult to maintain standardized assessment administration conditions when students work in groups, manipulate equipment or engage in related activities designed to precede the actual assessment.

The following examples illustrate some of these issues. Each example illustrates the potential for unfair advantage or unfair penalties to some students but not others.

Suppose the assessment requires students to measure their classroom and calculate the cost of wall-to-wall carpeting. Suppose further that in most classrooms the students are given yardsticks for measuring but that one classroom does not have yardsticks available so the teacher gives each student a ruler. Although it is possible to complete the task with a ruler, the task is more difficult and measurement with a ruler may take more time. Measurements made with inefficient measuring instruments are also more prone to error. Thus, failing students who were forced to use a ruler rather than a yardstick could claim unfair treatment because they were required to complete a harder task than the other students. Similar arguments can be made in any situation in which equipment used for a high-stakes assessment task differs from classroom to classroom or school to school.

A related issue involving assessment task equipment is the training students have received in using the equipment. For example, a student once had a calculator that featured reverse polish notation. By entering a series of operations in reverse order, certain parentheses could be eliminated. Though cumbersome at first, once learned this feature permitted faster calculation and eliminated errors due to forgetting to enter a last parenthesis. However, facility on this type of calculator made it difficult to remember and revert back to the procedures necessary to use a conventional calculator. If the student had needed to use a conventional calculator on a high-stakes assessment, the student might have become frustrated without some prior practice. This frustration might have negatively impacted performance. Similarly, any equipment which students may be required to use on an assessment and which is different from what they normally use may pose unfair difficulties due to unfamiliarity and lack of practice. Some schools which furnish equipment for an assessment may lack the resources to purchase products of higher quality or having more optional features.

A possible solution to the above issue would be to allow students to use their own equipment. But this alternative also has standardization problems which

copied a social studies response from an encyclopedia. These efforts were obviously clumsy and readily detected. In yet another incident, small groups of students administered reading and writing open-ended tasks by the same instructor had identical responses. Such incidents, in the absence of a written policy on assessment security and a failure to systematically check for security violations, leads one to wonder how many other students may have received unauthorized assistance in completing their responses. One also wonders why these incidents were not detected by assessment administrators. Did some teachers believe outside sources of information could be utilized? Or did they believe they would not be caught if they looked the other way when their students did so? Did some teachers suggest what to write or give students hints?

Clearly, students with access to reference materials during administration of the assessment would have an unfair advantage over those students who did not have such access. Consequently, schools with procedures lax enough to permit such access would have an unfair advantage over those schools that did not. Due to the high-stakes for schools and lack of random, unannounced auditing by KDE, there was an incentive to gain an advantage by ignoring the issue or encouraging teachers to allow whatever assistance they deemed appropriate. In the absence of clear and specific guidelines from KDE, it could reasonably be expected that school and teacher interpretations across the state might differ significantly.

The confounding of social and academic skills discussed earlier also has implications for OFS. A school's score will partly reflect the degree to which the school is successful in mixing students of different abilities and in arranging for the more able students to assist the less able. Related concerns regarding the preparation and scoring of portfolios are described in detail in Chapter 4.

A two-page description of appropriate assessment practices is contained in Appendix C of the 1992-93 KIRIS Technical Report. The document describes two general standards, professional ethics and educational defensibility, from an article by Popham (1991). Interpretation and specific application of these standards to KIRIS is left largely to the reader.

A form at the bottom of the page must be signed by all school personnel involved with KIRIS. The form certifies that the signatory has read the assessment security document. Local districts are responsible for keeping the signed forms on file. Although the document indicates actual test items are not to be photocopied or used in instruction, apparently one school did just that for the 1994-95 KIRIS. Clearly the procedures described above were not adequate for guaranteeing equal opportunity for success to all schools. These procedures also did not adequately protect assessment security.

A.8 Assessment Security

Standard 15.3
Reasonable efforts should be made to assure the validity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means. (p. 83) **(Primary)**

Standard 15.7
Test users should protect the security of test materials. (p. 84) **(Primary)**
Comment
Those who have test materials under their control should take all steps necessary to assure that only individuals with a legitimate need for access to test materials are able to obtain such access. (p. 83-4)

Standard 16.7
Under certain conditions it may be desirable to cancel a test taker's score or to withhold it because of possible testing irregularities, including suspected misconduct. The type of

> evidence and procedures to be used to determine that a score should be canceled or withheld should be explained fully to [those] whose scores are being withheld or canceled. (p. 86) (**Primary**)

As the above Standards indicate, measurement experts recommend tightly controlled test administration conditions for high-stakes assessments. A high-stakes assessment is one which is used for accountability, including the award of diplomas to individual students or monetary awards to districts, schools or teachers. Administrators of high-stakes assessments have a duty to ensure, insofar as possible, uniform and equitable assessment conditions for all students and all schools. Included within this responsibility is the maintenance of secure assessment materials, adherence to specified directions for test administration, adequate proctoring during test administration, accounting for all test booklets and answer folders, refraining from inappropriate test preparation activities or unauthorized assistance, and maintenance at all times of a climate of integrity and ethical conduct devoid of any hint or appearance of a violation of these duties.

It is an unfortunate fact that accountability through high-stakes assessment provides an incentive for unethical behavior. Past experience has demonstrated that some educators will engage in inappropriate assessment activities when the stakes are high. Teachers and administrators have been caught practicing with actual test items, practicing with similar items, erasing and changing student answers, pointing to incorrect answers while students were taking the test, advising low achieving students to stay home the day of testing, teaching only the content sampled on the test rather than the total curriculum the students were supposed to learn, giving a parallel form of the test just prior to scheduled statewide testing, and supplying answers to students (see *Schools For Scandal*, 1992; Mehrens & Kaminski, 1989).

Some critics of testing have even gone so far as to suggest that when educators disagree with the purpose or content of assessments, they are justified in disregarding assessment security procedures. Linn (1983) observed: "For some of those who are philosophically opposed to a [basic skills test] requirement for graduation, cheating may be seen as a case of the ends justifying the means" (p. 127). Beck (in press), Mehrens, Phillips & Schram (1993), Phillips (1994a; March 11, 1993), Popham (1992; 1991), Mehrens (1992), and Mehrens & Kaminski (1989) strongly refute such justifications for unethical behavior while acknowledging that pressures to do so are strong. Their conclusions suggest that states implementing high-stakes assessments must be vigilant and proactive in minimizing such activity in order to promote fairness for all participants. It is naive and irresponsible to ignore the potential for unethical behavior in a high-stakes environment.

In response to the reality of such unethical practices, most high-stakes statewide assessment programs have instituted a variety of policies to deter unethical behavior. One of the most significant deterrents has been the development of sophisticated assessment security procedures together with strict policies for sanctioning violators. Phillips (1993) provides the following recommendations for assessment security in high-stakes applications:

1. Ship test booklets so they arrive only a few days before testing. Require a responsible administrator sign a form acknowledging receipt and assuring that the materials will remain locked in a storage area with very limited access.

2. Allow only the minimum necessary time for testing and require all sites to test on the same day(s).

3. Require all testing materials to be returned immediately after testing.

4. Seal and number all test booklets and shrink wrap bundles of test booklets.

5. Require written assurance from test administrators at each site that test booklets were opened only by examinees when told to do so during testing and that no booklets were photocopied.

6. Require test administrators to account for all testing materials before examinees are allowed to leave the room for lunch breaks or at the conclusion of testing.

7. Arrange for multiple-proctors in each testing room and allow only one student at a time to leave during testing.

8. Have all test administrators keep records of irregularities at the test site.

9. Investigate all reports of breaches of test security and sanction those involved in confirmed incidents.

10. Randomly audit test sites unannounced to ensure proper procedures are being followed.

11. Request that the legislature enact a statute or the state board adopt an administrative rule defining and providing sanctions for individual educators who engage in inappropriate test preparation activities and cheating.

12. Examine answer documents for tampering, excessive erasures, copying, and other signs of cheating. Screen group statistics and repeat testers for unusually large performance gains. Use suspicious findings to trigger appropriate investigations. (p. 31-2)

Enforceable test security procedures require a written policy which conforms to recommended measurement practice, notice to all affected educators of the policy and penalties for violations, and consistent application of the policy and its associated penalties. The courts normally assume notice of codified laws whether or not the violator has actually ever seen or read them. However, compliance is more likely when a written policy has been distributed and schools are on notice of penalties for violations.

When penalties have been specified a priori, it is reasonable for such penalties to be applied to undisputed violators. This sends a strong message to other schools that the state is serious about maintaining the integrity and fairness of the assessment program and provides an incentive for offending schools to follow such policies exactly in the future. Such action also supports the integrity, credibility, and value of the assessment results for those who did follow the rules.

Schools cannot be allowed to pick and choose which rules they will follow; all written procedures must be followed by all schools to maintain an equitable and fair statewide assessment program. Allowing even minor deviations opens the door to a gradual erosion of standards because it then becomes difficult to draw the line on how much deviation is too much. Random, unannounced auditing is necessary to ensure that schools are following established guidelines.

A.8.1 Application to KIRIS

Due to the seriousness of the consequences associated with KIRIS performance, there is an incentive for individuals to compromise the validity of the examination. If an individual perceives that desired rewards may more easily be attained by deception than by the hard work required to achieve curricular goals, unethical activities may ensue. (Examples from OEA memo re '94/'95).

AS has written procedures for handling the shipping and receipt of assessment materials. Returned materials are counted and deficiencies questioned. Schools are responsible for proctoring but are not required to fill out irregularity

forms. AS scorers noted irregularities and referred them to the Kentucky Board of Review for investigation. Booklets are not numbered or sealed to prevent access to questions prior to the scheduled assessment period.

Despite the potential for violation of assessment security in a high-stakes environment, KDE has no capability for investigating irregularities. KDE requests that OEA investigate any reported irregularities using an honor system in which the districts provide unconfirmed evidence. There are currently no provisions for proactive auditing or seeking information on potential security violations. In contrast, the Texas Education Agency maintains a full time investigator to handle security violations in its high-stakes student assessment program (Phillips, April 1991).

Anecdotal reports indicate that a guidance counselor copied the 1994-95 grade 8 KIRIS assessments, and distributed the copies to teachers who used these items in classroom instruction. The students recognized the items as repeats when the assessment was given. This is a serious security violation. However, due to the honor system and the fact that the proposed Code of Ethics for State Testing Program was never enacted, appropriate sanctions may not be imposed. This sends a message to schools of tolerance for "cheating" and disregard for fairness to schools which participated honestly.

No data regarding KIRIS irregularities is available for 1991-92 or 1992-93. A memo summarizing the disposition of assessment irregularities for the 1993-94 KIRIS (Owen, Jan. 19, 1995) listed 16 incidents which were investigated by the Board of Review. All but one were irregularities detected by AS during the scoring process. One was initiated by a letter from a teacher.

For each reported irregularity, the Board sent a letter to the district requesting a response. Based on the districts' responses, the Board maintained the original scores in 9 incidents and invalidated responses and changed student classifications to novice in 7 incidents.

The bulk of the incidents involved open-ended responses to reading, science and social studies tasks. Allegations involved different handwriting across tasks for the same students, copying answers, use of or copying from reference materials, inserting additional information for portions of the assessment completed earlier, looking ahead and taking notes on tasks in a portion of the assessment to be administered the next day, and small groups of students with identical responses to open-ended tasks.

The Board of Review results indicate that AS was the source for nearly all information on irregularities and this information was based entirely on analysis of student responses. Further, the resolution of all incidents was entirely dependent on districts investigating themselves. It is unlikely that this procedure would have detected failure to follow directions for administration, teacher coaching, unauthorized use of references which were paraphrased in the student's own words, peeking ahead to sections to be administered later and obtaining background information regarding content from parents, teachers or friends prior to responding. Also not likely to be detected were any breaches of security involving educators reviewing assessment tasks prior to administration and providing background information, practice on parallel tasks, or overuse of released common items for drill and practice sessions.[8]

One of the ways other states have addressed the security of assessment materials is to make all copying illegal. In such environments, possession of any copies of assessment materials constitutes a violation. However, Kentucky has allowed copying of student responses to common items before assessment materials are returned to AS. Thus, unauthorized copying of matrix items and responses has been made easier and anyone caught with such copies could argue they simply made a mistake while legitimately copying the common item responses.

To the extent that security violations occurred during Biennium I, there may be an incentive for escalation of unethical behavior in subsequent biennia since the

baseline for the next biennium will be determined from the inflated scores from the previous biennium.

A.9 Assessment Disclosure -- Free Exercise of Religion/Freedom of Speech

Some parents of Kentucky students required to take KIRIS have requested an opportunity to review the assessment. Parents who request assessment review appear to be concerned about the possibility of a systematic attempt by the state to influence their children's beliefs through religious, political, or values-oriented content in the assessment questions.

A.9.1 Legal Background For Parental Challenges

The legal basis for claims by parents of a right to review assessment instruments prior to administration to their children is summarized by Phillips (1994):

> **Free Exercise of Religion and Freedom of Speech**
>
> [Test items] may . . . be subject to first amendment freedom of speech and free exercise of religion challenges. The first amendment to the U.S. Constitution provides in part that "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof." Both the establishment and free exercise clauses of the first amendment have been used to challenge specific requirements imposed by states and school districts.
>
> In particular, parents with definite religious or political views may object to any [test item] which appears to require the student to espouse a specific belief or point of view. Performance tasks which require students to read materials about controversial topics and form judgments about alternative actions may trigger such parental challenges. Social studies materials which deal with politically controversial topics may also be targets for challenge.
>
> For example, suppose a reading passage excerpted from a news magazine article described a survey of American religious beliefs in which a random sample of Americans from a variety of denominations responded to questions such as "How often do you attend church services each month?" or "Do you believe in the theory of evolution?". Suppose further that a reading comprehension question following the passage asked which denomination reported attending church services most often each month. Some parents might object to this question as implying a value judgement about the superiority of one denomination over another or as an attempt to influence students for or against their particular denomination.
>
> Similarly, an essay question which asked students to compare and contrast Americans' beliefs about the theory of evolution by religious affiliation might garner the same criticism from parents who object to any attempts by public schools to influence students' religious beliefs. The problems described above could be corrected by substituting a news magazine article describing a survey of the number and types of automobiles owned by a random sample of Americans. Presumably, asking students to summarize information or form judgments about automobile ownership would be much less controversial.
>
> The first amendment to the U.S. Constitution also provides that "Congress shall make no law . . . abridging the freedom of speech." As long as the speech is not disruptive or vulgar, the courts have held that students are entitled to the free speech rights guaranteed by the first amendment (*Tinker v. Des Moines Independent Community School Dist.*, 1969; *Bethel School Dist. No. 403 v. Fraser*, 1986). While schools have been allowed some control to maintain student order, complete censorship of the content of student speech is not permitted (*Board of Educ. v. Pico*, 1982; *Hazelwood v. Kuhlmeier*, 1988). Thus, students have been allowed to wear armbands to protest the Vietnam war (*Tinker*) and to review R-

rated movies in the school newspaper (Desilets v. Clearview Regional Board of Education, 1994) but not to make campaign speeches containing sexual innuendo (*Fraser*). The Supreme Court has also upheld a student's right not to speak (*West Virginia State Bd. of Educ. v. Barnette*, (1943).

[Test items] may be challenged as violating the first amendment to the extent that they prescribe or proscribe certain viewpoints. For example, the draft physical education standards developed by the National Association for Sport and Physical Education include assessment criteria under a diversity objective which states that students "will not actively voice or exhibit exclusionary behavior" (Education Week, October 19, 1994). Suppose a school has adopted this standard and a student is found to have violated it because the student advocates separate education for students with severe disabilities. While such a statement is contrary to the beliefs of those advocating full inclusion, in specific cases it may satisfy the least restrictive environment dictates of the Individuals with Disabilities Education Act (I.D.E.A., 1991) and is a position advocated by some parents of disabled students (Education Week, October 5, 1994). Requiring the student to support one viewpoint over the other may be seen by some parents as a violation of the student's free speech rights.

The arguments for free speech in schools are similar to those used to strike down hate speech codes (*Doe v. University of Michigan*, 1989). No matter how reprehensible the viewpoint, the Supreme court has held that the first amendment protects the speaker from punishment. [Giving low scores to students who express unpopular viewpoints or who fail to express politically correct viewpoints may be seen as such an inappropriate restriction on free speech.]

Parental Rights

Even if the content of multiple-choice questions or performance tasks does not infringe on free speech or religion, parents may still demand the right to preview such assessments prior to administration. A decision is expected soon in a Texas case in which parents have demanded the right to review all state-mandated assessments prior to administration to their children. They argue that denial of assessment review requests violates the constitutional rights of parents explicated in other federal cases.

However, there are no specific parental rights enumerated in the Constitution. The reference to parental rights appears to originate from the fourteenth amendment due process clause together with the ninth amendment which provides that "[t]he enumeration in the Constitution, of certain rights, shall not be construed to deny or disparage others retained by the people." In *Pierce v. Society of Sisters* (1925), the Supreme Court established the right of parents to direct the education and upbringing of their children. Courts have also held that parents have the right to homeschool their children, exempt their children from curricula such as sex education which may be contrary to their religious beliefs, and obtain special education assistance for qualifying students who attend private schools. In addressing parental rights issues, the Supreme Court has balanced the concerns of the parents against the policy objectives of the state or school district. (p. 15-19).

One factor to consider when balancing the interests of the state against those of a parent is whether the collective benefit to many students outweighs the alleged infringement on parental rights. One could argue as follows. The state has a duty to ensure students are learning the curriculum. The only way the state can meet this obligation is through valid assessment and remediation. Valid assessment requires secure test materials. Breach of the state's duty may be educational malpractice and may provide grounds for a suit by parents of students who did not learn and were not remediated.

A.9.2 Policy Issues Related to Public Test Disclosure

A number of issues and their policy implications should be considered when formulating a test disclosure policy. These include: equal opportunity for success -- fairness, professional standards, threats to test validity, and the security risk of post administration disclosure.

Equal Opportunity For Success -- Fairness

The concept of ***opportunity for success*** was discussed in an earlier section. The following paragraphs expand on its relevance to assessment disclosure.

Consider the following hypothetical example which describes how a public disclosure policy may result in violation of the opportunity for success requirement. Suppose parents of students with low achievement in mathematics at School K reviewed the upcoming KIRIS assessments to be administered to the students at School K. Suppose further that, after viewing the KIRIS mathematics assessment, these parents briefed the teachers in School K about the mathematics content and skills for which they expected their children to have difficulty. This would provide an unfair advantage to the teachers in School K and an unfair penalty to teachers in other schools that did not receive such targeted instructional information. The teachers in School K would be teaching to the specific content sampled by the assessment while the teachers in other schools would be teaching to the general domain of all possible content that could be assessed. This scenario might also create tension among teachers or between teachers and parents who would employ different ethical standards when determining how such "inside" information should be used.

There is an even more insidious form of cheating that could occur if large numbers of parents were allowed to review KIRIS assessment tasks. Similar fraudulent schemes have been documented in some state licensure testing programs. In such schemes, the perpetrators effectively steal portions of the assessment. In the KIRIS context, such a scheme might work like this. A group of parents who wish to have access to the assessment arrange to review it. Prior to the viewing, they assign each individual a small set of items to memorize while reviewing the assessment. For example, person 1 is assigned items 1-5, person 2 items 6-10, person 3 items 11-15, etc. With enough individuals participating, the number of items assigned per person can be relatively small. After leaving the assessment review area, each person immediately writes down the assigned items from memory. The information is collected from all persons and assembled into an intact assessment form. A modified version of such a scheme was used by parents in another state who believed the test included inappropriate values questions. In that case, students who took the test either recited items from memory after the test or smuggled out notes containing copied questions.

Professional Standards

Professional standards demonstrate concern for potential fraud in high-stakes testing programs. The *Code of Fair Testing Practices* (1988) states:

> Test Developers or Test Users Should: . . . 18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. ***Strive to make such information equally available to all test takers.*** (p. 4, emphasis added).

Relevant *Test Standards* 15.3 and 15.7 were cited and discussed earlier in the section on *Assessment Security*.

Threats to Assessment Validity

Meaningful and fair decision-making can only occur when assessments are valid and reliable. Unethical test preparation threatens assessment validity and reliability because some students receive spuriously high scores which do not

accurately reflect achievement of the skills measured by the assessment. Phillips (1994) elaborates on the security concerns referred to in the *Test Standards* and by the *Code of Fair Testing Practices* which may threaten test validity under a policy of public disclosure:

> With respect to review of assessment questions and tasks, states and school districts have a compelling interest in maintaining the security of the assessment instrument. Because it would be unworkable to attempt to limit review to only those parents with sincerely held religious or political objections, the [assessment review] would have to be open to all parents. If large numbers of parents were given access to secure assessment instruments, the potential for compromising security would be greatly increased, Elaborate procedures would be necessary to monitor the assessment review to ensure that no copies of the exam were lost and no assessment questions were copied. Despite careful monitoring, some parents may still attempt to memorize assessment questions or content. Performance assessments are particularly vulnerable because they contain fewer questions to remember than a multiple-choice exam. If some examinees know the questions or even the topics covered by a performance assessment ahead of time, they will have an unfair advantage over others who do not. Those who argue that this would never happen might recall survey results reported by Popham (1991) in which 57% of California teachers and 60% of California superintendents indicated that practicing on actual test questions from previous forms was appropriate test preparation activity. (p. 19-20).

As indicated above, if some educators (who should know better) believe cheating on high-stakes tests is acceptable, certainly some parents might also share that view. It only takes a single unethical act to compromise an entire assessment program. If an assessment program is compromised, the resulting data is invalid, the results are unusable, and the substantial resources used to develop and administer the assessment program have been wasted.

Security Risk of Post Administration Disclosure

Even if parent viewing occurred after the KIRIS was administered, it would pose a security risk. Phillips (1994) summarizes the arguments against post administration disclosure of assessment items:

> It has been suggested that [security concerns] could be alleviated by review of the assessments after administration with an option to invalidate student answer sheets if parents identify objectionable content. However, such an option would not alleviate the need for security to preserve the technical quality of the assessment program and could adversely affect state and district attempts to evaluate the effectiveness of educational programs. Additionally, . . . significant expenditures [would still be incurred in] providing adequate facilities, scheduling, and monitoring for parental review sessions.
>
> There are three major reasons why [the state] may need to maintain assessment security after a high-stakes assessment is administered: reuse of items, maintaining equivalent passing standards, and pretesting of items. Thus, compromise of assessment security after administration could still have grave consequences for the quality and viability of an assessment program.
>
> Because tasks and items are expensive to develop, they are often reused in subsequent assessments. The inability to reuse items and the resulting need to develop all new items each year might seriously strain the resources of the assessment program. This

> could result in a decrease in item quality and could adversely affect the reliability and validity of the assessment instrument.
>
> In addition, some overlap of items from year to year is needed to equate assessment forms so that a consistent passing standard can be maintained. Failure of the equating process could lead to nonequivalent passing standards across consecutive years. Such nonequivalence could result in greater numbers of students failing in a subsequent year because the assessment was more difficult that year. . . . Some educators and parents might view such circumstances as arbitrary, capricious and unfair.
>
> Finally, states and districts have a compelling interest in pretesting assessment items. Items are pretested by imbedding small numbers of new items among the scored items on an assessment instrument. Students do not know which items count and which are nonscored pretest items. This technique provides the most accurate data possible for judging item quality because student motivation and assessment conditions are the same for scored and pretested items. (p. 20-21).

A.9.3 Application to KIRIS

A KDE Staff Note on Assuring the Quality of the KIRIS Assessment Program states the KDE policy on assessment disclosure. KDE has:

> Established [a] procedure for any citizen to review all test items, provided the citizen sign[s] a non-disclosure statement to protect the security of items needed for equating across years. (p. 1)

At a panel meeting in Kentucky in March 1995, KDE stated that the policy allows any citizen to arrange with the local school district to review KIRIS. No guidelines for such reviews have been developed for the districts and KDE has no statistics available on how many individuals have reviewed the KIRIS assessments.

Kentucky has already faced a lawsuit over a student's refusal to take KIRIS based on religious and free speech objections. The student's parents have alleged that KIRIS "[compels] a child to speak against his or her wishes by selecting morally objectionable responses" (Associated Press, 6/23/94).

Because students who do not take the test are counted as novices, thus lowering the school's accountability index, the school named in the lawsuit had made taking (but not passing) KIRIS a graduation requirement. The school further justified the requirement by asserting that KIRIS assesses only academic skills.

In another incident, the parents of 28 out of 90 fourth grade students at Reidland Elementary School in McCracken County School District kept their children from taking the KIRIS transitional tests. These students were classified as novices for accountability index calculations. On appeal, the school requested that these students be excluded from accountability calculations. KDE recommended that this appeal not be granted and the Educational Improvement Advisory Committee (EIAC) which heard the appeal concurred stating, "While this was an unusual circumstance, the school or district could have instituted policies that would have prevented or minimized this incident" (Appeal 22, Summary Appeals of Accountability Cycle I Performance Judgments, March 23, 1995, p. 26). The Commissioner followed the EIAC recommendation and denied the appeal.

The key questions in this debate are:

1. How does one draw the line between appropriately assessed academic skills and inappropriate religious, political or values-oriented content?

2. Do assessments which require students to respond to controversial topics violate religious freedom or constitute compelled speech in violation of the First Amendment? and
3. What evidence is there that KIRIS contains objectionable content?

To provide data for considering these questions, the content of the 12 forms of the 12th grade KIRIS assessments for 1993-94, 1992-93 and 1991-92 were examined. Multiple-choice and open-ended questions were considered.

Reading

The pool of 47 passages for the KIRIS reading assessments were classified into nine broad theme areas. The following table provides a breakdown of the content of the reading passages for each assessment year.

	1991-92	1992-93	1993-94
Experience of historically disadvantaged groups (immigrants, African-Americans, persons with disabilities, women, Native Americans); multiculturalism	16%	20%	23%
Respect for animals; animal rights	19%	20%	13%
Health (physical, mental, nutrition, fitness, safety, consumerism)	25%	30%	35%
Environmentalism (endangered species, global warming, appreciation of nature, hazardous waste, etc.)	9%	10%	17%
Criticism of traditional education	9%	7%	7%
Evolution			3%
Work & job searches	9%	7%	3%
Art/History	3%	3%	
Science fiction, space, computers	9%	3%	

All but the last three themes represent controversial topics for which members of different religious and ethnic groups have strongly held views. For example, some groups object strenuously to the portrayal of women in nontraditional roles; some groups feel strongly about preservation of the environment while others believe the economic welfare of landowners and inhabitants should take precedence; and some groups believe that animals should provide food and clothing to humans while others disagree. The use of so many controversial themes in the KIRIS Reading Assessments has the potential to spark challenges from those parents who do not want their children exposed to views which are not consistent with the parent's views or the teachings of the student's religion.

To illustrate, consider the themes in the reading passages from the 1993-94 KIRIS assessments. Each senior who took the 1993-94 KIRIS Reading Assessment was exposed to 8 selections, 6 of which were common to all students and 2 which were unique to the particular form administered to the student. Of the 6 common selections, 2 were about members of historically disadvantaged groups, 2 were about health, 1 was about environmentalism and 1 was about respect for animals. The question of importance to parents challenging the content of KIRIS might be paraphrased as: "Is there a state-sponsored message in the assessment selections which attempts to influence student beliefs?"

It appears that the KIRIS selections are pro multiculturalism, pro women's rights and nontraditional roles, pro animal rights, pro healthy diet and exercise, pro consumer rights and pro environment. The average Kentucky parent probably will agree with some of these positions while disagreeing with others. KIRIS would be on much safer ground if the content of its reading selections focused on noncontroversial themes and appeared more neutral and balanced. Interestingly,

the 1991-92 passages appeared to be more balanced and to have less of an agenda than the 1993-94 passages.

There are a couple of passages in the pool that raise particular red flags. One is a passage listing 10 commandments for a safe prom and graduation. Although the message on alcohol consumption is laudable, use of a serious religious symbol in this context may be offensive to some people. Another narrative which depicts a kid chain smoking in the presence of an apparently accepting parent probably would also be seen as inappropriate by some parents.

The inclusion of controversial themes increases the likelihood that some groups of parents will challenge the KIRIS assessment program. Such challenges can lead to extensive disclosure of secure assessment items. Reading passages tend to be memorable and those who view them may remember enough to impact the prior knowledge of future students administered KIRIS.

Concern over the security of the pool of KIRIS reading passages is also of concern because the existing pool is relatively small. The following tables illustrate this point.

Passage	'91-92	'92-93	'93-94	Passage	'91-92	'92-93	'93-94
1	C *			25	M	M	C
2	C			26	M	C	
3	C			27	M	M	M
4	C			28	M	C	
5	C			29	M	M	M
6	C			30	M	M	M
7	C			31	M	C	
8	M	M	M	32	M	C	
9	M	M	M	33		M	C
10	M	M	C	34		M	C
11	M	M	M	35		M	M
12	M	M	M	36		M	
13	M	M	M	37		M	M
14	M	M	M	38		M	M
15	M	M	M	39		M	C
16	M	C		40		M	M
17	M	M	M	41			M
18	M	M		42			M
19	M	M	C	43			M
20	M		M	44			M
21	M	M	M	45			M
22	M			46			M
23	M	C		47			M
24	M						
				TOTAL	32	30	30
				COMMON	7	6	6

* C = common; M = matrix

		Common Items				Matrix Items						
					91	only						91/
Year	Passages	New	91	92	92	91	92	93	91/92	91/93	92/93	92/93
1991-92	32	7	-	-	-	1	-	-	7	2	-	15
1992-93	30	-	6	-	-	-	1	-	1	-	7	15
1993-94	30	-	-	3	3	-	-	7	-	1	4	12

These data indicate that for the 1993-94 KIRIS Reading Assessment, 3 of the 6 common passages appeared on both of the previous assessments, 3 common passages appeared on the '92 assessment, none of the common passages were new, 12 of the matrix sampled passages appeared on both of the previous reading assessments, 4 matrix passages appeared on the '92 assessment and 1 appeared on the '91 assessment. Thus, of the total of 30 reading passages in 1993, 50% had been used three years in a row, 23% had been used on the previous year's assessment, 3% had been used on the first year's assessment, and 23% were new passages that had never been used before.

Overall, there are 47 passages in the KIRIS reading assessment pool. Nineteen of the 47 were common passages and have been released. Of the remaining 28 passages, 12 have been used all three years, 6 have been used in 2 of the last 3 years and 10 have been used only once before. If, for security reasons, the 12 passages used 3 years in a row were not reused again in 1994, only 16 usable passages would remain in the pool. Based on a continuation of the same design, fourteen new passages, twice as many as were used in 1993, would then be needed to assemble the 12 forms of the reading assessment. These data suggest that the possibility of data contamination from parents, teachers or students "remembering" passages is significant in the Kentucky high-stakes environment.

Social Studies

Social studies is another area where parents have challenged assessment content. The most directly objectionable items are ones which deal specifically with religious content. Forms 1-12 of the 1993-94 KIRIS Social Studies Assessment contained three items explicitly dealing with religion: one asked about the leader of the Moslem religion; two others contained an excerpt from an 1848 Declaration of Women's Rights in which the word "God" was used and the text argued that the Creator intended equality for women. A closely related item about the Salem witchcraft trials might also be viewed by some parents as objectionable on religious grounds.

Though not per se inappropriate or referring to religion, 14 other questions contained content similar to that objected to by parents challenging other statewide assessment programs. Two were common items which appeared on all forms of the KIRIS Social Studies Assessment and dealt with the characterization of traditional family roles and ways in which those roles have hindered equal opportunity for women.

Among the remaining 12 items which might be judged objectionable by some parents due to their conflict with religious teachings, there were 4 on feminism and women's roles, 1 based on an excerpt from the Reader's Guide on evolution and 1 on childcare when both parents work. Six others involved controversial topics including immigration policy, HIV testing, acceptable behavior related to beliefs/values, gun control, and the interpretation of culture from gargoyle masks and a painting of a woman's face.

Writing

Writing prompts require students to choose sides on some controversial subjects (e.g., AIDS testing, euthanasia, parenting, citizens owning firearms, banning smoking in public places, rating lyrics in popular music, effects of moving away from traditional family structure, high school students marrying before graduation, added responsibility outside of school). When responding to these prompts, students are encouraged to cite evidence, personal experiences and examples.

Some parents may consider these topics too personal and sensitive. Such parents may feel that students are being coerced into speaking and sharing personal information. Some of the prompts specifically instruct students to share their feelings about an issue. These requirements may conflict with the family values and religious teachings of some parents.

Summary

Generally, it is advisable to remove any content which may be challenged based on religious or free speech objections. It is also desirable to eliminate controversial topics which deal with issues and values that may be offensive or controversial for some groups. Such sensitive content might be better handled, if at all, in the classroom with human intervention than on an impersonal statewide assessment.

A.9.4 KIRIS Sensitivity Guidelines

The following concerns were raised by some members of a statewide Sensitivity Review Committee which issued a report in 1994:

* Taken as a whole, the test favors a liberal viewpoint.
* Many questions appear to seek a politically correct response.
* There is a lack of patriotic and/or traditional literature used on the assessment.
* Religion is often cast in a negative light.
* Multiculturalism is over emphasized.
* Environmental issues are addressed in Science, Vocational Studies/Practical Living, and Social Studies with preference given to the environmentalist viewpoint.
* The feminist viewpoint is given preference over traditionalist views. (KIRIS Review Committee Report, June 29, 1994, p. 1-2)

With the exception of overemphasis on environmentalism, these recommendations were discounted by KDE as a minority viewpoint. However, the entire committee did recommend:

* Incorporate more patriotic and traditional stories in the assessment.
* Be as sensitive to the Christian right as we are to the other ethnic groups in our presentation of subject matter.
* The continuous assessment has four somewhat depressing items at the beginning; *The Flowers, Acid Rain,* and two poems: *Husband and Wives,* and *Ties.* Perhaps these could be spread farther apart or balanced with something more upbeat.
* Use legitimate/proven scientific articles about astronomy, physics or chemistry.
* Test items should not use company names such as Levi Strauss or Nordic Track since they imply state endorsement of company products or policy.
* The Reading passage, *The Flowers,* contains subject matter some parents may consider inappropriate for statewide assessment and should be eliminated. (p. 4-5)

The recently developed *Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development* (no date) addresses many of the issues described above and provides helpful guidance for the selection and writing of assessment materials. However, for some parents it may not go far enough in eliminating objectionable or offensive content. For example, creation/evolution, violence,

animal rights, religion and family issues appear on a list of topics labeled "may or may not be suitable" for KIRIS. According to these guidelines, birth control and witchcraft may be used in a broad context. It would seem a wiser strategy to avoid such topics. Surely there are enough noncontroversial topics that can be used to minimize potential legal challenges and decrease the threat to assessment security from large numbers of persons requesting to review the assessments.

A.9.5 Independent Review

In the event the Board of Education determines that it is necessary to provide assurance to the public that KIRIS does not assess inappropriate religious, political or values-oriented content while preserving assessment security, a review of secure KIRIS assessment materials by a psychometric consultant may be commissioned by the Board. The consultant could be asked to prepare a written report specifically addressing the issue of inappropriate content in assessment items.

Ideally, the selected consultant would be mutually agree upon by the Board and concerned parent groups. If an acceptable consultant could not be agreed upon by both parties, each could select a consultant to review the content of the KIRIS.

Prior to the review, the consultant(s) would be asked to sign a nondisclosure agreement. KIRIS secure assessment materials would be reviewed on site at the Department of Education under the supervision of its designated representatives. The consultant would not be allowed to copy any secure assessment materials nor to reproduce actual assessment items in the report. The charge to the consultant(s) would be to report the numbers of items found to contain objectionable religious, political or values-oriented content. In the report, the consultant could identify the item number and general topic of any question judged objectionable.

To provide a representative sample of KIRIS content, consultant(s) could be asked to review the KIRIS assessments for 1993, 1994 and 1995. The consultant(s) could also be asked to determine if each KIRIS assessment item is a valid measure of the core skills intended to be measured. The consultant(s) could be requested to include a summary of this information in the written report.

The written report(s) from the consultant(s) could be submitted to the Commissioner of Education for review and comment and then forwarded to the Board with recommendations for further action. It would be the prerogative of the Board to take further action as it deemed appropriate.

A.10 Accommodations For Persons with Disabilities

The Americans with Disabilities Act (A.D.A.) requires that reasonable accommodations be provided to disabled students. This means that disabled students must be considered when writing goals or standards which apply to all students, when developing assessment items or tasks and when determining passing or other reporting standards.

A.10.1 Physical Versus Cognitive Disabilities

Originally, testing accommodations involved compensation for physical disabilities such as blindness or use of a wheelchair. But recent challenges to assessment practices have focused on accommodations for cognitive disabilities including learning disabilities such as dyslexia, dyscalculia, processing deficits, etc. For cognitive disabilities, it is much harder to separate the effects of the disability from the skill being measured.

A.10.2 Valid and Invalid Accommodations

For assessments and their corresponding score interpretations to be valid, the assessment must measure what is intended and scores must have the same interpretation for all students. Such is not the case if some students read a reading comprehension test themselves while others have the test read aloud to them. For the latter group, the test is measuring listening comprehension rather than reading comprehension. Similarly, a math computation test for which some students can use calculators and others may not does not measure the same skills for all students. The *Test Standards* state:

> unless it has been demonstrated that the psychometric properties of a test, or type of test, are not altered significantly by some modification, the claims made for the test . . . cannot be generalized to the modified version (p. 78).

The courts have clearly indicated that reasonable accommodations must compensate for aspects of the disability which are incidental to the skill being measured but that test administrators are not required to change the skill being measured to accommodate a disabled examinee (*Brookhart v. Illinois State Bd. of Educ.*, 1983; *Pandazides v. Virginia Bd. of Educ.*, 1992; *Southeastern Community College v. Davis*, 1979). Drawing the line between a valid and an invalid accommodation requires consideration of the assessment specifications and the uses for the resulting scores. Articulation of an accommodations policy may involve a clash between the goal of inclusion of all students and the validity of the assessment scores for high-stakes decisions.

The most important requirement when balancing inclusion of the disabled against the validity of the assessment is to develop a comprehensive written policy outlining the procedures for requesting accommodations and detailing how decisions will be made regarding specific requests. Both case law and Office for Civil Rights (OCR) opinion memoranda indicate that accommodation decisions must be made on a case-by-case basis considering the type and degree of disability (*Hawaii State Dept. of Educ.*, 1990).

A.10.3 Excluding Scores From Nonstandard Administrations

There is still some debate about whether score notations for nonstandard administrations may be reported. Privacy statutes preclude reporting the specific disability. However, it may be permissible to report nonstandard administrations when they affect the meaning and interpretation of the assessment score. The *Test Standards* (1985) summarize the dilemma as follows:

> Many test developers have argued that reporting scores from nonstandard test administrations without special identification (often called "flagging" of test scores) violates professional principles, misleads test users, and perhaps even harms handicapped test takers whose scores do not accurately reflect their abilities. Handicapped people, on the other hand, have generally said that to identify their scores as resulting from nonstandard administrations and in so doing to identify them as handicapped is to deny them the opportunity to compete on the same grounds as nonhandicapped test takers . . . (p. 78).

The *Code of Fair Testing Practices* (1988) recommends:

> Test Developers Should: . . . 16. . . . Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores. (p. 3).

Future cases brought under the A.D.A. may provide guidance on this difficult policy issue. In the meantime, policymakers must define "reasonable accommodations" based on all available evidence regarding whether they are

impact the content being measured, the policy specifically permits a read-aloud accommodation for the reading assessment. This means that some students will take a listening comprehension assessment while most are required to take a reading comprehension assessment.

Using the qualifications for special education as the criterion may be unfair to low achieving students who are poor readers and might perform much better with the read-aloud accommodation. The policy may also be unfair due to differential application across schools. That is, students with identical disabilities may receive different accommodations depending on what schools they attend. Further, the policy has no provision for standardizing such accommodations, for example, by using a cassette tape for the read-aloud accommodation.

The policy states that alternative portfolio assessment scores are to be included in a school's results for Biennium II. Because the APA is different from the requirements for nondisabled students, scoring at the proficient level on the APA will not have the same meaning as proficient on the standard KIRIS assessments. Therefore, it is invalid to combine these scores. Effectively, those students who take KIRIS with accommodations or who take the APA will have taken different assessments and the inferences from their scores will not be the same.

Several assessment irregularity incidents investigated by the Board of Review (Owen memo, Jan. 19, 1995) illustrate some of the difficulties with the current accommodations policy. These reported irregularities involved the use of recorders for disabled student's responses to open-ended items. One district responded that the recorder was not a special education teacher, had not been given instructions on appropriate procedures, and had had difficulty keeping up with the student's dictation. As a result, the recorder paraphrased parts of the response.

In several other investigated cases, handwriting differences across responses for the same student suggested that multiple recorders had been employed. Sometimes the student wrote some of the responses while using recorders for the others. The reason for the student's ability to provide written responses to some questions but not others was not explained.

When recorders are used, stringent training and protocol are required to ensure that the recorders do not differentially assist students through facial expressions, tone of voice, comments, encouragement, paraphrasing and the like. Even when the recorder transcribes the student's exact words, an adult recorder is more likely to use correct spelling, grammar and paragraphing which may affect the score. For example, one irregularity involved an individualized educational program (IEP) which allowed the teacher to provide revised responses for scoring. The IEP stipulated that open-ended responses could be checked for accuracy so the student would be measured on what the student knew and not the student's writing ability. But such procedures are not permitted for disadvantaged students who are simply poor writers.

For students who are unable to provide written responses, a more standardized accommodation would be to have the student use a cassette tape recorder to orally record responses, including punctuation and paragraphing. Afterward, the oral responses could be transcribed verbatim by a secretary. Alternatively, detailed procedures and training for recorders could be developed and employed for all KIRIS assessments. This might include the use of two recorders for each student: one to write and the other to monitor and verify the use of correct procedures.

However, the problems noted with recorders in the irregularity reports are only the tip of the iceberg of issues related to KIRIS accommodations. KDE allows any accommodation provided in the student's IEP to also be used for KIRIS. The student's school is responsible for implementing such accommodations when the student is assessed. The irregularities identified by AS scorers are probably only a small fraction of the variations in accommodations across Kentucky schools. That is, the accommodations provided and the ways in which they are

implemented could be expected to differ depending upon which district/school the student attends.

KDE appears not to distinguish between valid and invalid accommodations and appears to treat scores from nonstandard administrations as equivalent to those from standard administrations. KDE has argued that the number of affected students is too small to make a difference. However, schools benefit by not having disabled students automatically classified as novices. Thus, in the future, there may be an incentive for schools to identify more students as disabled so those students can be assessed with accommodations. The recent experience of the Scholastic Achievement Test, where the number of examinees requesting accommodations has increased dramatically in recent years,[9] supports the likelihood that KIRIS will experience such increases. Allowing any accommodation a school chooses to offer in a high-stakes environment and combining the accommodated or alternative scores with those from standard administrations produces invalid and uninterpretable aggregate scores.

A.11 Summary and Conclusions

No large-scale state assessment program can be expected to be perfect. The political process which creates such programs must by necessity reflect compromises among constituencies with opposing views. The responsibility of assessment developers in such circumstances is to educate legislators on the consequences of alternative courses of action and to implement legislative policies within acceptable boundaries of professional standards and legal requirements.

In Kentucky, it appears that AS and KDE went beyond justifiable claims of the benefits of performance assessments and the adequacy of existing measurement technology for implementing performance assessments in a high-stakes school accountability system. KDE and AS have not produced evidence that KIRIS assessments conform to professional standards while the panel has developed evidence that KIRIS does not meet many important professional standards. Further, the evidence that was produced by AS and KDE was not used effectively to revise the assessments or to strengthen the defensibility of KIRIS.

Saying an assessment is reliable and valid does not make it so nor does the face validity of "authentic, real-world tasks." Anecdotal evidence of changes in instructional practices should not replace systematic study of consequential validity, including unintended consequences and what has been deleted from the curriculum to make room for teaching the "new" processes required for the assessments.

Professional standards and the legal requirement for fundamental fairness require an assessment to be tailored to its purpose. Kentucky has articulated two conflicting purposes, developed its initial assessments without clear assessment specifications, failed to match assessment items/tasks to Kentucky goals until after the fact, and mandated processes while measuring confounded outcomes. The relative importance of processes versus products, and content-based versus generic skills in the Kentucky assessment system has not been articulated. Kentucky has failed to provide evidence to contradict the potential claims by a school that the assessment system failed to provide adequate notice of content and emphasis, was implemented too quickly to allow schools adequate time to prepare, failed to follow professional assessment development standards, changed assessment system rules in midstream, failed to demonstrate that it is assessing clearly defined teachable skills, failed to provide adequate security to ensure fairness to all schools. Furthermore, avoidable errors were not eliminated and unavoidable errors were not minimized.

In sum, there is good reason to believe KIRIS may be vulnerable to a legal challenge. How will Kentucky react to this concern? A proactive stance may place Kentucky in the unique position of assisting the courts in establishing appropriate standards for school accountability. Complacency and a presumption

that the courts will reject all challenges to KIRIS may result in Kentucky being on the defensive and unprepared.

It is possible that the professional development benefits of KIRIS could be obtained and the accountability system be modified to meet professional and legal standards. This would require suspension of the current reward structure and a complete reexamination of all facets of the KIRIS accountability system. As an aid to this process, the following recommendations are offered.

A.12 Recommendations for KIRIS

Kentucky could decide to use its statewide assessment program in the low-stakes context of professional development. If so, more flexibility relative to adherence to professional standards and legal requirements is possible. However, if Kentucky wishes to continue high-stakes, school accountability assessment, it should do the following:

1. Define a single, clear purpose for KIRIS. Do not require one set of assessments to satisfy multiple, conflicting purposes.

2. Redesign the KIRIS assessments to directly address the stated purpose. Begin by developing assessment specifications which clearly communicate content/skills to be assessed, the role of process variables in the assessments, and the relative proportions of each content/skill and process to be included in the total assessments.

3. Provide adequate notice to schools of the specific content/skills to be assessed prior to implementation of the actual assessments. Provide sample items/tasks but discontinue release of actual assessment items/tasks each year.

4. Follow appropriate professional standards for assessment development including review of each assessment's match to its specifications, review by representatives from historically disadvantaged groups, and content sensitivity review for religious, political or other controversial topics.

5. Choose cost-effective assessment formats that satisfy the purpose for assessment. Reinstate multiple-choice items where appropriate and save the resource intensive performance tasks for content/skills which require that format for valid assessment.

6. Design a parsimonious equating procedure devoid of ad hoc procedures and reasonably calculated to maximize stability, minimize error and maintain equivalent standards across years.

7. Develop and implement written assessment security and assessment accommodations policies which are consistent with professional standards. Enact a statutory code of ethics, including penalties, for the assessment program and develop investigative and auditing capabilities within KDE.

8. Remove confounding of measurement due to group participation and lack of task standardization. If social skills are to be included, measure them separately. Consider issues of confounding of reading and writing skills with other disciplines and develop a written policy outlining the state's intent with respect to such measurements.

9. Discontinue use of unreliable portfolio scores for school accountability. Use structured, on demand writing tasks for the writing accountability assessment.

10. Reconsider the use of a single index to describe school success. Consider distributing rewards separately by discipline (e.g. reading, mathematics, science, social studies, writing) so that schools can more directly tie instructional changes to rewards for success.

11. Reconsider the use of gain scores for school accountability. Gain scores are too unreliable for high-stakes decisions. Consider rewarding all schools that meet statewide targets which increase each year.

12. Set classification standards based on a compensatory model applied to initial, intact forms of the assessments. Equate subsequent forms to the base forms to determine equivalent standards. This will eliminate the need for multiple, weighted calculations to determine school proportions. If any weighting is necessary, use standard scores to ensure that actual weights have the intended values.

13. Do not change the system during an accountability period. Provide adequate notification to schools for all modifications.

14. Collect systematic evidence of the effects of the assessment program on schools, teachers and students. Consider unintended consequences and the implications for content no longer being taught.

15. Resist attempts at parental disclosure. Use expert review, if necessary, to assure parents that objectionable content has not been included. Do not include controversial material in the assessments.

16. Use standardized procedures and equipment to address opportunity for success concerns.

17. Monitor adverse impact consistently across all years. Analyze the adverse impact of classification rates, matrix items, performance events, and portfolio measures.

A.13 References

Allen v. Alabama State Bd. of Educ., 612 F. Supp. 1046 (M.D. Ala. 1985).

AERA, APA, & NCME. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: APA.

Americans with Disabilities Act (A.D.A., 1990). Pub. L. No. 101-336, 42 U.S.C. §§ 12101 et seq.

Anderson v. Banks, 540 F. Supp. 472 (S.D. Fla. 1981), *reh'g*, 540 F. Supp. 761 (S.D. Ga. 1982).

Associated Press, (June 23, 1994). Livingston parents sue school board over test, *Herald Leader*; Livingston Couple sue local school board over performance test, *Courier Journal*; Parents claiming state tests violate religious beliefs, *State Journal*.

Beck, M. (in press). "Authentic Assessment" for Large-Scale Accountability Purposes: Balancing the Rhetoric. *Educational Measurement: Issues & Practice*.

Bethel School Dist. No. 403 v. Fraser, 478 U.S. 675, 106 S.Ct. 3159 (1986).

Board of Educ. of Northport-E. Northport v. Ambach, 436 N.Y.S.2d 564 (1981), *aff'd with mod*, 458 N.Y.S.2d 680 (A.D. 1982), *aff'd*, 457 N.E.2d 775 (N.Y. 1983).

Board of Educ. v. Pico, 457 U.S. 853, 102 S.Ct. 2799 (1982).

Brookhart v. Illinois State Bd. of Educ., 697 F.2d 179 (7th Cir. 1983).

Clarizio, H.F. & Phillips, S.E. (1989). Defining Severe Discrepancy in the Diagnosis of Learning Disabilities: A Comparison of Methods. 27 *Journal of School Psychology*, 383.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part*, 644 F.2d 397 (5th Cir. 1981); *on remand*, 564 F. Supp. 177 (M.D. Fla. 1983) *aff'd*, 730 F.2d 1405 (11th Cir. 1984).

DePascale, C. (1992). Factors influencing the cognitive emphasis of classroom tests, AERA annual meeting, San Francisco.

Desilets v. Clearview Regional Board of Education, ___ F. Supp. ___ (1994).

Diegmueller, K. (October 19, 1994). Draft Standards for Health, P.E. Are Released. XIV(7) *Education Week* 8.

District and School Test Manual (1994). State Department of Education, Jackson, MS.

Doe v. University of Michigan, 721 F. Supp. 852 (E.D. Mich. 1989).

Gibson, G. (April 8, 1995). Ex-Corbin teacher sues for part of school bonus, *Lexington Herald Leader*, A1, A4.

Hawaii State Dept. of Educ., 17 EHLR 360 (O.C.R. 1990).

Hazelwood v. Kuhlmeier, 484 U.S. 260, 108 S.Ct. 562 (1985).

Individuals with Disabilities Education Act (I.D.E.A., 1991). Pub. L. No. 102-119, 20 U.S.C. §§ 1400 et seq.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education*, Washington, D.C.: Author.

Kentucky constitution, Section 183.

Kentucky Department of Education. (March 1995). *Folder of Assessment Information from Regional District Assessment Coordinator Meetings.*

Kentucky Department of Education. (1995a). *KIRIS Biennium I Technical Manual*, Frankfort, KY.

Kentucky Department of Education. (1995b). *1992-1994 Kentucky Accountability Results For Schools and Districts*, Frankfort, KY.

Kentucky Department of Education & Advanced Systems in Measurement and Evaluation. (1991-92; 1992-93). *KIRIS Technical Report*, Dover, NH.

Kentucky Education Reform Act (KERA; 1990). House Bill No. 940; School accountability provisions: Kentucky Revised Statutes 158.645 et seq. reproduced in KDE, *1994 Kentucky School Laws Annotated*, Banks-Baldwin Law Publishing Company, 1994.

Linn, R. (1983). Curricular Validity: Convincing the Courts That It Was Taught Without Precluding the Possibility of Measuring It, in Madaus, G. (Ed.) *The Courts, Validity, & Minimum Competency Testing.*

May, L. (1995). Poor schools as likely as rich to hit KERA goals, earn rewards. *Lexington Herald Leader.*

May, L. (Feb. 8, 1995). Educators haven't decided yet how to divvy up, spend money, 13(37) *Lexington Herald Leader A13.*

Mehrens, W.A. (1992). Using Performance Assessment for Accountability Purposes, 11(1) Educational Measurement: Issues & Practice 3.

Mehrens, W.A. & Kaminski, J. (1989). Methods For Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?, 8(1) *Educ. Measurement: Issues & Practice* 14.

Mehrens, W.A., Phillips, S.E. & Schram, C. (1993). Statewide Test Security Practices, 12(4) *Educ. Measurement: Issues & Practice* 5.

Nowak, J., Rotunda, R. & Young, J. (1986). *Constitutional Law.* 3rd ed. §§ 10.6, 11.4 & 14.1-14.3.

Owen, C. (January 19, 1995). *Summary of Actions Taken as a Result of Board of Review Recommendations*, KDE memo.

Pandazides v. Virginia Bd. of Educ., ___ F. Supp. ___ (E.D. Va. 1992).

PASE v. Hannon, 506 F. Supp. 831 (N.D. Ill. 1980).

Personnel Administrator v. Feeney, 442 U.S. 256, 99 S.Ct. 2282 (1979).

Phillips, S.E. (1994a). Legal Defensibility of Standards: Issues & Policy Perspectives. Paper commissioned for the NAGB/NCES Joint Conference on Standard Setting for Large-scale Assessments, Washington, DC, October 5-7.

Phillips, S.E. (1994b). High-Stakes Testing Accommodations: Validity Versus Disabled Rights. 7(2) *Applied Measurement in Education* 93.

Phillips, S.E. (1993). *Legal Implications of High-Stakes Assessment: What States Should Know.* Oak Brook, IL: North Central Regional Educational Laboratory.

Phillips, S.E. (March 25, 1993). Testing Accommodations for Disabled Students. 80 *Education Law Reporter* 9.

Phillips, S.E. (March 11, 1993). Legal Issues in Performance Assessment. 79 *Education Law Reporter* 709.

Phillips, S.E. (1991). Diploma Sanction Tests Revisited: New Problems from Old Solutions. 20(2) *Journal of Law and Education* 175.

Phillips, S.E. (April 1991). Implementation of Statewide Test Security Policies, NCME annual meeting symposium, Chicago.

Phillips, S.E. (Dec. 20, 1990). The Golden Rule Remedy for Disparate Impact of Standardized Testing: Progress or Regress?. 63 *Education Law Reporter* 383.

Pitsch, M. (October 19, 1994). Critics Target Goals 2000 in Schools 'War'. *XIV(7) Education Week* 1.

Pierce v. Society of Sisters, 268 U.S. 510 (1925).

Popham, J. (1992). *The Perils of Responsibility Sharing, 11(4) EDUC. MEASUREMENT: ISSUES & PRACTICE.*

Popham, J. (1991). Appropriateness of Teachers' Test-Preparation Practices. 10(4) *Educational Measurement: Issues & Practice.*

Raudenbush, S.E. & Willms, J.D. (in press). The Estimation of School Effects.

Rose v. Council for Better Education, Inc., 790 S.W.2d 186 (Ky. 1989).

Schnaiberg, L. (October 5, 1994). Chicago Flap Shows Limits of 'Inclusion,' Critics Say. XIV(5) *Education Week* 1.

Schools For Scandal. (April 27, 1992). *U.S. News & World Report* 66.

Southeastern Community College v. Davis, 442 U.S. 397 (1979).

The Assessment Team of Consultants (Redfield, Kifer, Millman, Rindone, Wiggins). (June 27, 1991). *Recommendations on Proposals to Implement an Interim and Full-Scale Assessment Program for the Commonwealth of Kentucky.*

Tinker v. Des Moines Independent Community School Dist., 393 U.S. 503, 89 S.Ct. 733 (1969).

U.S. Constitution, Amendment I; Amendment XIV, section 1; Amendment IX.

United States v. LULAC, 793 F.2d 636 (5th Cir. 1986).

Village of Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252, 97 S.Ct. 555 (1977).

Wards Cove Packing Co. v. Antonio, ___ U.S. ___, 109 S.Ct. 2115 (1989).

West Virginia State Bd. of Educ. v. Barnette, 319 U.S. 624, 63 S.Ct. 1178 (1943).

Willms, J.D. & Raudenbush, S.W. (1989). A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability, 26(3) *Journal of Educational Measurement* 209.

A.14 Endnotes

1. Portions of the general descriptions of legal issues related to high-stakes assessment that are contained in this chapter originated under contract number RN92107101 of the Department of Education/National Assessment Governing Board pursuant to presentation at the Joint Conference on Standard Setting for Large-scale Assessments, Washington, DC, October 5-7, 1994.

2. Note that the computations described as average dropout rates across grades and years are really total proportions. For example, from the KDE Staff Note regarding the formula for determining successful schools, 1994-96 biennium, p. 3, the average 1993-94 dropout rate across grades was (d9 + d10 + d11 + d12)/4 = (3+5+7+12)/4 = 6.75 . The "average" reported on that page is the total proportion of dropouts for all grades = (d9 + d10 + d11 + d12)/(n9 + n10 + n11 + n12) = (3+5+7+12)/374 = .072 = 7.2% . The estimated total dropout (7.2% * 4) = 78%) is uninterpretable.

3. Taken from *Scoring Rubrics for Grade 12 Common Open-Ended Items*, Sossman memo, AS, May 24, 1994 and KIRIS 1991-92 Performance Level Guides for Open-Response Common Items Grade 12, *1991-92 Technical Report*, Appendix I, KDE, Frankfort, KY, p. 11.

4. Samples taken from KIRIS 1991-92 Performance Level Guides for Open-Response Common Items Grade 12, *1991-92 Technical Report*, Appendix I, KDE, Frankfort, KY, p. 11-12.

5. Note that employment assessments, unlike graduation tests, are subject to Title VII of the Civil Rights Act and the corresponding Equal Educational Opportunity Commission (EEOC) *Uniform Guidelines for Employee Selection* which provide for a presumption of discrimination when the success rate for disadvantaged applicants is less than 80% of that for nondisadvantaged applicants.

6. Data taken from *Biennium I Technical Manual*, KDE, 1995, p. 272-84.

7. Data taken from *1992-1994 Kentucky Accountability Results For Schools and Districts*, KDE (1995).

8. It is curious that the 3 forms of the continuous assessment designed for practice in grades 9-11 were identical except for a few of the open-ended questions.

9. Nancy Cole, ETS President, personal communication, March 22, 1995.