Kentucky Instructional Results Information System: A Technical Review

January 1998

Authors

Professor James S. Catterall, University of California at Los Angeles

Professor William A. Mehrens, Michigan State University

Professor Joseph M. Ryan, Arizona State University West

Mr. Eugene J. Flores, Director of Testing, Long Beach (CA) City School District

Ms. Pamela M. Rubin, Director of Research, *Los Angeles Educational Alliance for Restructuring Now (LEARN)*

Produced under a contract between the authors, Coopers & Lybrand, L.L.P. of Louisville, Kentucky, and the Commonwealth of Kentucky Legislative Research Commission.

Table of Contents

Kentucky Instructional Results Information System: A Technical Review

J. S. Catterall, W. A. Mehrens, J. M. Ryan, E. J. Flores, & P. M. Rubin

January 1998

Section 1. Background and Methods

Authorization. This report was produced by the authors under a sub-contract between James S. Catterall & Associates (JSC&A) and Coopers & Lybrand, L.L.P. (C&L), pursuant to a contract between C&L and the Kentucky Legislative Research Commission (LRC). The subcontract to JSC&A was approved on October 4, 1997.

The main contract between the LRC and C&L called for an audit of the contracts and contractual relationships between the Kentucky Department of Education (KDE) and Advanced Systems for Measurement and Evaluation, of Dover, New Hampshire (ASME) between 1991 and fall 1997. The call for this audit by the Kentucky Legislature and Legislative Research Commission was made shortly after the termination of the KIRIS testing contract between KDE and ASME in the late spring of 1997. The request for an audit also emerged in the context of a steady stream of both heightened local and national interest in Kentucky's innovative testing system as well as long-standing controversies over KIRIS' technical qualities witnessed in various technical reports, academic articles, and the popular press over KIRIS' life span. Among the technical reports most influential in the history of KIRIS was a 1995 expert panel review of the measurement quality of KIRIS sponsored by the Kentucky Office of Educational Accountability (Hambleton, et al., 1995). As outlined in our more detailed scope of work below, our report pays close attention to the recommendations of the 1995 OEA Panel Report and the degree to which KIRIS has been responsive to specific recommendations of this panel.

Charge. One of four specific tasks described in the audit RFP issued by the KY Legislative Research Commission was the focus of the work reported here. Originally stated as Task 3 in the RFP, this task was the following:

> "Determine if the development, administration, scoring, equating, and reporting by ASME of the performance events and other assessment components of the Kentucky Instructional Results Information System (KIRIS) met a threshold of quality that would indicate, when compared to the reasonable standards of quality generally accepted by the psychometric evaluation trade community for similar services and contractual circumstances, that there was or was not substantial compliance with the contracts."

Contracted Scope of Work. Our negotiated and accepted proposal for work to accomplish Task 3 refined the original statement of Task 3 from the RFP, and called for a focused set of analyses by our review team. The following statement outlines the original negotiated and approved scope of work under this sub-contract:

> "Work under Task 3 ... is intended to examine and evaluate progress of KIRIS on eight specific recommendations made in 1995 by an expert review panel (Hambleton, et al.). ... Of central interest is the validity, reliability, and appropriateness of specific KIRIS components and practices as measures of student achievement and performance as well as measures of school success in producing student achievement and performance. This review will focus on developments and activities related to KIRIS between the formal issuance of the Hambleton report in 1995 and the administration of KIRIS assessments in the spring of 1997. We will attend to KIRIS policies and results from earlier testing and accountability cycles as needed to address the nature of responses and developments related to the eight recommendations. We will not address reports of results from the spring 1997 tests within this scope of work."

At the time our scope was proposed, about September 1, 1997, KY still awaited the release of revised 1996 KIRIS test scores and there were no firm dates set for release of 1997 KIRIS scores. Corrected 1996 KIRIS scores became available in mid-September 1997.

KIRIS score analysis added to scope. Because the 1997 KIRIS School and District Test Scores were released to the public on December 4, 1997 and because these scores bore potential relations to the topics of this review, the subcontractor obtained a change order to include analyses of 1997 KIRIS test scores in the context of this review. Several such analyses are included in this report.

A second, small expansion of scope was granted in this same change order. This was for the sub-contractor to engage in exploratory interviews with a random sample of 1997 KIRIS test scorers, as well as a random sample of past KIRIS test scorers no longer employed by ASME. The purpose of these interviews was to capture the nature of the scoring processes for KIRIS tests and the workings of scoring quality controls from the scorers' points of view. These interviews were carried out in late December 1997.

Central Focus of Our Work.

This report focuses on an overriding question regarding the technical quality of KIRIS, and within this attends to specific recommendations of a 1995 technical review sponsored by the Kentucky Office of Educational Accountability.

Focus point 1: The technical capacity of KIRIS to do what it intends to do.

The overriding goal of this report is to provide a global appraisal of the following fundamental question:

What is our review team's appraisal regarding the validity, reliability, and appropriateness of specific KIRIS components and practices as measures of student achievement and performance as well as measures of school success in producing student achievement and performance?

Focus point 2: Recommendations of the 1995 OEA Panel Report.

This question took on an operational form as we clarified our work plan and scope in discussions surrounding the contract. The operational mission of our review team became the appraisal of the responses of KIRIS -- especially the responses of its primary policy and administrative actors, KDE and ASME -- to eight specific recommendations made by an expert review panel commissioned in 1994 by the Kentucky Office of Educational Accountability (Hambleton, et. al, 1995)) -- hereafter the OEA Report. We included in our concept of mission the freedom to provide our own appraisals of the OEA Report recommendations where we felt this to be important. The eight specific recommendations for focus were selected by the Legislative Research Commission during negotiations concerning the scope of our review.

Methods:

Before turning to a summary of the key recommendations reviewed and a summary of our primary appraisals of progress related to each recommendation, we describe here our methods of going about this work.

There were three main components of our work plan -- searching for and procuring documents, interviewing various players and analysts involved with KIRIS, and making our assessments based on the information and insights obtained. The first task was to engage in an extensive search for documentation related to a wide range of technical aspects as well as impacts of KIRIS. This was accomplished through requests for documents from KDE, ASME, and OEA, and secondary requests for additional documents as we became aware of such through our reading and interviews. We also secured and read media accounts of developments regarding KIRIS since 1995. In all, we collected a total of about 200 documents, ranging in nature and heft. They included substantial technical manuals; validity, standards-setting, and equating studies;

discussion papers prepared for the National Technical Working Group; and individual correspondence regarding specific issues.

We also engaged in an extensive program of interviews with various officials and educators involved in KIRIS.

Our interviews included: Dr. Brian Gong, Associate Commissioner, KDE, Dr. Ed Reidy, Deputy Commissioner, KDE, Dr. Richard Hill, President, ASME, Dr. Neil Kingston, Dr. Lisa Ehrlich, Dr. Stuart Kahl, Mr. Richard Rizzo, all of ASME; Dr. Ken Henry and Dr. Doug Terry, KY OEA, Ms. Penny Sanders, former director of OEA, Profs. George Cunningham, University of Louisville, Larry Sexton, University of Eastern KY, Dr. Faurest Coogle, KSBA, Mr. Richard Innes, Kentucky citizen, about 10 District Assessment Coordinators gathered in a focus group, about 12 scorers who scored 1997 Grade 11 KIRIS tests for ASME, and about 12 previous KIRIS scorers no longer in the employ of ASME. Our request to interview 1997 Grade 4-5 and 7-8 scorers employed by Data Recognition Corporation of Minnetonka, MN was turned down. Two of our team members also attended a fall 1997 meeting of District Assessment Coordinators in Louisville.

In the course of our work, individual members of the team also interacted with several members of the National Technical Working Group, the main advisors to KDE and ASME concerning technical issues surrounding KIRIS. These members included Prof. Ed Haertel of Stanford University, Prof. Skip Kifer of the University of Kentucky, Dr. Jonathan Dings of KDE, and Dr. Scott Trimble of KDE.

Our team took up certain specializations for this engagement. Prof. William A. Mehrens was our lead analyst and writer on issues of standards setting and validation. Prof. Joseph M. Ryan was our lead analyst on issues of test equating and linking. Eugene Flores gave particular attention to issues surrounding student portfolios. Pamela Rubin participated in some interviews and focused on instructional effects of KIRIS. Prof. James S. Catterall managed all arrangements with C&L and LRC, led most face to face interviews, managed document search and distribution, produced analyses from 1997 KIRIS test score files, and organized and drafted much of the first draft of the final report -- incorporating substantial sections drafted by other team members.

Acknowledgments.

At the outset, we would like to acknowledge the extremely high levels of responsiveness to our requests for information from several offices involved in our search for relevant documents. These include KDE, and especially Dr. Brian Gong, who furnished us with requested materials at repeated intervals throughout our work -- right up to December 23rd, 1997. We also wish to thank officials at ASME for making time available for interviews in Dover, NH, for several follow-up discussions of issues with our team members, and for helping to arrange for interviews with ASME test scorers in a short time frame. We also thank Drs. Ken Henry and Doug Terry of OEA for providing many background documents and interview leads during our search for information and for engaging with us in discussions along the way about some of what we were seeing and concluding. We benefited also from preliminary conversations with Professor Joseph Petrosko of the University of Louisville and Ms. Wynn Egginton, University Coordinator, School Reform Initiatives, University of Louisville. We also thank Professor Marvin C. Alkin of UCLA who helped catalogue our fast-accumulating library and who also served as an internal reviewer of our draft report. Also, we would like to acknowledge the assistance of Ms. Susan Henry of the Indianapolis Office of C&L, who was our prime contact person and issue-handler on all matters of contract arrangements;

Ms. Henry was also our bridge to information about the progress of the C&L audit under the remaining three tasks. Finally, thanks go to Ms. Siobhain MacCarthy, a doctoral student working with Professor Catterall at UCLA, who managed the document library and various logistics of the team's work for this engagement.

Section 2: Main Review Question and OEA Panel Recommendations

In this section, we list the main question of our review along with the eight specific recommendations of the 1995 OEA Panel which we were asked to address. Following each, we provide an outline of our main views and conclusions.

Overarching Question Facing Our Review:

What is our review team's appraisal regarding the validity, reliability, and appropriateness of specific KIRIS components and practices as measures of student achievement and performance as well as measures of school success in producing student achievement and performance?

The two questions – (1) Does KIRIS generate appropriate measures of student performance? and (2) Does KIRIS produce appropriate measures of school success in producing student performance? – are intimately connected. Appraising the performance of schools in producing student performance with reliability and validity depends in turn on the reliability and validity of the measures of student achievement constituting the building- blocks of the assessment system. This report addresses several of the key technical issues surrounding the quality of those building blocks.

A global answer to the question hinges in our opinion on the tolerance of the KY education system and its constituents for the inevitable errors introduced at key points in KIRIS . The answer hinges as well on acceptance or rejection of a testing system that frequently shows results for which the role of actual student learning gains is both impossible to decipher as well as called into question. It is likely that different constituencies will look upon the error-proneness of KIRIS with different attitudes and thresholds of tolerance.

Hambleton et al. (e.g. p. A-21) described the technical challenges facing KIRIS in this regard generally as minimizing avoidable error in the system, and then leaving questions of unavoidable error up to policy makers who must make decisions involving trade-offs.

A testing system capable of measuring student achievement and school performance in producing achievement must be judged to be sound on several technical criteria.

1. **Standards.** The system must competently develop, set, and validate standards for student performance.

KIRIS has undertaken additional standards setting and validation in partial response to the OEA Panel report. As a practical matter, the two 1996 standards setting and validation studies were generally designed to confirm existing standards and were interpreted by KDE and ASME as doing just this to their satisfaction. We project that no net impact on the likelihood of avoidable error resulted from this process. We suggest that there are future needs for standard setting, especially if KIRIS takes stock of its accomplishments and adapts and re-sets itself for a more technically defensible future.

2. **Equating.** The system must equate tests across forms and across years, essentially so that the tests constructed for two sequential years would be expected to produce the same distributions of results if given to the same population of students.

The various issues associated with KIRIS equating are described in detail in Section 5 of our report. The equating demands for KIRIS have been substantial since they include many subject areas, different assessment formats, and have had to deal with program and policy changes over the life of the program. Equating across assessment forms within a year has been fairly routine, whereas equating assessments across years has represented a serious challenge.

Equating methods employed early in the program's development have been criticized as *ad hoc* and arbitrary. The choice of technical procedures used in the early phases of KIRIS seem to represent a decision to adapt conventional technical procedures and criteria to support the initiation and early development of an unconventional and innovative program.

A certain consistency has evolved in the technical procedures used to conduct many of the KIRIS equating tasks. In addition, KIRIS has generally been responsive to the spirit, if not the details, of the recommendations made by Hambleton et al.

An important characteristic of KIRIS equating is revealed in recent equating activities related to performance events and grade shifts for certain assessments (KY administered former 4th grade tests in reading, science, and writing to 5th graders for the first time in 1997, and former 8th grade mathematics, social studies, arts/humanities,

and practical living/vocational studies tests to 7th graders for the first time in 1997). In both cases, policy decisions were made before there was any evidence that there were adequate technical procedures to support the policies. Subsequent technical analyses revealed that performance events could not be adequately equated. A variety of equating procedures were explored for adjusting performance standards to accommodate the grade shifts. Equating procedures were selected that produced "plausible" results, a criterion not generally used in judging the adequacy of equating. KDE suggests that added criteria were used along the way; our reading of the grade shift equating documentation is that "plausibility" was a major defense of the final approach used.

The examples of equating related to performance events and grade shift adjustments illustrate a particular relationship in KIRIS between educational policy decisions and technical support procedures. To date, policy decisions seem to have been made with the assumption that some psychometric procedure existed or could be devised to support any policy decision. Those responsible for the technical aspects of KIRIS have been very conscientious and resourceful in devising technical procedures that produce results that are reasonable or "plausible" enough to support the decisions. This *post hoc* use of technology to support decisions rather than inform potential policy choices may be appropriate early in the development of an innovative educational program to insure that the demands for technical rigor do not stifle innovation. KIRIS, however, has had sufficient time to mature and, at this point in the program, technical procedures should be used in two ways: 1) to inform policy decisions before they are made; 2) to evaluate the adequacy and precision of technical procedures that are already in place.

Finally, KIRIS policy makers seem very resistant to the suggestion to re-establish baseline years for scaling, equating, and standard setting, despite a long series of changes in the program. Many statewide assessment programs and other assessment programs have reset their baseline year or re-normed their scales after a period of implementation. The task of establishing a new baseline year may present challenges (such as the possibility of introducing a new initial year that schools could "game" to keep their baselines low, as well as the general challenges accompanying a major redirection of the program), but it is not overwhelming. Given the long chain of adjustments and equating links that holds the KIRIS program together, it would seem timely and prudent to begin the process of re-establishing a baseline year and re-setting standards as soon as possible.

Recommendations. Our professional opinion is that 1997 marks a point in the evolution of KIRIS where a fresh baseline might be struck, and the recent major grade shift is a good part of the reason. Not only are new grade levels now involved in accountability testing, but as discussed immediately below, the fact that students are writing about half as many test items also creates a new base level from which performance changes should be judged.

Our recommendation for a new baseline and accountability clock for KIRIS is based on both technical considerations centered on equating as well as on evolving changes in KIRIS administration conditions described elsewhere.

In conjunction with establishing a new baseline year, we also recommend that technical procedures and analyses be used to inform policy decisions before they are made and to evaluate the adequacy and precision of technical procedures that are already in place. In particular, we recommend the use of equating procedures that allow for a more rigorous empirical evaluation of equating precision and adequacy.

3. **Administration and scoring consistency.** The system must ensure that test administration and scoring conditions remain constant across years.

Administration of KIRIS tests. Consistent, standardized administration is fundamental to "standardized" forms of testing. Even with clear performance standards and suitable equating, student test performance from year to year would be expected to respond to differences in conditions surrounding administration of the tests.

It may be in this domain where there is most reason to question the validity of performance change scores from year to year under KIRIS – administration conditions seem to have evolved both naturally and as matters of policy; administration conditions were subjected to marked changes in 1997, including the significant grade shifts described above.

On the informal side, many observers note and expect some learning about a new test on the part of both teachers and students to influence test performance over time. This became a featured part of the discourse surrounding KIRIS when in its second year, the 1993 scores showed approximately 20 percent improvements across the board in comparison to 1992 scores. Much of the assumed reason for this leap was that teachers were now more familiar with items, and students had gotten more practice with KIRIS type response forms as part of their regular instruction. How much of this should be called inappropriate "teaching to the test" and how much of this was salubrious "teaching to a test well worth teaching to" was then, and remains, a matter of debate.

Studies discussed in our *Effects of KIRIS* section report non-trivial instances of teacher assistance with students during testing of one form or another – not wholesale cheating by any regard, but a substantial suggestion that some students are benefiting from some administration practices that serve to push results up.

Lexington Herald accounts of the many high score gains on KIRIS reported for 1997 contain a predominance of reports by KY school administrators suggesting that their students were better motivated for the tests through more positive administrative conditions – free doughnut breakfasts for test

takers, special rewards for students who remained focused on their test tasks, and so on. To be fair, these reports also mention special initiatives for subject matter learning. None mentions special reading initiatives at the high school level to help explain the 61 percent gain in the Grade 11 reading test for 1997, about which we have much to say in Section 7 of this report.

Formal changes probably impact scores. Beyond conditions spontaneously arising in the field as a response to KIRIS, Kentucky continues to build-in changes in KIRIS test administration conditions expected to boost scores that remain unaccounted for in explaining annual test results. This is a potentially substantial problem of unknown magnitude, in our view. Two clear examples arise from the structure and guidelines for the 1997 KIRIS tests.

a. Administration guidelines. 1997 KIRIS DAC guidelines expressly allow spreading testing in a given subject (e.g. middle school science) out over longer periods of time (up to 3 days), and allowing students to write until they themselves feel they are finished with a test paper.

b. The Grade Shift. Even more important, 1997 witnessed a major change in the administration of grade 4 (elementary) and grade 8 (middle school) tests. This was the splitting of the tests at these two grade levels to additional grade levels described briefly above.

This split seems a reasonable response to the growing perception that KY children in the three tested grade levels were absorbing too much of the annual testing burden under KIRIS – an example being 4th graders each year writing as many as 8 open response items in each of 6 academic areas, a writing prompt response, and compiling and maintaining a writing portfolio. The split saw reading, science, and writing tests in grade 8 transferred to grade 7 and four tests at grade 4 transferred to grade 5.

Studies of the impact of KIRIS have made note concerning how "tired of writing" grade 4 and 8 students were becoming generally. One can only imagine how students feel at the end of a three week KIRIS testing window that has seen dozens of writing prompts aimed at all subjects fly across their desks.

We can only believe that under the 1997 design allowing students to concentrate on half as many writing tasks in only three or four as opposed to seven academic areas each year, students would perform better.

While this fact seems to be appreciated by officials at KDE and ASME, we could find no attempts to make adjustments for the improved performances expected under such grade-shifting.

The 1997 KIRIS scores for all grades within and thus for elementary and middle schools almost certainly include an unknown inflation factor due to grade shifting. We should note that this unknown inflation factor will interact with other influences on shifted-grade scores, such as equating errors. The net effect of all influences in the new test grades could be plus or minus. Note that in Section 9 below, we show that 5th grade gains in index scores for 1997 are lower than 4th grade gains, on average; and that 7th grade gains are lower than 8th grade gains, on average.

c. **Possible inclusion of KIRIS scores on transcripts.** There is ongoing discussion of a future policy change that would have KIRIS scores reported on student transcripts. This discussion seems justified because there is reason to believe that student motivation for performance on KIRIS tests is variable and may suffer from the fact that there are few material consequences to students themselves tied to their KIRIS performance. The various discrepancies between KIRIS scores and student scores on the ACT, SAT, Advanced Placement, and National Merit Scholarship qualifying tests may have much to do with the fact that these "comparison" tests are in fact high stakes tests for students whereas KIRIS tests are not high stakes for students.

The possible inclusion of KIRIS scores on student transcripts is a change that, first, demands additional technical evaluation of the reliability of student level scores – especially the equating of tests across forms since students in the same grade respond to different open response items within each tested subject. Such a policy demands that student level score reliability is demonstrated.

Second, the inclusion of KIRIS scores on student transcripts would also elevate the importance of KIRIS tests to students and their families and would likely have the effect of elevating reported average scores. Such an increase should be considered an artifact of changing test administration conditions and not an increase of average student or school performance. As such, a policy of including KIRIS scores on student transcripts would provide a substantial added rationale for establishing a new accountability baseline to coincide with the implementation of such a policy.

Scoring

An additional requirement of effective determination and reporting of student and school performance changes is consistent scoring of tests *vis a' vis* established standards and scoring guides from year to year.

We were unable in the scope of our work to conduct original analyses of scoring quality; in addition, the documentation of processes related to KIRIS does little to produce evaluatable studies of scoring accuracy for given tests and across years.

We did interview samples of scorers as part of our work. Within severe limits of this "audit," it appeared that various procedures to guarantee scoring consistency and accuracy are consistently maintained during the scoring of KIRIS tests. These procedures include substantial initial training and qualification of scorers through gauging their performances on sample tests in each area scored, monitoring of scorer performance through regular "back-reading" of their work, having scorers regularly score previous year

tests throughout the scoring process to monitor equating of scoring from year to year, and the maintenance of a reasonable work flow – scorers consistently report scoring 15-25 papers per hour, depending some on grade level and subject.

Non-trivial scoring changes. As described under our discussion of Panel Recommendation 6 and in our report of our scorer interviews, small changes appear to have impacted KIRIS scoring in non-trivial ways over the years. As of 1991-92, open response items within a subject area such as mathematics needed an average score of 3 "and no scores below 3" to be classified as "proficient." By the end of Accountability Cycle 1, the "no score below 3" proscription was relaxed. This change probably contributed to gains on KIRIS between the first base year and the end of Cycle 1. Another scoring "relaxation" was described by a seasoned and senior KIRIS scorer regarding the 1997 scoring of 11th grade science – this was the elimination of what was described as the "no flaws clause" for assigning a score of 4 (out of 4 possible points) to a given open response. Prior to 1997, a paper, no matter how high in quality, could not be given a 4 if it contained an error of fact.
(Note: We do not believe we have cognizance of all such changes that may have been made between 1992 and 1997.)

4. **Reporting scores and confidence intervals.** A testing system should report gain or change scores in ways that allow observers to grasp in a basic way the potential for error in the system.

This was an essential recommendation of Hambleton, et al. This recommendation was marginally responded to in the published reports of 1997 KIRIS scores disseminated by KDE.

Two footnotes to the 1997 KIRIS Score Reports fall into this category. One was a note about error bands stating that a school's true index score would be expected to fall within a 3 point band around the reported score about 2/3 of the time. The other was a note that the 1997 tests were administered under different conditions, including the grade shift, for the first time in 1997.

The changed-administration conditions announcement is largely unhelpful to KY citizens, since it does not go beyond the disclaimer to discuss expected or known implications of the shift.

The "error-band" disclaimer is largely unhelpful as well, because it does not describe any implications. The error-band announcement in the 1997 KIRIS score report implies potential issues that should be examined and laid out for KY citizens. First, by customary statistical standards, the statement implies that a 95 percent confidence interval for school index scores would extend to about plus or minus 3 school index points – a 6 point error band.

Given the pattern of small changes in KY scores between 1995 and 1996 (the scores most recent as the error band was being estimated and its implications considered), this 6 point error band should be held up against change scores for 1995-96 to reveal the percentage of school index score changes that were at least as likely to have been random as achievement-generated. The test

score change reports for 1996 indicate that about 38.5 percent of KY schools reported changes in KIRIS index scores ranging between -3 and +3. An added 43.4 percent of schools showed declines between 1995 and 1996 exceeding 3 points. And about 18.1 percent of schools showed increases greater than 3 points. We acknowledge that KDE and ASME have been more attentive to school classification error possibilities than to more general error in reported scores for all schools. KDE and ASME also, appropriately, devoted analytic energy to the error-proneness of score reports for small schools.

Note that if errors across the confidence interval are random, the percentages of schools gaining and declining should remain the same if true scores were known . But according to published gain scores, there will simply be numbers of schools near the cut-offs of the confidence interval which are misclassified as gainers and losers.

Because of dramatic score increases overall reported for 1997, the change figures for this year have a different look when held up to the error-band or confidence interval. Between 1996 and 1997, about 30.2 percent of KY schools had changes in index scores falling between -3 and + 3. About 7.1 percent of schools declined by more than 3 points, and another 62.7 percent of schools gained by more than 3 points.

While the confidence interval is mentioned in passing in the 1997 KIRIS score reports, there is no explanation in the 1997 announcement of KIRIS scores that many KY school change scores would not be considered statistically different from zero according to the standards outlined in any introductory statistical inference textbook.

If the true 95 percent confidence interval is wider than 6 points, then the numbers of schools for which changes might be considered non-significant would be higher. This is a distinct possibility given the uncertainties of estimating a composite error band on a test such as KIRIS. A larger than reported true confidence interval was suggested by the OEA panel which claimed that estimated confidence intervals took only student cohort variation into account and did not make allowances for task variance or scorer instability. Kingston and Dings, 1995, provide an analysis and projection based on a simulation that includes task variance. We consider the science of estimating KIRIS error bands to be suggestive and not well established by precedent.

5. **Construct validity of change scores.** A testing system should report primary gain or change scores in ways that responsibly reflect estimates of changes in student achievement levels.

This means that when KY reports changes in middle school science index scores amounting to 27 percent or high school reading score gains of 61 percent, that KDE should promote guidelines facilitating reasonable interpretations of such statistics. We show analyses in Section 7 of this report indicating that the actual student performance level changes required to precipitate given percentage increases in index scores for academic subjects can be reasonably thought to be considerably lower than the

percentage changes in index scores reported. We begin with a reasonable set of assumptions about individual student KIRIS scores and the student achievement changes required to boost students across performance classifications enough to produce the 61 percent Grade 11 reading index gain reported. We then calculated that student achievement levels would need to advance by just less than 24 percent -- this assuming all of the gain was due to student achievement effects. Under an extreme set of assumptions about achievement distributions and their change from 1996 to 1997, the 61 percent reading change could have been prompted by a reading achievement change of less than 5 percent!

KIRIS thus appears to have a tendency to promote unrealistic impressions of achievement gains by its central and almost exclusive focus on its artificial school index score.

Even if the test were technically perfect, with all measures on target, valid, and reliable, KIRIS would have this problem. When index scores in reading go up by 15 percent and CTBS or NAEP scores only go up by 3 percent, there appears to be a huge gulf between the two tests. Our analyses suggest that these two hypothetical results could in fact be fairly equivalent on the basis of the achievement changes needed to produce such results -- yet there are no translations or discussions offered in score reports to sort out the inevitable confusion.

KIRIS is thus set up for accusations of inflated score changes under the very best of circumstances. Only changes in score interpretation and primary reporting will fix this problem.

6. **Construct validity of classifications.** A testing system should use classifications in describing the status of schools or individuals that meaningfully convey the performance levels and categories of performance levels determined by test results.

KIRIS has established a system of goals for schools based on overlapping accountability cycles. Schools meeting high performance criteria against their targets are rewarded. Schools declining against targets sufficiently are sanctioned.

KIRIS at present has documented neither strong enough ties between student performance and school index scores, nor strong enough ties between changes in school instructional practices and increased performance. In the absence of such evidence, and especially in the absence of the latter evidence, the school reward and sanction classifications are suspect and currently allocated rewards and sanctions appear to have an indefensible spurious quality.

We also take issue with a particular feature of the school classification scheme -- the arbitrary designation as "In-Crisis" any schools falling under its

essentially mechanical definition. More particularly, we doubt that very high performing schools in a given Accountability Cycle who slip to levels bringing an "In Crisis" label should be called "In-crisis." Such schools may remain among the highest performing schools in the state. Their scores probably suffered from what statisticians call "regression to the mean." In essence, this means that many of the highest (and lowest) performing schools in any distribution at any given time are not expected to remain there over time -- events usually conspire for some schools in a testing cycle to the point where most things that could have gone right did go right, and for other schools to the point that most things that could have gone wrong did go wrong, thus accounting for some of the extreme performers. Next year becomes a different story, with some outliers gravitating toward the mean performance level. KDE officials express reasons for wanting to keep up pressure on all schools to increase performance, but the in-crisis designation for very high performing schools that fall back in a given cycle seems unjustifiable. Its consequences for school staff (formal probation, for example; and takeover by a Distinguished Educator) seem both unwarranted and also unjustly critical of the qualifications of the educators involved.

Note. This general problem also visits the ability of KIRIS to classify schools as Reward at the very low end of the spectrum. For analogous reasons, much growth in schools by the very lowest performing schools expected in a given cycle may be caused by regression upward to the mean.

Note also: The use of two year averages for baseline and accountability scores tends to temper, but not eliminate, the degree to which regression to the mean effects should be considered a problem for interpreting score gains and declines at the extremes of the performance band.

<u>At very least, KIRIS should review its criteria for "In Crisis" school designations at the high end of school performance bands.</u>

Summary Response to the Main Question

We believe the KIRIS cognitive assessments as presently constituted and implemented are marginally adequate for reporting back to schools -- although portions such as the KIRIS writing portfolios are somewhat suspect with respect to psychometric characteristics. By this we mean that KIRIS scores bear plausible but not technically defensible relations to student and school performance levels and should be considered and used only in the context of the many limitations we identify in this report. We do not think KDE or its contractors have produced evidence tying gains or losses in school index scores -- the primary arbiters of reward and sanction under KIRIS -- to specific causes; most important, there is a shortage of evidence that KIRIS school scores change mainly in response to instructional initiatives and efforts designed to enhance student performance. We also believe that the accountability index is itself flawed (see for example, Cunningham's example of a school being both best and worst), and that we would be uncomfortable giving schools rewards and sanctions based on measures with such limiting characteristics.

KIRIS lies at an important junction in its evolution and three options seem available at this time:

1) Stay the course by continuing KIRIS substantially as-is;

2) Abandon KIRIS and install a traditional assessment program using commercially available standardized tests;

or 3) Continue KIRIS with changes based on the recommendations of this report.

The first option -- staying the present the course with KIRIS -- would perpetuate the dissemination of indefensible information about school and student performance. The second alternative -- abandoning KIRIS and using an off-the-shelf standardized test instrument -- also seems unwise at this time. The Commonwealth of Kentucky has made a substantial investment in KIRIS-related curriculum and instruction, as well as in assessment-related professional development, materials, and procedures. It seems premature and unnecessary to abandon this investment at this time, since weaknesses in KIRIS assessments seem to have remedies that can be applied and evaluated before a drastic action is considered.

We recommend that Kentucky should seek to capitalize on its investment in its unique, evolving, and nationally visible assessment and accountability system by learning from its accumulated experiences and making appropriate modifications in order to minimize the technical limitations we discuss. We make this recommendation in the full realization that KIRIS is vastly more expensive than a commercial test approach to statewide student assessment; thus the validity and consequences of the assessment and accountability system must be shown to be more positive if the KIRIS program is to prove cost-effective.

We turn now to the eight specific recommendations of the OEA panel we considered in our review of KIRIS.

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

Writing portfolios continue to be used at grades 4, 8, and 11 and are included in the school accountability index scores. Mathematics portfolios have not been used as part of accountability indexes and are under consideration for future use, pending working out of technical difficulties in their scoring.

Writing portfolios continue to be scored by teachers at test-takers' own schools. These scores are subject to audit. Schools showing large gains in writing portfolio scores from year to year are audited; additional schools are randomly selected for audit each year. Scores are adjusted if audits find them out of line with trained expert scores.

NB: OEA Panel Report Recommendations 2, 3, 7, and 8 were not included in our scope of work.

Early audits of portfolio scores in 1992-94 showed systematic and often dramatic differences between teacher scores and external expert scores. The 1996 writing portfolio audit suggests that the gap is narrowing to reasonable levels of agreement between teacher scores and expert scores, even though an inflationary bias ranging between 5 and 10 percent in the writing index score remains. A 1997 writing portfolio scoring audit showed gaps in a similar range.

Field studies show both positive and negative effects of the inclusion of writing portfolios in KIRIS – they are time consuming and subject to high levels of structure and proscriptive detail when it comes to the kinds of feedback teachers can provide students. A 1996 code of ethics for teachers adopted since the OEA report provides guidelines for work with portfolios.

The general attention to student writing brought by the writing portfolio component of KIRIS and the overall open-response design of KIRIS are considerable pluses in the minds of many KY educators – and on balance are seen as potential pluses to this review team to the extent that they are documented over time.

Professional judgments have varied regarding the weighing of the benefits and drawbacks of portfolios in KIRIS. Some professionals (e.g. Haertel and Wiley) wish portfolios to be kept. Others, such as the OEA Panel, recommend removal of portfolios from the accountability index portion of the assessment. Our team is not unanimous with respect to a recommendation.

If writing portfolios are to remain a part of KIRIS and the accountability index, efforts to bring writing portfolio scoring into closer line with what would be expected to result from independent scoring should be considered. One possibility is that the audit mechanism should be reviewed and its effectiveness in keeping scores in line with independent scores enhanced – perhaps with added penalties for discovered score inflation or more general adjustments of all scores based on annual audits. Another possibility is to enlist independent scorers, external to each school, as part of teams scoring all portfolios.

OEA 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.

Certainly the original OEA recommendation was justified. There did need to be more validation work on the assessments. Additional construct validation evidence was needed to support the various uses of the data.

Issues related to validity are immersed throughout this document. For example, the ability to equate well across forms within a year and across years and the use of local district teachers to score the portfolios are related to validity and are covered in other sections. Therefore, this section should not be read in isolation.

Nevertheless, our summary conclusion is that we do not believe enough validation research has yet taken place. Given our understanding of what validity evidence has been gathered since the OEA report, we believe more should have been done in the interim and much more remains to be done. The April 1997 draft technical manual was sketchy in its presentation of validity evidence, and perhaps could be described as a bit self-serving. However, some work has been done – e.g. the content validity work done by Nitko reported on in Section 5. Moreover, there is no philosophical opposition from the KDE to gathering more validity evidence. Indeed, they requested HumRRO to prepare a validity research plan. While that plan did not contain detailed research designs, it did present a list of research questions that should be investigated. We note, somewhat pessimistically, that many of these questions should have ideally been answered prior to applying rewards and sanctions. However, on an optimistic note, the planning document exists. Given sufficient commitment of time and money from the KDE, these questions can be addressed at a more comprehensive level than they have been to date.

We add one construct validity concern of our own to this discussion. This is the validity of calling "In-Crisis" very high performing schools whose Accountability Index Scores decline during an accountability cycle. Schools remaining at extremely high "performance" levels under the Accountability Index formulas can be designated as "In Crisis" under existing rules, and we question the construct validity of that designation.

OEA 5: The design for equating assessments should be strengthened and the ad hoc procedures eliminated. The DoE 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.

KIRIS is an assessment system that was set in place very rapidly and continues to evolve. During Cycle I, in particular, and also in Cycle II, a variety of technical decisions were made to get KIRIS up and running. Hambleton et al. characterized a number of the decisions as "ad hoc" and arbitrary, a characterization that seems reasonable and fair. The "ad hoc" decisions made to support KIRIS psychometric procedures in Cycles I and II were made to help initiate an innovative program and to support the program in its formative years. The choice of technical procedures used in the early phases of KIRIS represent a decision to take liberties with conventional technical procedures and criteria to initiate and nurture an unconventional program.

The once "ad hoc" decision rules used in equating open-response items have been used consistently over the life of the KIRIS program. The "*ad hoc-ness*" of these procedures and associated decision rules certainly diminishes with repeated use and these procedures now seem to add to the argument that consistent procedures are being employed.

KIRIS seems to have been responsive to the recommendations of Hambleton et al. in several ways and two, in particular, stand out: a) the inclusion of multiple-choice items; b) OEA's implied advice to not use performance events in the absence of better equating evidence. Multiple-choice items, however, are not yet being used in the way that Hambleton et al. recommended, although groundwork for such use was laid by including MC items in the 1997 tests; and the decision to drop performance events was made only after several attempts to justify the equating of these assessments proved inadequate.

The technical procedures used to support KIRIS are fundamentally the same as those decided upon at the very beginning of the program. There has been some tinkering and refinement of these procedures, but there has been no exploration of any approaches that examine the technical issues from a fundamentally different perspective. The major specific technical aspect of KIRIS that could be and should be improved is increased rigor in checking the precision and stability of the equating procedures. A variety of procedures that are more powerful in checking the adequacy of linking and equating are available and should be explored. Included in the consideration of this issue should be the possibility that, in at least some cases, parallel analyses would be conducted using different approaches as a means of cross-validation.

The KIRIS relationship between educational policy decisions and technical support procedures needs to be changed. To date, policy decisions seem to

have been made with the assumption that some psychometric procedure existed or could be devised to support any policy decision. Those responsible for the technical aspects of KIRIS have been very conscientious and resourceful in devising technical procedures that produce results that are reasonable or "plausible" enough to support the decisions. This post hoc use of technology to support rather than inform policy decisions may be appropriate early in the development of an innovative educational program to insure that the demands for technical rigor do not stifle innovation. KIRIS, however, has had sufficient time to develop and, at this point in the program, technical procedures should be used in two ways: 1) to inform policy decisions before they are made; 2) to evaluate the adequacy and precision of technical procedures that are already in place.

The apparent resistance of KIRIS policy makers to recognize the need to re-establish baseline years for scaling, equating, and standard setting is difficult to understand. Many statewide assessment programs and other assessment programs have reset their baseline year or re-normed their scales after a period of implementation. The task of establishing a new baseline year may have its challenges but it is not overwhelming. Given the long chain of adjustments and equating links that holds the KIRIS program together, it would seem timely and prudent to begin the process of re-establishing a baseline year and re-setting standards as soon as possible.

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

We agree with the substance of the criticisms of the original standard setting raised by the OEA panel. Subsequent to their report, two standards-related studies were conducted. These include a 1996 Standards Setting Study for Arts, Humanities, and Practical Living and a 1996 Standards Validation Study in Math, Reading, Science, and Social Studies (both authored by ASME). These studies are significantly but only partly responsive to the recommendation of the OEA Report. Standards were not reset, but were "validated" in math, reading, science, and social studies. The validation was not without merit nor logic -- and we discuss its approach and results in Section 6 of this report. But this response is not what was recommended by the OEA panel.

We do not conclude with any specific recommendations regarding future standard setting. One problem is that any improvement or re-doing of the standard setting means instability in the program across years. There is some delicate balance between improving a complex testing system over time and keeping some stability. We are not convinced that the KY program has found that balance, and if pressed for judgment we would say that lack of stability has been a substantial problem for KIRIS.

OEA 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.

Across our review, we detected little activity that was or could be formally labeled "auditing" of aspects of KIRIS. The one exception is the annual audit of writing portfolio scores reported for 1996 and 1997.

The National Technical Working Group serves as an informal auditing board, according to our reading of its functions and activities. We had access to their meeting minutes and procured many discussion papers, memoranda, and data files from equating studies discussed at NTWG meetings.

These meetings seemed to produce open and occasionally critical discussions of developments and issues surrounding KIRIS and to be genuinely concerned with, and active in resolving, important issues of technical quality. The meetings were typically attended by the NWTG members, senior ASME officials, and KDE assessment leaders. At the same time, these meetings seemed occasionally overly-focused on defending KIRIS from its various critics over the years, including the OEA report.

We learned through written materials and interviews about the various key processes of item development , testing, scoring, equating, and so on. But we learned of no procedures that could be called auditing, or third-party review or observation, of these processes.

Auditing in general seems an area where KIRIS has been unresponsive to the OEA panel report, and an area where KDE and NTWG might give some thought to where investments in auditing resources could contribute to the integrity of the testing system.

Potential audit targets include:

Increasing Use of Sub-Contractors. A target ripe for auditing lies in the many KIRIS-related activities taken on by the increased number of outside sub-contractors in recent years. When we walked through the 1996-97 and final 2-month 1997-98 ASME contracts with ASME officials, we learned that a sizable number of tasks contracted between KDE and ASME had been sub-contracted to other firms, including Data Recognition Corporation, HumRRO, and Allegiant. These tasks included major open-response item scoring contracts (all of Grade 4-5 and Grade 7-8 tests were scored by DRC in 1997, for example). We also learned from 1997 ASME scorers about science test items

developed by outside contractors. The 1996 contracted validity research plan was taken on by HumRRO.

If KDE expects to manage a system as large and complex as KIRIS, and to have major chunks of the work handed over by its contractors to sub-contractors across the nation during the first year of a major contract, then it would appear that KDE should step up its efforts at quality assurance in so far flung a system through a planned program of systematic audits and observations of practices. If all of KY's 4th grade science tests are being graded in Minnetonka, MN over a three week period in July, a KDE representative should drop in once in awhile and review the proceedings – and file a written report.

Test Score Gains. The need to audit observed gains in KIRIS scores specifically mentioned by the OEA Panel became something of a moot point in the 2 years following their report – that is, up until the issuance of the 1997 scores. KIRIS scores went through a relatively flat period between 94-95 and 95-96 – with more declining scores than advancing scores at the school and subject area levels. The 1997 scores provide new reasons to consider auditing test score increases. More than 78 percent of KY schools increased their index scores between the Cycle 3 baseline (1994-96 averaged together) and 1997. Moreover, more than a third of KY schools met their entire 2 year growth target in one year and nearly 7 percent of schools exceeded their growth targets by 100 percent. Several subject area scores increased phenomenally – High School reading by more than 60 percent and Middle School science by nearly 30 percent, for example.

We discuss 1997 scores at more length in Section 7 of this report.

OEA 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.**

Multiple choice items were added to open response items in subject matter tests for 1996-97. These items were not used in calculating student or school index scores for 1996-97, but are being considered for use in Accountability Cycle 4, for which 1997 and 1998 scores will serve as a baseline. KIRIS has thus responded to this recommendation of the OEA panel. We present some discussion of the multiple choice items in the section on equating and linking below.

In our analysis in Section 5, we refer to the controversy over multiple choice items as a debate between those believing multiple choice items to be "corrupters of curriculum" and others (such as the OEA Panel) who believe them to be the "saviors of technical rigor." This debate cannot not be established as matters of fact; these two views, in fact, simply represent differences in curricular philosophy and educational-political beliefs. The

KIRIS program, to date, has eschewed the use of multiple choice items for reasons of curricular validity despite the possibility that such items might add technical rigor and, possibly, even increased technical and curricular credibility. KIRIS is moving toward the use of multiple choice items with the 1997 test, though not for accountability purposes yet. This decision should not be evaluated as "right" or "wrong", but should be viewed on the continuum of being more or less useful in supporting the goals of the KIRIS program.

As a review team, we favor the inclusion of appropriately developed and selected multiple choice items in the cognitive tests for KIRIS as well as their inclusion as appropriate components of the school accountability index

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

Our review team was eventually able to find considerable documentation regarding equating and standard setting (or validation). The forms in which such information is kept, the somewhat ad hoc as well as prolific nature of its generation, and the lack of any cataloguing of this documentation are worth comment.

We began our search for documentation materials in early October, 1997. We received an outpouring of items from OEA and KDE, followed by more documents provided by ASME in Dover in early November. As we read through these documents, we discovered references to numerous studies, memos, discussion papers, and other documents that we needed to pursue our work. This generated another round of requests for materials which were generally forthcoming. But these new materials led us to new paper trails, and so on. A great many documents, including main KIRIS technical reports long since issued, appear only in forms clearly marked, "DRAFT."

Given the high stakes and visible nature of KIRIS, it seems inevitable that someone or some office in the system will have reason periodically to sponsor an outside review of important technical aspects of the testing system. There are many reasons to anticipate internal needs for information supportive of technical reviews also.

Keep a comprehensive, organized documentation library. Under such circumstances. we earnestly recommend that KDE keep an intact, organized library of technical documents in an accessible location. This library would include both special studies as well as the numerous reports related to questions of equating tests from year to year and across forms.

Produce an intermediate technical manual each year. We also suggest that KDE produce or sponsor creation of an annual KIRIS technical information manual written for non-experts who nonetheless have important interests in the technical quality of the testing system. This audience includes legislative

leaders, KDE members, local school board members and educators, and citizens who take a strong interest in KIRIS for various reasons. This manual would discuss any special studies (such as grade shifting studies or validity research plans) as well as equating data showing just what adjustments were being made in which tests for what reasons each year. And the manual should have a parallel form each year, with a special edition at the completion of each accountability cycle.

The major sections below covering standard setting, validation, and equating and linking offer additional specific comments about documentation in these areas.

OEA 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, the accuracy of the estimates of change upon which KIRIS focuses—be more fully evaluated.**

There appear to be two important components lying within this recommendation. One is the possible need to build more content knowledge into KIRIS assessment tasks, which were criticized sharply by the OEA Panel for being lacking in such content. The second addresses the validity of claims that KIRIS produces various positive instructional effects, claims that must be subjected to ongoing empirical scrutiny if they are to serve as a foundation for KIRIS.

Measuring more content. KDE responded to the issues addressed in this recommendation by the OEA Panel. KDE moved in 1996 to produce a more detailed specification of subject matter content for KIRIS testing in a report called the Core Content for Instruction (KDE, 1996). An apparent goal of this document, itself based on previous content specification documents such as the "Valued Outcomes" which had evolved since the start of KIRIS, was to ensure that specific skills and content knowledge were being incorporated into KIRIS open response items. This would mean, for example, that science items would require some recall and reporting of specific scientific content, or that math items would require knowing specific formulas, whereas earlier tests seemed to be more concerned with quality of writing, reasoning, and application and less concerned with specific subject knowledge.

Studies of instructional effects. Recent studies, such as Koretz et al. (1996) have appeared since the OEA panel report and which shed light on the instructional consequences of KIRIS. These are discussed at some length in Section 8. To us, the findings of these studies are mixed. There is no question that KIRIS has raised attention to questions of instruction and curriculum in KY schools. KIRIS is a high stakes system and schools, teachers, parents, and public

officials are more conscious of its workings and implications -- and *report* more workings and implications -- than we would expect to see in most other state testing systems in our awareness. This is a plus. The main counterbalances seem to be burden on teachers and students, and over-concentration in instruction on student writing at the expense of subject area content (because of the open-response test format).

Ongoing study of the effects of KIRIS seem absolutely important. At many turns, the sponsors of KIRIS -- KDE, ASME, the National Technical Working Group, offer reactions to technical concerns with KIRIS tests that effectively claim that the results are worth the limitations. The OEA Panel, and to some degree this 1998 review of KIRIS find little difficulty pointing to important technical limitations. If curricular and instructional effect is to be the response, the burden is on the responder to document such effects.

Summary of Recommendations

Recommendation 1: Bring consistency to KIRIS. Beginning with test year 1996-97, maintain a consistent structure of test components, grade-level testing distribution, and test administration regulations.

Recommendation 2: Begin a new accountability clock. If a commitment is made to maintain a consistent testing program corresponding to the structure implemented in 1996-97, use test year 1996-97 as the first year of a new accountability cycle. Suspend the accountability process for Cycle 3 and use accumulated test results for information only.

Recommendation 3: Begin new cycle of equating analyses to accommodate this new accountability cycle. Use more equating procedures for which their adequacy and precision can be tested rigorously.

Recommendation 4: Re-set standards following established procedures to accompany the start of the new baseline.

Recommendation 5: Publish more complete and informative guides for interpreting school accountability index score changes. Include information about estimated error of the accountability index as well as information about connections between index score changes and estimated changes in student performance levels.

Recommendation 6: Review the basis of "Reward" and "In-crisis" school accountability classifications to assure their construct validity in all cases where they are applied.

Recommendation 7: If writing portfolios remain in the school accountability index, maintain and strengthen the annual audit of portfolio scores in ways that serve to minimize error between teacher-produced scores and audit-generated scores.

Recommendation 8: Maintain the program of development and inclusion of multiple choice items in the cognitive tests for KIRIS and in the accountability index.

Recommendation 9: Proceed to enact the elements of the Validity Research Plan developed for KIRIS.

Recommendation 10: Establish more routine audits of key processes engaged by KIRIS.

Recommendation 11: Maintain and catalog a library of technical documents related to KIRIS for internal and external review purposes. Produce an annual technical report for audiences including educators, testing coordinators, parents, and legislative leaders.

Recommendation 12: Maintain a vigorous ongoing program of research and documentation of the effects of KIRIS in Kentucky schools.

We believe our technical assessment of KIRIS and the recommendations summarized above generally reflect appraisals to be expected in a mid-course evaluation of the complex, comprehensive, and developing assessment and accountability system embodied by KIRIS. Recent developments discussed in this report do reflect responsiveness and even greater sympathy on the part of KDE to the recommendations of the OEA panel; there is also much to be done.

We turn now to seven analytic sections which address the OEA Panel recommendations respectively in more depth and providing added background and rationale for our recommendations.

Section 3. The Role of Portfolios in KIRIS

The OEA Panel's Recommendation 1 urged that portfolios not be included in the KIRIS accountability index.

Status of Recommendation: Since 1995, writing portfolios at grades 4, 8, and 12 remain a part of the accountability index. These portfolios continue to be scored annually by teachers at the schools generating KIRIS writing portfolios. Mathematics portfolios, while contemplated for inclusion, are not presently used as part of the school accountability index. Audits of writing portfolio scoring in 1996 and 1997 brought additional data (and some relief) to concerns regarding upward bias in the scoring of writing portfolios at their own children's schools.

So some observers, portfolio assessment captures a vision of educational reform by integrating instruction with assessment. Portfolio advocates have long argued that a student's classroom work and the accompanying student reflection on that work provides a truer picture of a student's competencies than do traditional forms of assessment. Advocates also point out the impact portfolios have on challenging teachers and students to focus on "big ideas" and meaningful outcomes. Portfolios also support the assessment of long term projects over time, and encourage students to show

what they have learned through various products and presentations suitable for portfolio assessment.

Given KERA's ambitious agenda for instruction, assessment, and accountability, it is not surprising that Kentucky included portfolio assessments as part of KIRIS. However, questions persist over the role and impact portfolios should have in the KIRIS accountability process and in the school accountability index. Most concerns have centered on the reliability of portfolio scores, and the implications of unreliability for the continuing use of portfolios in KIRIS. The focus of this review is to evaluate the recommendations set forth in the report from the Office of Education Accountability (OEA) (Hambleton et al., 1995) and specifically the OEA Panel recommendation that portfolios should not be used in the accountability index.

Portfolios In The Kentucky School Accountability Index

Kentucky school's students are measured on cognitive and non-cognitive outcomes; the data are then calculated to provide each school's performance. Cognitive data are measures of student academic performances in seven areas; mathematics, reading, science, social studies, and writing, each weighted 14% of total score, and Art & Humanities, and Practical Living/Vocational Studies – each weighted 7%. Cognitive measures collectively represent 84% of the accountability index. Non-cognitive data are non-academic measures of school performance comprising attendance rates, retention rates, dropout rates, successful transition to adult life. Non-academic data comprise 16% of the accountability index.

Cognitive data are presently derived from "on-demand tests" comprised of open-response items in various subjects and essay responses derived from writing prompts, as well as through scoring "portfolios" of student writing. The portfolio score is based solely on one item; the portfolio itself.

Writing and mathematics portfolios were collected from students as part of KIRIS, but only writing portfolios have been a part of the accountability score for each school since 1991-92. (A mathematics portfolio was administered at grades 8 and 12 for accountability purposes during the second accountability cycle, 1992-93 through 1995-96, but math portfolios were withdrawn from formal accountability mechanisms in order to undergo further research and development. The mathematics portfolio is scheduled to be reinstalled during the 1998-99 school year.)

Students in grades 4, 8, and 12 were assessed with writing portfolios containing a collection of writing from one school year. State guidelines describe the types of items to be included, and the portfolios are scored by teachers in that school. KDE conducts random audits using trained readers, e.g. other teachers, to validate scores. Each portfolio is given an overall rating of novice, apprentice, proficient, or distinguished, rather than scoring each individual piece of student work separately.

Concern With The Use Of Portfolios

The OEA report (Hambleton et al., 1995) clearly recommended against the use of portfolios in the accountability index, listing this as the first of its twelve

recommendations. The basis of this recommendation is what the OEA Panel report describes as the lack of evidence about the **reliability of the KIRIS portfolio scores** (p. 4-3). This report and the recommendations are focused mostly on psychometric issues in scoring, though concerns are noted about the operation of the KIRIS portfolio program in general.

A major concern with the operation of the portfolio program is the "highly unstandardized" aspect of its operation (p. 4-6). The report criticizes the permissive guidelines offered to teachers; the way in which scoring is conducted at school sites; the pre-scoring training model; and the ambiguity of the scoring rubrics. While the OEA report recommends against the use of portfolios for both psychometric and operational issues, a review of the operational issues is necessary before discussion of psychometric questions.

Operational Issues. Hambleton et al. criticized the general guidelines of the portfolio program for its vagueness concerning directing teachers regarding which pieces of student writing to include. The KIRIS Accountability Cycle II Technical Manual, April 1997 specifically describes the KIRIS writing content requirements in Table 3-19, (p. 3-23). Similar evidence is found in the District Assessment Coordinator Implementation Guides (1995, 1997), as well as in the Kentucky Writing Teacher's Handbook. In these works, specific requirements for grades 4, 8, and 12 guide the number and possible types of writing to be included. Hambleton et al., criticize the "substantial leeway" in terms of types of products portfolios may contain. However, while there is permitted variability in product selection, (i.e., directing teachers to select two pieces that either predict an outcome, defend a position, analyze a situation, solve a problem, explain a process, draw a conclusion, create a model), there is also specificity: teachers are directed to select one personal narrative, and one short story, poem, play/script or other piece of fiction.

What may have been overlooked in the criticism of product selection articulated by the OEA Panel, or underestimated in its evaluation of the portfolio, is the focus on the "quality criteria" and the pertinent academic expectation the portfolio program is designed to elicit. The purpose of the writing portfolio is for the student to "demonstrate the ability to communicate through writing their ideas and information to a variety of audiences for a variety of purposes in a variety of written forms" (technical manual). This standard of academic expectation allows for variety in product selection. The "quality criteria" used to judge product selection is to evaluate the students' performance in the areas of Correctness (spelling, punctuation, capitalization), Language (usage, word choice), Sentences (correctness, structure, effectiveness), Organization (logic, coherence, transitions), Purpose, and Audience.

While the OEA Panel Report is helpful in pointing out the need to clarify the articulation of standards applied in judging best work versus progress versus process, it is less helpful in revisiting the original purpose of the portfolio program in evaluating student work.

Another concern was expressed by the OEA Panel about factors that may vary across students and across programs; such as initial instructions from teachers; the amount of time allowed; the use and amount of pre-teaching tasks using similar formats and prompts; and the amount of assistance provided by teachers, parents, peers, and others. Together these concerns point to the question of "who's work is it?" that is contained in a portfolio and thus driving portfolio scores. This is not a new question in the evaluation of portfolios and expressions of such concerns are spread across the

literature (Condon & Hamp-Lynos, 1991; Herman, et al., 1993; Koretz, McCaffrey, Klein, Bell, & Stecher, 1993; Stecher & Hamilton, 1994). But it remains a valid question.

New Code of Ethics. At the time the OEA Panel deliberated and composed its 1995 report, the KIRIS portfolio schemes were vague on the key questions such as the amounts of time allowed for development of items and directions from teachers for initiating or revising student work. However, the "Code of Ethics for Appropriate Testing Practices for School and District Personnel" (1995, 1997), lists specific ways to establish ethical working relationships between teachers and students over portfolio related work.

This Code of Ethics specifies the student "as the sole creator, author, and owner" of the work, and by definition establishes the portfolio as "an exhibition of an individual's achievements." Clearly stated is the proviso that any assistance or intervention should be for the purpose of encouraging the student to make choices that strengthen rather than diminish the ownership. The Code of Ethics guidelines recommend that teachers:

* provide supportive environments;
* merge instruction and assessment activities;
* allow time for drafting, developing, and revising student work;
* ensure student ownership through revision activities; and
* discuss and model processes and strategies to develop writing.

One specific guideline is directed at the role teachers are to play in the revision process. Specifically, "Teachers shall not at any time actually make corrections and revisions on a student's work; teachers may guide students by noting or asking questions about errors or possible improvements in the students' work, rather than by making direct corrections."

These guidelines, along with the *KIRIS Teacher Certification of Appropriate Testing Practices,* (a document teachers must sign) go far in bringing uniformity to the portfolio program without dictating statewide the actual assignments, allowable time, working conditions, and teacher instructions or verbal directions. Certainly, other such guidelines and program recommendations from KDE can assist in the reduction of concerns over program administration. Guidelines about the use of time, and the types of assignments and the conditions and context for student work development may be of assistance. Suggestions for portfolio procedures have been made before (Gearhart and Herman, 1995, Webb, 1994). Such standardization may very well bring about a greater conformity and provide less concern over score variance (a major concern to be addressed), but great care and caution would need to be exercised to avoid deviation from the natural and expected role of teacher-student interaction involved in teaching and learning. Recent surveys of teachers and school administrators (Wilkerson and Associates, 1996; Koretz et al., 1996; AEL 1994) have acknowledged the difficulty and increased stress brought on by the use of portfolios. Most notable are concerns about the amount of time required to administer the portfolio program, the burdensome nature of portfolios, and perceived diversion of instructional attention away from the mechanics of writing. Despite distinct concerns regarding administrative aspects of the portfolio program, responses acknowledge the value of the portfolios in KERA, and in increasing the opportunities for students to write, demonstrate students' improvements in writing, and in improving students' thinking skills. There are indications in these studies that portfolio assessment is having a desired effect on the instructional practices of teachers in Kentucky.

It is imperative that Kentucky maintain a program of auditing or regular research to provide a continuous assessment of the impact of portfolios in the field. Because it is only evidence of considerable positive instructional impact that would support the continued inclusion of portfolios in KIRIS, given some technical concerns to which we now turn.

Measurement Errors And Concerns

The portfolio program is not without measurement error. It is precisely this issue that prompted the OEA Panel recommendation to ban portfolio scores from the Accountability Index. Some of the measurement concerns are generic to portfolio assessments in general and others specific to the Kentucky program. The operations guidelines and new Code of Ethics described above are likely to reduce some forms of error. However, general concerns over the use of portfolios in any accountability system will inevitably flow from the portfolio's shortcomings as a measurement instrument. In Kentucky, these shortcomings have been recognized by the KIRIS contractor, the technical advisors, and the Kentucky Department of Education. Despite these shortcomings, it is the estimation of these sponsors that the influence of high stakes portfolio-based assessment on instruction is likely to be sufficiently beneficial to outweigh the disadvantages (Haertel and Wiley, Response to the OEA Panel Report, 1995; Haertel, personal communication, August, 1995). Similar recommendations were made by an Advanced Systems appraisal (Awbrey, 1996). As we discuss in Section 8, the evidence on whether effects are positive or negative is equivocal.

Portfolio Audits

During the summer of 1993, a select group of Kentucky teachers was convened to re-score writing portfolios that had been originally scored throughout the state by teachers. A randomly selected sample of portfolios from schools throughout the state was scored. A second sample of portfolios from teachers whose scores were inconsistent with the rest of their district was scored. This special summer scoring took place in very structured conditions that did not replicate the scoring conditions under which the portfolios were originally scored by teachers. The general findings were that teachers were inconsistent in scoring portfolios. The results showed that the portfolio score assigned by an average teacher in Kentucky was considerably higher that the score assigned by teachers participating in re-scoring. The differences were a source of concern. However, while it was clear that the summer scores were giving portfolios lower scores than did Kentucky teachers from throughout the state, no evidence was collected to show that the differences indicated that the summer scores were more accurate. Measures were taken to assure consistent and accurate scoring of writing through a "writing audit" system. The purposes of the audit were to determine whether existing feasible measures for identifying schools with potential scoring difficulties were effective and to reinforce the message that the accountability system requires and insists on rigorous application of scoring standards by teachers. Procedures were instituted across the state to assist teachers in scoring protocol. Most dramatically, schools that had been identified as having the greatest discrepancies from the summer 1993 scoring later exhibited the most accurate scores a year later.

A similar audit was conducted in the summer of 1996 by KDE and its assessment contractor, ASME. The two purposes of the audit were to monitor the scoring accuracy of schools statewide, and to adjust scores for those schools for which scoring was discrepant. Recalling that last time auditing was conducted was during 1993, all schools by 1996 had received substantial scoring feedback via the 1994 and

1995 Writing Portfolio Scoring Analysis sessions. The selection list included 98 schools (originally 100, but two schools were removed for different reasons) -- 49 purposefully selected because of recent high score gains and 49 randomly selected. Scorers were trained using the Kentucky Writing Portfolio Teacher's Handbook, 2nd Edition with benchmarks, exemplars and high-end samples, along with the Kentucky Writing Portfolio Holistic Scoring Guide. Training procedures were the same as those used to train teachers in Kentucky during the 1995-96 school year.

The results of the audit showed that schools across the state had increased their accuracy in scoring writing portfolios. Schools that were purposefully selected demonstrated a 73% rate of exact agreement. Randomly selected schools demonstrated a slightly higher rate of 77%. Rates of exact agreement ranged from 98% to 8%; however, 68 of the 98 schools demonstrated agreement rates of over 70% (Awbrey, 1996).

These differing rates of agreement resulted in different consequences for individual schools. The purposefully selected schools had a -10.89 change in the Writing Cognitive Index as compared to only a -5.16 change for randomly selected schools. Sixty-three of the ninety-eight schools included in the audit had changes of 10 points of less on their writing cognitive index.

When the results of the 1996 audit are compared to the results of the previous audit, the data show a substantial improvement in scoring. During the 1992-93 audit year, the average grade 4 change of index for the purposefully-selected schools was a -53.5, while the 1996 average audit change was -8.6. Similarly the grades 8 and 12 changes show respective substantial improvements -- a -40.6 in 1993 versus a -13.2 in 1996; and a -32.8 in 1993 versus -13.4 (Awbrey, 1996) These results demonstrate that with training and experience, teachers can improve their scoring accuracy. These results also demonstrate that self-scored writing portfolios do involve systematic inflation of true writing scores that must be monitored annually. A 1997 writing portfolio audit showed similar results to the 1996 audit, with slightly larger scoring discrepancies (about 10 percent) for randomly selected schools.

The District Assessment Coordinator Implementation Guide offers schools six options for conducting their writing portfolio site scoring. All six options begin with individual scores being assigned by the teacher, and five of the six options include a blind second score. However, all six options allow for the scorers to "meet to discuss the discrepancies and come to consensus" when scores are different. This process of discussion and consensus is much less formal and rigorous than the "read to resolution" process used in the 1996 audit. By the nature of the consensus process, it allows for more leniency and less accuracy than three or four blind scores, contributing to the lower coefficient of inter-rater agreement cited by the OEA Report.

As noted earlier, the potential shortcomings of portfolios as measurement instruments make them suspect in a high stakes assessment system. Given the concerns captured in the OEA Report, it is not surprising that Hambleton et al., would recommend the discontinuation of portfolios in the accountability index. However, the potential shortcomings of portfolios as measurement instruments make them suspect in a high stakes assessment system. Professional judgments have varied regarding the weighing of the benefits and drawbacks of portfolios in KIRIS. Some professionals (e.g. Haertel and Wiley) wish portfolios to be kept. Others, such as the OEA Panel, recommend removal of portfolios from the accountability index portion of the assessment. Indeed, our team is not unanimous with respect to a recommendation. Some prefer keeping the portfolio because of the purported instructional benefit; others

favor removing portfolios because of their psychometric characteristics. If portfolios are maintained as part of the accountability index, this is an area where meaningful auditing of portfolio scores in ways that encourage accurate scoring as well as high quality field documentation that verifies or falsifies alleged net benefits on instruction seem to be wise prescriptions for KIRIS.

Section 4: Assessing the Validity of KIRIS

OEA Panel Recommendation

The OEA Panel Recommendation Number 4 was that "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." Some more specific recommendations were also included within the OEA Panel report. For example, "Conduct or sponsor studies to investigate to what extent, if any, unwanted factors are being measured by the assessment tasks." (p. 2-3).

As the OEA report makes clear, because the KIRIS data are used both for accountability purposes and for the purpose of impacting curriculum and instruction, there are two different sets of criteria against which the assessments must be investigated: the traditional measurement quality of validity which has to do with the accuracy of the inferences made from the scores -- about both the individual students' levels of achievement and the school districts' assessments; and the consequences of any actions taken as a result of the assessment on curriculum and instruction.

Chapter 4 of the OEA Panel report raises specific questions about the validity of the KIRIS Writing Portfolio scores. The authors suggest that "evidence about the validity of the KIRIS portfolio scores is limited..." (p. 4-21). However, as they report, "some evidence pertaining to the validity of KIRIS portfolio scores is available, and taken together, it suggests skepticism about the validity of the scores." (p. 43-22).

The OEA Panel discussed whether gains in KIRIS represent real improvements in student learning. They concluded 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. ... 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." (p. 8-4).

KDE Response

The KDE (July, 1995) agreed with both the main recommendation to expand validation work and the more specific recommendation to conduct or sponsor studies to investigate what unwanted factors are being measured by the assessment tasks.

Reactions of the National Technical Working Group

Little could be found in the minutes of the NTWG regarding validation. The March 1-2 1996 minutes did have a section on Comments on Research Agenda Items. In that section they commented on the draft factor-analytic study of construct validity, a proposed consequential validity questionnaire, the proposed NAEP item re-administration, and a study on school decision consistency.

With respect to the factor validity study the minutes state that "this study appears more likely to provide evidence of factorial validity than evidence of construct validity per se, so the title should probably be revised for clarity. ... It would be desirable for this study to employ a confirmatory approach rather than simply an exploratory one."

The NTWG recommended that KDE not participate in the SCASS Technical Guidelines for Performance Assessment group validity questionnaire because of concerns they had about the instrument. The proposed NAEP item re-administration study was postponed until 1997.

In the December 13-14, 1996 minutes, it is reported that the NTWG received an update of the validity research agenda. The minutes of that meeting devote about one page to summarizing the discussion of the NTWG. The issues of validity research was again discussed in the March 14-15, 1997 minutes.

It seems clear that while there are some legitimate differences of opinion about whether certain research endeavors would be worthwhile and/or whether they are designed most effectively, both the KDE and the NTWG are, at an abstract level, in favor of additional validation studies.

KIRIS Accountability Cycle II Technical Manual

The draft Cycle II Technical Manual (April, 1997) has a Section devoted to Generalizability. In that section, there are separate chapters on reliability (Ch. 13), validity-related evidence (Ch. 14), and consequential validity (Ch. 15). Taken together, these chapters provide some data regarding reliability and validity. The chapter on validity-related evidence posits that

> "high consequential validity is probably not possible without high degrees of traditional reliability and validity, but high degrees of traditional reliability and validity cannot guarantee high consequential validity." (p 14-2).

We would certainly agree with this position.

The technical manual devotes two paragraphs to content-related validity evidence. Within these two paragraphs, there is reference made to the section of the manual that describes how the components of the KIRIS assessment were derived. The manual suggests that "there is substantive validity-related evidence in the process by which KIRIS assessments were constructed." (pp. 14-2 and 14-3). This seems like a reasonable statement. However, another statement is more troublesome: "Arguably, most performance events embody essential characteristics of cognitively complex tasks, requiring the student to demonstrate problem-solving." (p. 14-3). If it is arguable that most performance events do this, should there not have been some evidence gathered that these particular performance events do that? No new research is reported.

The technical manual does report on some research under the heading "construct-related validity evidence." In that section, they report on a factor analysis of the common open ended items in reading, mathematics, science, and social studies. Excluded from the analysis were matrix-sampled items in those areas as well as multiple choice items, performance events, the items from the arts and humanities and practical living/vocational studies, and the holistic scores from the writing and mathematics portfolios. The results showed that:

> "each rotated factor is clearly defined by loadings on one and only one KIRIS test. ... [however] it also appears that a general achievement factor may also underlie the data at each grade level. ... The amount of common variance explained by the first initial factor extracted at grade 8 was 82%. In addition, the inter-factor correlations among the four rotated factors ... were moderately high, again reflecting the general achievement factor. ... Results of the above factor analyses provide favorable evidence with regard to construct validity" (p. 14-5).

We would consider this extremely weak construct validity evidence and would note that the NTWG had recommended this not be called a construct validity study and that the factor analysis be confirmatory rather than exploratory.

The manual devotes two pages to concurrent validity-related evidence. This is mostly a reiteration of validity evidence reported in the Accountability Cycle I Technical Manual.

Chapter 15 of the draft technical manual reports on the "consequential validity" of KIRIS. We discuss that chapter more under a section of this report on "Instructional Consequences of KIRIS."

In general, we do not feel the evidence reported in the draft April, 1997 KIRIS Accountability Cycle II Technical Manual has been responsive to the request from the OEA panel for more validity evidence.

Other individuals' views of the validity of KIRIS

Cunningham has been quite critical of the validity of KIRIS. Some of his comments are quoted below:

> "KIRIS is valid if the schools that are in crisis are staffed by lazy incompetent teachers that need to be punished and reward schools are staffed by wonderfully dedicated teachers who are deserving of cash rewards. If you don't agree with these assertions then you are questioning the validity of KIRIS because the purpose of KIRIS is the allocation of rewards and sanctions." (Cunningham, no date, p. 2)

> "In Kentucky there is a tacit agreement that there will not be a state curriculum, there is therefore no instructional goals with which to compare the KIRIS items." (p. 3)

> "The best evidence for evaluating the criterion-related validity of KIRIS comes from the NAEP Trial State Assessments. In fourth grade reading the KIRIS scores increased 81 percent from 1992 to 1994 while the comparable NAEP scores dropped by a half percent. In math, between 1992 and 1996 KIRIS math scores increased 115 percent while NAEP scores increased by 2 percent." (p. 3).

> "The absence of any relationship between the two cycles is evidence for a lack of construct-related validity." (p. 5).

> "This system has created absurdities like Brown Middle school. ...which was in reward for Cycle 1 and had the third highest accountability score of any middle school in Jefferson County in Cycle 2. It has been labeled a crisis school and its teachers have been placed on probation and labeled incompetent. Anchorage school district is one of only two school districts in the state that have been labeled in decline, and of the two it is the most in decline. It also has the highest

> accountability score of any district in the state. Anchorage School District is therefore either the best school district in the state, because it has the highest accountability score or the worst because it has the lowest score in comparison with its baseline. Only under KIRIS can a school district simultaneously be the best and the worst in the state. These are only two examples, there are many more. What these examples do is illustrate the lack of validity of the KIRIS assessment." (p. 5).

Petrosko (May, 1997) has also raised issues about the validity of KIRIS. In reviewing the work of others he concludes by stating that

> "Evaluators have criticized several aspects of KIRIS including: (a) the uncertain relationship between 57 academic expectations and KIRIS items, (b) the validity of portfolio scores, (c) the reliability of the accountability index used to place schools into performance categories, and (d) the low degree of relationship between KIRIS and other outcome measures, especially ACT college placement scores." (p. 28).

Validity Research Plan for the Cognitive Component of the Kentucky Instructional Results Information System

Hoffman (September, 1997), an employee of Human Resources Research Organization (HumRRO), issued a report with the above title which had been prepared for the KDE. This is a 26 page document written as a "comment draft." This report is presented in three sections: documenting the intentions for the uses of KIRIS, test validity questions regarding the meaning and interpretation of KIRIS scores, and the systematic impact of KIRIS. In general, we think this is a well written report listing some important validity questions such as the following:

Are the contents of KIRIS consistent with intended contents?

Do KIRIS items elicit the kinds of performance intended?

How do KIRIS scoring processes affect the meaning and accuracy of student scores?

Are KIRIS scores appropriately related to other measures of educational achievement?

Are KIRIS scores appropriately related to other factors logically associated with educational achievement?

Are scores influenced by extraneous or inappropriate factors?

System impact questions are as follows:

Are assessment results impacting on instructional practices and student achievement as intended?

Are assumptions about the chain of relationships leading from the uses of KIRIS to the goals of KIRIS supported empirically?

Does use of assessment or accountability results have unintended consequences that detract from targeted objectives?

Do KIRIS assessments leave out important and intended areas of competence?

Are long-term systemic changes being stimulated?

The validity research plan concludes by suggesting that "This document presents a comprehensive statement of the intended uses of KIRIS and the research needed to confirm the validity of those uses. The purpose of the document is to provide a checklist for tracking and coordinating completed, on-going, and needed research." (p. 26).

Conclusions regarding OEA Panel Recommendation 4

Certainly the original OEA recommendation was justified. There did need to be more validation work on the assessments. Additional construct validation evidence was needed to support the various uses of the data.

Issues related to validity are immersed throughout this document. For example, the ability to equate well across forms within a year and across years and the use of local district teachers to score the portfolios are related to validity and are covered in other sections. Therefore, this section should not be read in isolation.

Nevertheless, our summary conclusion is that we do not believe enough validation research has yet taken place. Given our understanding of what validity evidence has been gathered since the OEA report, we believe more should have been done in the interim and much more remains to be done. The April 1997 draft technical manual was sketchy in its presentation of validity evidence, and perhaps could be described as a bit self-serving. However, some work has been done – e.g. the content validity work done by Nitko reported on elsewhere in this document. Moreover, there is no philosophical opposition from the KDE to gathering more validity evidence. Indeed, they requested HumRRO to prepare a validity research plan. While that plan did not contain detailed research designs, it did present a list of research questions that should be investigated. We note, somewhat pessimistically, that many of these questions should have ideally been answered prior to applying rewards and sanctions. However, on an optimistic note, the planning document exists. Given sufficient commitment of time and money from the KDE, these questions can be addressed at a more comprehensive level than they have been to date.

Section 5. Equating and Linking

1. Introduction

Section 5 of our report is presented in five subsections beginning with this introduction that describes the broader context for understanding KIRIS equating. The second part of this section provides a brief description and our review of equating procedures used in the various KIRIS components. The third section revisits and comments on the evaluation of KIRIS equating offered by Hambleton et al. (1995). This is followed by a brief section examining some other investigations related to KIRIS and KIRIS equating. In the fifth and final section we present twelve major findings regarding our evaluation of KIRIS equating based on a review of all available information.

Basic Issues in Equating

Equating or linking refers to a set of procedures used to determine the equivalence of students' performance on one assessment with their performance on another. Some form of equating is required unless all students in a given year or across several years take exactly the same test items/assessment tasks. There are many different procedures for equating and which one is used depends on the particulars of the specific assessment setting. Different types of equating may be used for different forms of assessment under varying circumstances.

Equating procedures for some forms of assessment are more well understood than for other assessment formats. For example, most measurement specialists can equate multiple-choice tests with any one of a number of different equating procedures. Furthermore, criteria for judging the adequacy and precision of equating with multiple-choice type items or tests are reasonably well established and can be demonstrated. Equating procedures for scores assigned to constructed responses, such as are found in open-ended assessments, performance tasks, and portfolios, are not as simple and well understood as are procedures for equating multiple-choice items. Measurement experts and educational practitioners have had limited experience in equating many forms of assessment advocated as part of several curriculum reform initiatives. Although psychometric or mathematical procedures for equating such items have been available for some time, practical experience in applying these procedures and in interpreting the results remains somewhat limited. It is further important to note that any number of procedures for equating data from constructed response items/tasks can be applied to such data but criteria for judging the adequacy and precision of the resulting "equating" are not well established.

A very important difference between multiple-choice items and open-ended items is the fact that multiple-choice items are designed to give repeated information about the same skill or ability. For example, a 40 item multiple-choice math test can be thought of as 40 "glimpses" of the same skill or ability. When taken together, the 40 multiple-choice items give a picture of the overall skill or ability, the stability of which can be assessed by looking at the consistency of the "40 glimpses." Data collected from constructed-response tasks are quite different. Students take many fewer such tasks in an assessment program and the assumption that all tasks reflect the same generalized ability is not supported by the research (Shavelson, Baxter & Gao, 1993).

When discussing equating and its practical implications, it becomes important to make a distinction between equating scores and interpreting the meaning of the scores. Equating procedures can be used to establish the relationship between the scores or

numbers on various open-ended tasks or assessments; the assertion that the "equated" scores reflect the same level or type of ability or skill, however, is not assured. In some instances it is more appropriate to talk about "linking scores " which means the relationship between the numeric values on two assessments are paired together. "Linking" does not imply that the paired (or linked) scores reflect the same level or type of ability.

Mislevy (1992) provides taxonomy for describing the various ways in which assessments can be linked. In Mislevy's framework, linking is the broad, generic term for pairing two scores together by variety of procedures. Claiming that two assessments are "equated" is a very strong claim. It posits that the assessments are written to the same content and format specifications. The foundation for establishing equating is not in statistical procedures but in the way the assessments are constructed. In this view, equating is more a substantive issue than an empirical demonstration. Mislevy has a category for linking assessments called "calibration". Tests that are linked via calibration are thought to measure the same thing but perhaps with different accuracy and in different ways. Results from different assessments " ...are mapped to a common variable, matching up the most likely scores of a given student on all tests." (Mislevy, 1992, p.22). In the Mislevy framework, most of the KIRIS "equating" would be characterized as "linking through calibration" rather than equating.

The Role of Scaling and the Basic Measurement Model

How assessment responses are scaled in the first place is central to the equating process since the initial scaling procedure strongly influences which equating method would be used and how easy it might be to assess its adequacy. Procedures based on Item Response Theory (IRT) are the standard among most measurement experts and many practitioners. There are many IRT approaches, however, and there continues to be some discussion and debate in the psychometric community about the most appropriate IRT method. In addition, there is certainly wide variation in the types of IRT procedures that have been successfully applied in numerous district, state, national, and commercial testing programs. The IRT procedures vary from the simple one-parameter model to the complex three-parameter model; all IRT models have a version for the simple dichotomous case (multiple-choice items scored as correct or incorrect) and the more complex graded-response or partial credit approach (in which students' responses can receive a range of scores from 0 to some higher value like 4 or 5.) The rationale for selecting an IRT model in a particular setting varies. In some cases, psychometricians seem to have become committed to a particular IRT model which they recommend in all cases. A more fundamental rationale has to do with selecting a model that can account for the data in a statistical sense. If data have certain systematic sources of variation, these sources of variation can be modeled, and the resultant model will "fit" the data better; the features or parameters added to the model have technical meaning and important statistical value but their educational meaning and rationale may not be obvious. A different approach, commonly advanced as the rationale for the one-parameter IRT approach, is that the model used to analyze the data should not be determined by the data, but should be predetermined based on some beliefs about the meaning of the parameters used in the model. In this circumstance, the central measurement question becomes whether or to what extent the data actually fit this model.

2. KIRIS Assessment Practices

It is essential to reflect and comment on the purpose and goal of the KIRIS program before examining the various technical details and issues involved in the review of KIRIS equating. KIRIS is an *educational program* designed to encourage and support Kentucky educators who are successful in helping students in their schools and districts attain the educational goals described in Kentucky's Learning Goals, 57 *Academic Expectations*, and *Core Content for Assessment* (KDE, 1996). Ultimately, the value of any technology used in supporting KIRIS needs to be judged in terms of the extent to which it seems to support real growth in students' attainment of these important goals and its fairness to school districts who are held accountable for students growth. As an educational program, KIRIS includes both an educational assessment component and an educational accountability component. It is important to understand that the technical procedures that might be adequate and/or appropriate for the assessment component might not be adequate and/or appropriate to support the accountability component and vice versa.

KIRIS Basic Scaling

All data in the KIRIS system are scaled or calibrated using the two-parameter graded response model for ordered categories (Samejima, 1969). The characterization of a measurement model as having "two-parameter" means that two characteristics of the item are examined and quantified. These are the difficulty of the item and the discrimination of the item (or the degree to which an item differentiates relatively low from relatively high achieving students). The KIRIS open-ended tasks are scored on a scale of 0 through 4; thus, the two-parameter graded response model estimates or quantifies 5 values for each item. There is a single discrimination value for the entire item and there are four difficulties (or 'b' parameters) that represent the difficulty of going from a score of 0 to a score of 1, from 1 to 2, 2 to 3, or 3 to 4.

Equating and Linking in KIRIS

The procedures used for equating and linking in KIRIS are well described in the KDE Equating Plan for Accountability Cycle II (1995), KIRIS Accountability Cycle II Technical Manual (1997), and in the critique by Hambleton et al. (1995), Chapter 8. A wide range of equating and linking procedures have been employed to respond to the complexities of the program and to certain changes that have been made over time in scoring and assessment approaches. A brief summary of the various types of equating and linking will be offered in what follows to provide a context for understanding our review of, and analysis of responses to, Hambleton et al.

Open-ended Math, Reading, Science and Social Studies

Equating across forms within a year. KIRIS uses an assessment design for reading, math, science and social studies open-ended questions in which there are multiple forms of the assessment and each form is composed of common or core questions found on all forms and matrix or sampled items unique to each form. The mix of core and matrix items has changed from time to time. This assessment design allows for breadth of coverage through the matrix or sampled items and stability through the core items. A simultaneous calibration is conducted for all assessment forms with the core items serving as the anchor or origin scale for all students. The measurement of the students is on a common scale since they all took the same common items which form

the origin of the scale. Raw score sample characteristics are described for each form sample as are the parameter estimates for the core items.

Linking across years, within content area. Procedures for equating results in math, reading, science and social studies within the same content area across different years are central to the accountability component of KIRIS. The basic procedures are described in the previously referenced documents. Some changes have been made in certain technical details used to detect outliers in the parameter values over years and in the original use of an iterative procedure that was subsequently dropped. The fundamental logic and basic architecture of the equating process, however, have remained unchanged.

The equating is based on items that were matrix items within a year but common across years. (These items have had least exposure to students since only a fraction of the students sees the items). Two sets of item parameters are estimated for these items, one for each year in which the items appear. The new parameter estimates ("difficulties" but not discriminations) are determined by calculating the slope and intercept that defines the relationship between the two sets of estimates (Equating Plan: Accountability Cycle II, 6B). The "new" estimates for a given year are compared to the values they are supposed to replicate and a series of decision rules are applied to identify "outliers." The "outliers" are items with values that differ from the value they are supposed to replicate. The decision rules about outliers include, or have included, the magnitude of the difference, the number of pairs of b's available, and the correlation between the b's. This process was originally designed to be iterative and would be continued until no more pairs of b's would be removed based on the stated criteria. (KDE dropped the iterative feature of the process in response to a recommendation from the NTWG). Final equating parameter values would thus be obtained and would then be applied to the new form parameter estimates to place them on the scale of the preceding form. The process was applied sequentially working across adjacent years in pairs. The final linear equating of the person performance scale (theta metric) is then determined across the beginning and end of the Assessment Cycle by combining the results of the pairwise year-to-year equating of the item parameters.

Review Comments

The procedures describe above reflect a reasonable application of linear equating methodology extended to the graded-response measurement context. The use of "arbitrary" criteria for determining stability (e.g., .30 or .40) and the minimum number of score points required (e.g., 32) are inherent in the type of scaling and equating that was conducted. Other values might have been chosen but the values used seem reasonable and there are no "magic" numbers that psychometricians recognize as "the correct values." There is, of course, a trade-off between gaining equating points by raising the criterion value used for excluding items that are defined as unstable.

There are several features of this equating process that warrant future consideration. First, it is important to recognize that this approach always produces an equating result. Given some choices about what score points are used and how many will be required, the procedure always generates a result, the adequacy and precision of which cannot be evaluated within the measurement approach. That is, the procedure is not internally falsifiable. A study reported in the KDE Technical Manual for Cycle II examines the errors of equating and suggests that the equating may be working well enough. This study, however, operationalizes "errors of equating" through a particular statistical procedure that may not be as directly related to the question as desirable. Second, separating the score points for a given item from each other and treating them as

if they were distinct binary items may be problematic. There may be an empirical basis for disarticulating the score points but the "items" that are used for equating are certainly not the intact "items" as the student took them. Third, it would be helpful to have some record of the nature of the score points that were lost by applying the stability criteria. It would be useful to know if score points are lost through the whole range of responses or at one or the other end of the response range. Lastly, the role of the discrimination parameters in the equating needs to be explored and explained. The equating appears to be conducted strictly through the "b" parameters and then applied to the discrimination parameters. It is not clear why a measurement approach that insists on the importance of the discrimination parameter for items uses information only from the "difficulty" parameters to conduct equating.

Arts & Humanities and Practical Living/Vocational Studies

The format and structure of assessment in the Arts & Humanities (AH) and Practical Living/Vocational Studies (PL) has changed over time. The AH and PL assessments contain a relatively small number of items and use matrix items only. The linking of the AH and PL tests is accomplished differently from the equating of the open-ended responses in Math, Reading, science, and Social Studies. The groups taking the various matrix sampled AH and PL assessments are treated as equivalent in basic ability relative to these measures. Item difficulties are estimated by calculating the log odds of the cumulative percentage of students at or above each score point. These difficulties are compared across years for repeated items and items that show "instability" in their cross-year difficulty estimates are eliminated from the linking. The cutoff for eliminating items has been raised "... from .40 to .60 because of the limited number of items available for linking the tests from year to year" (Linking KIRIS Assessments Accountability Cycle II: 92-93, 93-94, & 94-95, 1995).

Review Comments

This approach to the linking of the AH and PL assessments differs in technical detail from the equating procedures used for the open-ended assessments in the other content areas but is consistent in its basic philosophy and orientation toward handling data. First, the procedure is based on an important assumption, the central assumption about the equivalence of groups. This assumption must be based on the belief that the tests forms/items were distributed in a random (enough) fashion such that no one set of items is taken by a systematically more or less able group of students. This is a reasonable assumption, especially when the items/forms are mixed with classrooms or schools (or possibly even districts). Second, the choice of a criterion value for determining item stability is made based on the number of linking points the criterion will provide. The original value of .40 was raised to .60 to increase the number of linking points. As mentioned previously, this approach represents a trade-off between gaining equating points and decreasing the stability of the score points used. If the instability in the score points is random over years, more score points is desirable because the random fluctuations will tend to cancel out. If the instability is systematic over years, however, increased score points is undesirable because it will compound systematic errors in equating.

Three technical issues can be raised as topics for future considerations by KDE. First, the description of the "calibration" process, e.g., using the log odds of the cumulative frequency, suggests that the AH and PL data were not calibrated with the basic IRT model used for the open-ended items or with any typical IRT estimation procedure which would use an iterative approach with a convergence criterion. As a consequence, there does not seem to be an opportunity to examine data-model fit.

Second, there is clearly a dependency in the cumulative response categories used to obtain the initial difficulty estimates, but all IRT procedures assume stochastic independence of the items (in this case score points). Lastly, this approach seems to focus on the item difficulty only and ignores the discrimination of the items. Such an approach reflects the use of a one-parameter IRT model (for which the ln(q/p) is the estimate of difficulty when the data fit the model), and represents a departure from the use of the two-parameter IRT model.

Portfolios

Portfolios of students work have been and will continue to be part of the KIRIS program. At present, only writing portfolios are being used in the accountability index while math portfolios are optional. Portfolios have not been equated in the empirical psychometric sense nor are there plans to do so. Writing portfolios are scored using a rubric that leads to the portfolio being classified into one of the four KIRIS categories. The comparability of the use and meaning of the classifications across portfolios and across years is addressed through the training of the scorers, the monitoring of the scoring processes, and the use of checks on the consistency of the scoring.

Review Comment

This basic approach is certainly common in large scale assessment program in the scoring of constructed responses in writing. Its value in any specific setting, of course, rests on the demonstration that raters are scoring check papers to a fixed criterion of agreement and are consistent with each other.

Performance Events

A provision of KERA states that, "The State Board of Elementary and Secondary Education shall create and implement a statewide, primarily performance-based assessment program" (KDE 1995, p.5; KRS 158.6453). Lindle et al. (1997, p.5) raise the question of what actually constitutes "performance-based" assessment and examine the question of whether KIRIS actually constitutes a performance-based assessment program (they satisfy themselves that it does). Performance events have been part of the KIRIS system but are currently not being used. KIRIS performance events involve group activities with a follow-up component in which students make individual responses to activity focused questions (see Lindle, Petrosko, & Pankratz (Eds.), 1997, p.6, for an example). It was clear from the beginning that "equating" performance event across events and years would be a difficult task (ASME response to RFP, Attachment 5B, 1991, p.97). The basic approach that was explored involved using a judgmental process based on small samples of out of state students. The Deputy Commissioner for Learning Results Services, Ed Reidy, writes about, "Puzzling patterns of data for performance events ..." in a letter to OEA's Dr. K. Penny Sanders on January 31, 1997 (page 3, paragraph 6). "The patterns involved performance event results that were large in magnitude and inconsistent with results from open-response in the same content area, and inconsistent with results from performance events in the past". This letter and other correspondence and materials accompanying it, provide careful documentation of the unsuccessful attempts to equate results based on the performance events and the rationale for the decision not to use them in the accountability index. The attempted equating used small samples of students from outside of Kentucky. These students may not have been prepared for the performance tasks in the same way as Kentucky students might be and their expectations and concerns about taking the performance events seriously might not be comparable. The original equating design

called for the use of common subjects taking a pair of performance events. This could not be arranged and students only took one performance event in each content area. The assumption of random equivalent groups was made in an attempt to equate the performance events but the results provided so little support for the linking of the performance events that KDE in consultation with the NTWG decided that the performance events should not be used.

Review Comments

Given the nature of the performance events, a certain degree of fluctuation in the scores should have been (and perhaps was) anticipated. The initial design for equating performance events with common subjects taking pairs of tasks seems reasonable and the attempt to explore equating via the assumption of equivalent groups was also a sensible direction to explore. KDE's response to the inconsistencies and fluctuation in the data from the performance events reflects psychometric caution and technical good sense but may have created some policy issues and potential legal difficulties. Districts certainly could have devoted time to preparing students for performance events, time which may seem to have been wasted since scores on these events were not used in the accountability index.

At this time there appears to be continued interest in working performance events back into the KIRIS accountability systems. Three issues can be raised as topics for consideration by KDE in regard to the performance events. First, the construct or constructs being assessed by the performance events need to be carefully defined and demonstrated. This includes clarifying the task demands both in the group work and in the individual students' work. In addition, the extent to which skills in collaborative learning and group work influence the individual students' scores should be examined. Second, the role and influence of the of the characteristics of the group within which a child works needs to be identified and partialled-out from the student's individual performance. Lastly, some form of empirical evidence linking the scoring of these events across tasks within years and within tasks across years needs to be developed and evaluated.

A very unusual relationship between educational policy and technical procedures is illustrated in the KIRIS experience with performance events. The policy decision to include performance events was made for educational reasons in the absence of any proven technical procedures that could be used to equate or link scores across performance events. Performance events were administered and plans to use them in the accountability index were made. Only after the fact was the absence of adequate technical support recognized, confronted, and resolved by the decision not to use the performance events. The initial decision to use performance events seemed to assume that some technical procedure existed or could be devised to support the use of these assessments. There is a very different and more common relationship between educational policy and technical procedures. In many assessment programs, technical limitations are recognized and are used to inform policy decisions before they are made. In this approach, technical procedures that might support the innovative policy or program are developed and field tested. The results of field testing the procedures are used to inform policy makers about the wisdom and timeliness of implementing the policy or innovation.

Writing Prompts

Students in various grades take a "Writing Test" in which they are presented with two prompts, both calling for the same type of writing, e.g., narration. Students select one of the two prompts and construct a response based on the prompt. These prompts and the scoring of students writing are not equated in the empirical psychometric sense. Writing samples are scored using a rubric that leads to their being classified into one of the four KIRIS categories. In fact, the same rubric used for the Writing Portfolios is used for these writing samples. The comparability of the use and meaning of the classifications across portfolios and across years is addressed through the training of the scorers, the monitoring of the scoring processes, and the use of checks on the consistency of the scoring.

Review Comments

This basic approach is certainly common in large scale assessment program in the scoring of constructed responses in writing. One general issue can be raised as a topic for future consideration by KDE in regard to the writing prompts and the scoring of students responses to them. Specifically, as time and other resources allow, the investigation of empirical methods for equating should be explored. IRT methods are finding increased application to the analysis and scaling of students' writing and testing the consistency of raters' scoring and this may be a useful area of inquiry for KIRIS.

Equating Proficiency Levels Across Cycle I and II.

As described in Section 6 of this report, a decision was made to validate Accountability Cycle I standards rather than reset them for Cycle II. Two options were considered and a decision was made to "Determine the relationship between the original and revised 1992-93 scales using a common person method, and adjust the proficiency level cut-points accordingly" (KDE, Tech Report Cycle II, p. 9-10). The procedure for doing this equating is quite straightforward and well described. The procedure is based on certain reasonable assumptions and the resultant adjustments appear to be relatively modest.

KIRIS 1997 GRADE Shift Adjustments

Beginning in 1997, KIRIS on-demand assessments in Mathematics, Social Studies, Arts and Humanities, and Practical Living and Vocational Studies were shifted from 4th grade students to 5th grade students and Reading and Science on-demand assessments were shifted from 8th grade to 7th grade students. This change was designed to lessen the testing burden on individual students and to increase the number of students involved in KIRIS assessment. A special study which eventually explored numerous equating procedures was conducted to determine the appropriate performance standards for students taking the respective assessments at the new grade levels (Wise, 1997) .

The original design for adjusting performance standards across grade levels actually contained two cross-year counter-balanced equating procedures that could be used to cross-validate the results. This was a very useful and powerful design that the KIRIS program should incorporate more often. The results based on two approaches in this initial design, however, "...did not adequately cross-validate on the 1997 data" (p.2). Several hypotheses were offered as to why there were changes in grade differences.

A second equating approach was explored and recommended. This approach combined data across the two years and employed an equipercentile equating procedure. In this approach, category cut-points were adjusted so that the proportion of students in the new assessment grade were as similar as possible to the proportion of students in the corresponding category for the previously used assessment grade. This approach assumes that sources of variation that prevented the cross-validation of results in the original designed were counter-balanced across years and canceled each other out when the data were combined across years.

The procedures used to adjust standards across grades for Arts & Humanities and Practical Living/ Vocational Studies were different from the procedures used in the four core subject areas. Four different approaches were explored. A hybrid approach was selected. Equipercentile values were used for the Apprentice and Proficient cut-points. An additive constant approach was used for the Distinguished cut-point.

Review Comments

The procedures used to adjust performance standards across grades represent a very unusual amalgamation of approaches. The original equating design for the four core areas was a well planned and powerful design of the type that Hambleton et al. seemed to call for because it would allow for the cross-validation of the procedures. The results were clear: different approaches to equating produced different results. The response to such a finding might be the conclusion that results across grades should not be equated since equating results were inconsistent. The KIRIS response, however, was to explore yet another equating approach, namely equipercentile equating. The results of this approach generate plausible results and the approach was employed. A similar logic was used in selecting the procedures used for adjusting performance standards for Arts & Humanities and Practical Living/ Vocational Studies. Wise (1997, p.4) reports that, for these assessments, "The Equipercentile method yielded the most plausible (e.g., least extreme) adjustments for the Apprentice and Proficient cut-points." Selecting equating procedures based on the criterion of "plausibility" of the results that they produce is a very uncommon psychometric criterion. The application of the equipercentile method to the Distinguished cut-point was not possible because of the small sample of students in the Distinguished category and a different equating method was added for this cut-point. The combining of approaches, as seen in the procedures for adjusting standards for Arts & Humanities and Practical Living/ Vocational Studies, is very reminiscent of the "ad hoc" quality of the technical procedures that the 1995 OEA study suggested be avoided.

The KIRIS relationship between educational policy and psychometric technology is again illustrated in the search for a technical approach to support the policy decision to adjust standards to accommodate shifting grade levels for the various assessments. This policy decision seems to have been made with the assumption that some technical procedure exited or could be devised to support the policy. An equating study with a cross validation component was devised and conducted after the grade level shift had taken place. The study failed to cross-validate the equating. The results of the study could have been used to inform policy makers that adjusting standards across grades was not a very sound idea and that it might be prudent to revisit the standards at the new assessments grades. The results of the initial equating analysis were not used to inform policy in this fashion. and other equating approaches were explored. The research involved going through a repertoire of equating approaches until an approach (or combination of approaches in the case of Arts & Humanities and Practical Living/ Vocational Studies) produced a "plausible" result.

3. Are the scores comparable across Administrations? Hambleton, et al., 1995, Chapter 5.

The KIRIS equating procedures used up until 1995 were carefully reviewed in Hambleton, et al., (1995), Chapter 5. Hambleton's Chapter 5 provides a detailed description of the KIRIS technical procedures used up to that time and is substantially critical of the KIRIS equating procedures. Our review of Hambleton et al. will briefly examine and comment on 1) their two general conclusions, 2) their seven specific conclusions, and 3) their recommendations and KDE's response to these recommendations.

Hambleton et al. General Conclusions

The authors offer two general conclusions about the KIRIS equating procedures:

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

These two generally conclusions refer to a variety of ad hoc decisions (documented in Hambleton et al., Chapter 5) that were made in the scaling and equating procedures during the first cycles of the KIRIS program. In the early stages of the KIRIS assessments, there were a number of outcomes in the operational aspects of KIRIS and in the results obtained by applying the planned technical procedures that had not been anticipated and for which contingent technical plans had not been considered or developed. KDE staff and contractors made a number of decisions that the exigencies of the situation required. Most of these decisions appear to have been reasonable, but the circumstances that created the need to make such decision could have been anticipated.

The rapidity with which KERA/KIRIS went from legislative action to field based operation also bears on the issue of the adequacy of the early KIRIS technical procedures. KIRIS was up and running very quickly at a time when the psychometric procedures needed to support the various assessment components were not well developed. The implementation schedule for KIRIS did not seem to allow for a "research phase" during which various psychometric procedures could have been developed, field tested and evaluated.

Hambleton et al., Seven Specific Conclusions

Seven much more specific concerns are identified which Hambleton et al. believe "... need to be addressed if KIRIS remains a high-stakes accountability system." These are:

1. inconsistency of procedures across years;
2. use of inefficient equating design;
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 equating of performance events and use of non-equated portfolio scores;
7. combining data from noncomparable performance events, portfolio assessments, and alternative assessments for special education students with transitional assessment data. (Hambleton et al., p. 5.2.)

These seven concerns, taken as a group, reflect the two general conclusions previously mentioned. The status or timeliness of these concerns has changed over time. A brief discussion here will consider Conclusions 1 and 4 together. Conclusion 3 will be discussed in some detail because it helps frame a central feature in the evaluation of the technical aspects of KIRIS. Several other issues will be taken up when the recommendation of Hambleton et al. are considered in the next subsection.

Over the life of the KIRIS program, there has evolved a certain consistency in the application of technical procedures and this includes consistency in decision rules which were, originally, arbitrary and ad hoc. This is not to take a position on whether the procedures and decision rules are "correct", appropriate, or valid. It is to acknowledge, however, the concerted effort of all parties involved in managing the technical aspect of KIRIS to standardize the procedures and decision rules involved in the KIRIS scaling and equating processes so that the outcomes of applying the procedures would start to take on a consistent meaning. In this regard, it is useful to note that many decisions made about specific numeric criteria in the application of statistical and psychometric procedures are arbitrary to begin with, but take on meaning and value over time because of the consistency in their use.

The status and function of multiple-choice items in the KIRIS system continues to be reviewed and revised and must be considered relative to the goals and purposes of KIRIS. As mentioned earlier, KIRIS is an educational program that includes both an educational assessment component and an educational accountability component. The educational assessment component of KIRIS is designed to encourage and support Kentucky educators who help their students learn and master the educational goals described in Kentucky Core Content for Assessment (Version 1), 1996. Many educators involved in the initiation, development and on-going support of KIRIS feel very strongly the instructional goals and learning outcomes of KIRIS cannot be appropriately assessed with multiple-choice items. Indeed, their position would be that the use of multiple-choice items would corrupt the curricular intent of KIRIS by focusing attention on aspects of the curriculum domain that are decontextualized, discrete skills that are limited in scope and complexity. This perspective would not consider using multiple-choice items simply because they might improve the technical characteristics of KIRIS. Such improvement in technical rigor (if any) would be thought of as being purchased at the price of distorting the curricular intent of the program.

There are other educators who take a different point of view, a point of view not reflected in any KIRIS materials, documents or discussions. This position generally involves two issues. First, proponents of this "let's-use-multiple-choice-items" perspective would assert that carefully constructed varieties of multiple-choice items can assess a broad range of curricular goals and learning outcomes and are not limited to measuring discrete skills. Second, the psychometric technology for scaling and equating multiple-choice items is so well established that their inclusion in an assessment program may be justified on purely technical grounds. From this point of view, multiple-choice items could be included in KIRIS to provide technical defensible evidence that the year-to-year equating is working effectively. In the face of such evidence, doubts about the meaningfulness to year-to-year gains could be minimized and accountability decisions recognizing gains in learning could be made with more confidence.

The status and function of multiple-choice items as "corrupters of curriculum" or "saviors of technical rigor" cannot not be established as matters of fact; these two views actually represent differences in curricular philosophy and educational-political beliefs. The KIRIS program, to date, has eschewed the use of multiple-choice items for reasons of curricular validity despite the possibility that such items might add technical rigor and possibly increased technical credibility (and some would argue that multiple-choice items might also increase curricular validity). This decision about multiple-choice should not been evaluated as "right" or "wrong", but should be viewed on the continuum of being more or less useful in supporting the goals of the KIRIS program.

Hambleton et al. (OEA) Recommendations and KIRIS Response

The report by Hambleton et al. offers a series of recommendations about the technical procedures used for equating. KDE responded to these recommendations in "Summary of the KDE Response to KIER and OEA Report Recommendations to Improve the Kentucky Assessment and Accountability Program" (July 6, 1995). The KDE responses contained in this document are presented below. Hambleton et al. are referred to as OEA (Office of Educational Accountability), which sponsored the Hambleton et al. study.

The OEA recommendation number is listed followed by the specific recommendation being cited. The second line offers the page number from Hambleton et al. where the recommendation is made and the KDE description of its response to the recommendation.

OEA Recommendation #. Recommendation. /
OEA page Reference: KDE Response

#16. Use two-parameter IRT model and allow multiple mappings of the same raw score.
OEA, p 5-2; KDE Agrees.

#17. Use multiple-choice items to form stable form-to-form and year-to-year links.
OEA, p.5-2; KDE Disagrees

#18. Use larger sample sizes to improve the equating of performance events.
OEA p.5-3; KDE Agrees

#19. Check the stability of all items remaining in an equating link. Eliminate those items with large differences.
OEA p. 5-3; KDE Agrees

#20. Do not try to equate Grade 12 performance to Grade 11.
OEA p.5-3; KDE Agrees

#21. Establish a new baseline the first year the transitional tests are administered in Grade 11.
OEA, p. 5-3; KDE Disagrees

#22. Use an adequate number of linking items to allow for expected losses.
OEA, p. 5-3; KDE Agrees

#23. Design the placement of common items to ensure acceptable comparability of performance across forms and years.
OEA, p 5-3; KDE Agrees

#24. Calibrate the common and matrix items simultaneously using an IRT model.
OEA, p. 5-16; KDE Agrees

#25. 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.
OEA, p 5-17; KDE Partially Agrees

#26. The design for equating assessments should be strengthened and the ad hoc procedures eliminated.
OEA, p. 9-7; KDE Partially Agrees

#27. Set standards for Novice, Apprentice, Proficient, and Distinguished on an intact form and equate all other forms to it.
OEA, p. 5-2; KDE Disagrees

#29. Base a student's classification in each subject on the total set of multiple-choice and open-ended items.
OEA, p. 5-3; KDE Agrees

The KDE summary of recommendations made by Hambleton et al. is accurate and comprehensive. Furthermore, KDE decisions in regard to the recommendations seem quite responsive. The three recommendations with which KDE disagrees include using multiple-choice items as a mainstay of the between-forms and across-years linking; setting standards on intact forms and equating to these base forms; and, establishing a new baseline at Grade 11 when the transitional tests are first used at that grade. KDE's position on the use of multiple-choice items has previously been described and seems to reflect a curriculum concern more than a psychometric decision. The use of intact forms as the origins of the equating would actually require a complete departure from the KDE equating/linking approach. The rationale for not starting a new baseline year at Grade 11 is not clear but seems to reflect a concern about having to reset the entire system.

4. Other Related Analyses

There are two additional areas of inquiry that can be discussed under the general heading of equating (or linking) since they both concern the relationship of KIRIS assessments to other forms of assessment. Strong and Sexton (1997) provide an empirical investigation of the relationship between students' classification into the four KIRIS Performance levels and their classification on ACT Math when classified into four score groups (1-15, 16-17, 18-21, and 22-36). The authors provide a 4x4 contingency table (their Table 1) which they describe as showing that KIRIS does not adequately discriminate at the Novice and Apprentice levels and is only slightly better in differentiating students at the upper performance levels. This study is certainly well intended and studies of concurrent validity (such as this) and other studies designed to broaden the understanding of what it means to be in the different levels of the KIRIS classification system are certainly important. Strong and Sexton's work, however, has certain limitations. Their criteria for evaluating the overlap in data in the 4x4 table are never stated, the criteria are certainly not empirical, and no inferential tests are conducted. Of greater concern is the fact that no rationale is offered for discussing the degree to which classifications based on the two measures should agree and if the agreement should be the same at all levels. An estimate of the overlap between the ACT and the Kentucky Academic Expectations needs to be established before the "overlap" between ACT and KIRIS and can be interpreted. In general, the appropriateness of the criterion measure needs to be explored and established before the results of a criterion related validity study can be interpreted.

Nitko (1997) and Nitko et al. (1997) also offer investigations of KIRIS relative to a number of tests. Nitko (1997) provides very valuable and user-friendly discussion of the similarities and differences among KIRIS, the California Achievement Tests, 5th Edition (CAT5), the Comprehensive Tests of Basic Skills, 4th Edition (CTBS4), and TerraNova (CTBS5). These tests are discussed relative to thirteen questions, which deal with policies, procedures, and interpretation issues. Nitko notes a few gaps in KIRIS coverage of Kentucky's Goal 5- Think and Solve Problems and Goal 6- Connect and Integrate Knowledge (p. 39) , and expresses concern about the measurement properties of the performance events (p. 40).

Nitko et al. (1997) address the question, "How well are the Kentucky academic expectations matched to the KIRIS 96 assessments, CTBS4, and CAT5. The study found that "... KIRIS open-response questions, when viewed as a complete set, do assess most of the Academic Expectations" but the authors note a few exceptions. CTBS and CAT did not cover complex thinking processes as well as KIRIS. The authors found that CTBS and CAT did cover traditional academic basic skills in spelling, math, and language better than KIRIS. It does not appear that multiple-choice items were included in this review, the use of which might increase the extent to which KIRIS assessment will cover traditional basic skill areas.

Review Comments

The issue of concurrent validity and content/construct validity are raised in the three studies described in this section. These are certainly important issues and continued investigations should be encouraged. These issues of validity need to be framed, however, in the context of the KIRIS program. KIRIS is basically a curriculum referenced program, referenced to a set of Academic Goals and Academic Expectation, as described in the Core Content for Assessment (Version 1), 1996. As a consequence, studies like those of Nitko and Nitko et al. are central to evaluating KIRIS. Nevertheless, it is important to show that doing well on KIRIS assessments has some implications or consequences for students' academic progression. Support for KIRIS may be difficult to maintain if doing well on KIRIS assessments only means that students can do well on KIRIS assessments and has little or no relationship to other measures of students' academic attainment.

4. Our General Critique

1. KIRIS is an evolving assessment system that was set in place very rapidly.

2. During Cycle I, in particular, and also Cycle II, a variety of technical decisions were made to get the KIRIS up and running. Hambleton et al. characterized a number of the decisions as "ad hoc" and arbitrary, a characterization that was reasonable at the time.

3. The technical problems faced by KIRIS were (and continue to be) understandable, given the choice of psychometric model and equating design.

4. The "ad hoc" decisions made to support KIRIS psychometric procedures were reasonable (albeit arbitrary) in the context of the measurement model and equating design employed, and many such decisions are made when running any operational program.

5. The once "ad hoc" decision rules used in equating open-response items have been used consistently over the life of the KIRIS program. The "ad hoc-ness" of a procedures or decision rule certainly diminishes with repeated use and these procedures now seem to add to the argument that consistent procedures are being employed.

6. KIRIS seems to have been responsive to the recommendations of Hambleton et al. in several ways and two, in particular, stand out: a) the inclusion of multiple-choice items; b) the omission of performance events in the accountability index. However, the procedures used recently to adjust the standards for grade level shifts appears to be a return to the "ad hoc" approach to which Hambleton et al. strongly objected.

7. The decision to scale and equate multiple-choice items separately from open-response items is reasonable, although there is certainly another reasonable point of view that argues that the two formats should be to scaled together. KIRIS policy makers could choose to investigate this issue empirical and use information from a research investigation to inform policy.
8. The scaling and equating procedures used in the analysis of the open-response tasks in KIRIS do not appear to contain any major flaws that would jeopardize the program.

9. The technical procedures used to support KIRIS are fundamentally the same as those decided upon at the very beginning of the program (circa 1990?). There has been some tinkering and refinement of these procedures, but there has been no exploration of any approaches that examine the technical issues from a fundamentally different perspective.

10. The major specific technical aspects of KIRIS that could and should be improved is increased rigor in checking the precision and stability of the equating procedures. A variety of procedures that are more powerful in checking the adequacy of linking and equating are available and should be explored. Included in the consideration of this issue should be the possibility that, in at least some cases, parallel analyses would be conducted using different approaches. The results of the study using the "jack knife" analysis to assess the accuracy of the equating is informative and reassuring, but does not speak as directly to the verification of the equating as other approaches can.

11 The equating over the beginning and end of the assessment cycles is central to the use of KIRIS as an accountability program and efforts to demonstrate the adequacy of this equating have been narrowly focused on the psychometric audience. It may be necessary, but it is certainly not sufficient, to convince other psychometricians that the equating procedures are appropriate and adequate. With the resources available and the experience that has been accumulated, KIRIS staff, advisors and contractors should focus very serious attention on routinely building into the scaling and equating procedures straightforward mechanisms for demonstrating that apparent gains or losses in student attainment reflect characteristics of the students and are not artifacts redefining score points, recasting data, scaling or equating. In the long run, this type of "validation" of the scaling and equating procedures aimed at educators and policy makers may have far more serious consequences for KIRIS and the students it is designed to serve than technical validation efforts aimed at the psychometric community.

12. The KIRIS relationship between educational policy decisions and technical support procedures needs to be changed. To date, policy decisions seem to have been made with the assumption that some psychometric procedure exists or can be devised to support any policy decision. Those responsible for the technical aspects of KIRIS have been very conscientious and resourceful in devising technical procedures that produce results that are reasonable or "plausible" enough to support the decisions. This post hoc use of technology to support rather than inform policy decisions may be appropriate early in the development of an innovative educational program to insure that the demands for technical rigor do not stifle innovation. KIRIS, however, has had sufficient time to develop and at this point in the program, technical procedures should be used in two ways: 1) to inform policy decisions before they are made; 2) to evaluate the adequacy and precision of technical procedures that are already in place.

13. The apparent resistance of KIRIS policy makers to recognize the need to re-establish baseline years for scaling, equating and standard setting is difficult to understand. Many statewide assessment programs and other assessment programs have reset their baseline year after a period of implementation and the task, while difficult, does not seem overwhelming. Given the long chain of adjustments and equating links that holds the KIRIS program together, it would seem timely and prudent to begin the process of re-establishing a baseline year and re-setting standards as soon as possible.

Section 6. Standards Setting and Validation

OEA Panel Recommendation

Recommendation OEA 6 states that, "Performance standards should be re-established and full documentation of the process should be provided." This recommendation was based on the analysis contained in Chapter 6 of the Hambleton et al. (1995) report (frequently referred to as the OEA Panel Report). Without reviewing the chapter in detail, we note that the authors raise a variety of important issues which we summarize below. (Several sentences below are essentially quotes from Chapter 6 of Hambleton et al.).

KIRIS identifies students whose school achievement in a number of subjects warrants their classification into four categories: Novice, Apprentice, Proficient, and Distinguished. The proportion of students placed in each of the four categories for each of the subject areas forms the basis of the school's Accountability Index. The classification of schools depends critically on the validity of the procedures used to establish performance standards. 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. While there is no single best method of standard setting, the AERA, APA, NCME Standards (1985) requires that the method and rationale be presented in a manual and the qualifications of the judges should be documented.

Hambleton, et al. noted several problems (or at least limitations) of the standard setting process used for the 1991 baseline data. One is that the judgmental basis used to establish the performance standards was used for only the common open-ended questions. This leads to a tenuous result because of the small samples of items used (and, according to Hambleton et al., because of the undocumented qualifications of the judges). Kentucky then used analytic procedures to produce standards for the matrix-sampled open-ended items which the Hambleton et al. report describe as "although innovative and novel, are without judgmental or empirical foundation, depend on untested assumptions, and are sensitive to biasing statistical and measurement artifacts." (p. 6-6). A further problem raised is that the initial procedure used in 1991-92 used a model that was, in part, conjunctive in nature, however the performance standards at the end of the 1992-94 biennium used a compensatory item response modeling procedure. (In a conjunctive model, high scores on some items can not compensate for low scores on others. For example, to score Proficient in mathematics at grade 4, a student needed a total of 9 - 11 points across three items, but none of the items could have a score of less than 3. Thus, a student who scored 4, 4, and 2, for example would have a total score of 10, but not be considered proficient because for the one item the score was less than 3.) As the report suggested, "it is likely that some school classifications are a consequence of this change in models." (p. 6-7). This was almost bound to occur since the conjunctive model was, by its nature, more difficult to achieve. Thus, apparent growth in the area of mathematics could be due to a change from the conjunctive to compensatory model rather than due to a change in acquired knowledge or skill.

Thus, the Hambleton et al. report criticized several aspects of the standard setting process -- and the fact that a different model was used in the 92-94 results than was used for the 91-92 results. In addition, the report took issue with the Kentucky Department of Education's stance that standard-setting was a less critical issue for KIRIS than for most high-stakes tests. KDE had three rationales for their position: (1) the purpose is to drive instruction, and there is no reason to believe that such driving will be impacted by where the standards are placed; (2) Schools are held accountable for improvement, not initial score; and (3) because most schools start with low distributions, different placement of the standards would not have affected the Accountability Index or school placement. Hambleton et al. point out that the first argument is speculative and the second and third have been refuted by analyses conducted by Richard Hill of Advanced Systems in Measurement and Evaluation (ASME).

Further issues are raised in the Hambleton et al. report. We will only mention three here. First, the documentation of the procedures used in the initial standard setting were "sketchy at best" (p. 6-15). No information was provided about the training of the panelists, the specific questions posed to them, or the deliberations and discussions of the panelists. Adequate training was important because of the second point -- the performance level definitions adopted by the KDE were far from operational. More will be said about those definitions later in the subsection entitled Performance-level definitions.

The third point is related to how performance standards were set for performance events. Actually, there were no judgmental procedures used to set standards for the performance events. Rather, it was assumed that the percent of students at each classification level would be the same as the percent classified on the common open-ended items. Thus, "an empirical procedure that divided the distributions of scores on Performance Events so that they most closely approximated the desired distributions was applied." (Hambleton, et al., 1995, p. 6-25). As the OEA report concluded:

> "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." (Hambleton, et al., 1995, p. 6-26).

Thus, the Hambleton et al. report provided a strong basis for their recommendation that the standards be re-established and that full documentation of the process be provided.

KDE Response

In a Summary of the KDE Response to KIER and OEA Report Recommendations to Improve the Kentucky Assessment and Accountability Programs (July 6, 1995), the department agreed with the recommendation to re-establish performance standards. They developed a draft process plan which was sent to the NTAC (National Technical Advisory Council) which we believe was the same as the National Technical Working Group (see below). The proposal called for standards to be set for one subject at one grade with the results and documentation to be reviewed by the NTAC before proceeding in other subjects and graded.

Reaction of the National Technical Working Group (NTWG)

The KDE has a National Technical Working Group (NTWG) that has been advising them on technical matters. Two of the members of this group (Haertel and Wiley) wrote a response to the OEA report. In that response, they suggested that the OEA report was "troubling in its tone and facile in its recommendations." (Haertel and Wiley, no date, p. 1). For example, they suggested that to criticize something as being innovative and novel but without judgmental or empirical foundation implies that standard setting methods not plagued by untested assumptions exist and that "this is patently false." (p. 2). Haertel and Wiley argued that while the three cut points were set rather arbitrarily, "substantial meaning has accrued to those cut scores" (p. 4), and that the OEA panel's report suggesting that the "categories cannot be useful because some appendix failed to contain sufficiently elaborate biographies of the educators involved in their invention or that the silly rituals of the Angoff Procedure would have provided a firmer basis for their definitions simply do not make sense." (p. 3). They concluded by reiterating that "the meanings of standards inhere in their use -- Not in the rituals through which they are established." (p. 5).

It seems useful to note that none of the Haertel and Wiley rhetoric was a defense of the standard setting methods. Their response did not even address some of the issues -- such as assuming the same distribution on the performance events classifications as on the three item open-ended standards. Rather, they suggested (in a response that was troubling in its tone and facile in its recommendations) that the standards, although arbitrary, have acquired meaning.

Cunningham, in his Response to the Response to the OEA Panel Report (no date) questions the motivation of the authors of the Response. As he stated: "If they are paid consultants hired for the specific purpose of refuting the OEA Report, their criticism needs to be interpreted in that context." (p. 1). We do not know what the financial reimbursements were to Haertel and Wiley, if any. We accept their report as reflecting the authors' beliefs.

The September 1-2 (no year stated, but apparently 1995) minutes of the NTWG state that:

> "The committee disagreed with the conclusions set forth by the OEA panel, and was opposed to re-setting the standards....To reset standards to a new collection of descriptions and score scale cut-points conflicting with those existing in practice runs counter to the purpose of having standards at all. ...although the committee acknowledged that it would have been desirable to have used a greater number of judges and items and to have altogether avoided the use of a conjunctive model." (p. 8).

Decision made to Validate the KIRIS Standards

Taking the advice of the NTWG the KDE decided to "validate" rather than re-set the standards. This validation (in mathematics, reading, science & social studies) was undertaken by Advanced Systems in Measurement and Evaluation and they issued a final report on May 15, 1996 -- to be discussed in the next section. The report suggested that the standards validation study was conducted both because the KDE had originally stated that the standards would be reviewed and because of the OEA criticisms.

In a short paragraph summarizing the OEA criticisms the ASME report agreed they were well founded, but indeed, only alluded to the incomplete documentation, the limited number of items, and the small number of people involved. They did not mention the conjunctive/compensatory model change or some of the other concerns raised by the OEA panel.

Was this decision responsive to the OEA recommendation?

Whether the study to validate the standards is sufficiently responsive to the OEA recommendation is a matter that is not easily answered. Both the authors of the OEA report and the members of the NTWG are nationally recognized measurement experts -- and they disagree on what should have been done.

We believe the original OEA concerns were well founded. Some of their concerns (e.g. changing the model from the 1991 to the 1992-93 standards, basing the original standards on only a very few items, employing innovative but not well researched analytic procedures, not documenting the process as well as should have been done) are especially relevant. The fact that the standard setters' qualifications were not as well documented seems less important because there is no definitive literature on what standard setters' qualifications should be or whether it makes much of a difference. (However, not documenting the qualifications is contrary to recommended practice.)

We also agree with the authors of the OEA report that where the initial standards were set could influence a school's eventual classification even though the classification is based on a growth model.

However, the suggestion by the NTWG to validate the standards rather than to reset them is not totally without logic. It seems legitimate to argue that since 1991 "substantial meaning has accrued to those cut points and the categories they define."

Thus, we take no absolute position regarding whether the validation study was an appropriate response to the recommendation – although it obviously did not follow the recommendation. It can be considered a professionally acceptable response – there is documented evidence that a set of legitimate technical experts thought it to be the correct course of action. We now turn to a discussion of the validation process itself.

KIRIS Standards Validation Study (Mathematics, Reading, Science, & Social Studies) (May, 1996).

The KDE invited 352 judges to participate in the validation studies and 246 actually participated. Judges were divided into two groups: a confirmation group and a replication group. The confirmation group was shown descriptions of each standard and told how those standards apply to student work. They were then asked if they felt the described standard applied to that work, or if a different standard was more appropriate. The replication group did the classification without being told how the level of work was applied to the current standards. In both cases, the judges started with a set of definitions of what was meant by novice, etc. Thus, in one sense, the judges did not validate the standards – they were given the standards. What they did was to place student papers into one of the four classifications that has been given them. As the final report states:

> "The first step in the overall standards validation process was to ensure that the general and content-specific definitions were still appropriate. A group ... reviewed the existing general definitions for the four existing performance levels ... and made modifications where appropriate. These definitions were then used as the basis for content-specific definitions. The content-specific definitions were then determined by advisory groups ...These groups met to review the general definitions for the four desired performance levels ... and then create subject-specific performance level definitions. These definitions formed the basis for the judgments made by the standards validation panels." (p. 7).

Unfortunately (in our opinion) the process for reviewing the general definitions and creating subject-specific performance level definitions is not described further in the report – although the definitions are included in the Appendix of the report. At any rate, it is important to note that:

> "the role of the judges was to match individual sets of student responses to the general and content-specific performance level definitions. It was not the role of each judge to establish their [sic] own standards criteria. Their role was only to consistently apply the performance level definitions to student work being reviewed." (p. 7).

Thus, as stated above, the standards were not validated; rather student data were categorized into existing standards.

The results of the "validation" study indicated that judges "generally agreed with the standards as they were originally set" (p. 14). While we agree with what the data show, we would interpret this as saying that the judges, when using essentially the same standards used to make the initial classifications, placed student papers into the same categories as their scores (based on those initial standards) placed them. Recall that the judges did not establish their own standards criteria.

The validation study has been reasonably well described and the results reasonably well documented in the May 1996 final report. However documentation is lacking on the process of reviewing general definitions and creating subject-specific performance level definitions -- a topic we address further in the next sub-section.

Performance-level definitions

The OEA panel report criticized the original definitions of proficient, apprentice and novice indicating they were "quite brief and vague." (p. 6-10). A set of documents entitled World-Class Standards...for World-Class Kids: Performance Level Guides for Open-Response Common Items contains the performance level descriptions for grades 4, 8, and 12. Although the documents are undated, we were informed by KDE (which furnished these documents in response to our specific request) that these contain the original performance level definitions. These definitions have fairly common wordings across grades, but there are minor differences across some grade levels for some subject matter areas. We provide two examples below for the Novice level for grades 4 and 8 for mathematics:

> **Original Grade 4 Math Novice:** "A student at this level shows some basic understanding of problems and attempts to use strategies. The mathematical reasoning is not always appropriate and frequently leads to incorrect conclusions. There is a limited understanding of core concepts which leads to responses that are vague and lack detail." (p. 13 of 4th grade document).
>
> **Original Grade 8 Math Novice:** "A student at this level may show some basic understanding of problems and attempts to use strategies. The mathematical reasoning used is not always appropriate, and would not lead to a correct conclusion. There is a limited understanding of core concepts which leads to responses that are vague and lack detail." (p. 13 of 8th grade document).

1996 "Validated" Math Novices. The math definitions used in the validation study are apparently constant across grades and do not precisely reflect the wording or grade level differences found in the original standards. The definitions are listed in bullet form and, for math novice are as follows:

> Student shows limited understanding of problems.
> Student uses no or inappropriate reasoning.
> Student rarely or ineffectively uses mathematical terminology and/or representations (graphs, charts, etc.).
> Student demonstrates limited understanding of core concepts.
> (Validation study, 1996, p. 24).

We are not suggesting that these changes between the original articulation of student performance standards definitions and the definitional forms that showed up in 1996 make a difference. We do not know. Does it make a difference to say, as was done in the original document, that a student at the novice level "shows some basic understanding" and to say in the validation study definitions that a student at the novice level "shows limited understanding?"

The problem is that there is inadequate documentation of the process that was used in the validation study to make these changes. We do not know, for example, why the decision was made to have a constant set of words fit all grade levels (which is not quite what was done in the original standards), or why there was a switch to a bullet rather than keep a paragraph format in the definitions. However, the changes do seem minor and, as the validation report points out, the judges did not establish their own standards criteria -- they simply classified student responses into existing defined categories.

Of course the classification into "novice," for example, would depend a lot on what problems were given to the student. For example, a very "distinguished" fourth grader might have only "limited understanding of problems" if those problems were not appropriate for the grade level. Thus, the performance standards are really dependent on the test question quality and appropriateness.

Standard Setting for the Writing Portfolios

The 1992-93 Technical Report contains one paragraph on setting standards for the writing portfolios. It states that "the Writing Advisory Committee revised the KIRIS Annotated Holistic Scoring Guide for writing portfolios so that it would parallel the performance levels of scoring guides in other content areas. ...Though the scoring model changed, it is important to note that Kentucky standards have remained constant..." (p. 8-1).

This report is too sketchy for us to evaluate whether this change in the scoring model resulted in unchanged standards or not.

Standard Setting for Arts and Humanities, Practical Living, and Vocational Studies

The 1992-93 Technical Report devotes about four pages to this standard setting. As the report points out, "since each student completed only one item in these subject areas, performance on that one item was the basis for assignment to a performance level." (p. 8-2). The report states that "clear standards were difficult to establish at the Proficient and Distinguished levels, with some grades/subjects at those levels showing considerably mixed results. ... These difficulties may indicate a need to re-examine the standards in subsequent years of the program." (p. 8-5).

A report entitled KIRIS Standards setting study: Arts & humanities, practical living/ vocational studies dated October 18, 1996 describes a study conducted concurrently with the KIRIS Validations Study in mathematics, reading, science and social studies described above. (Note that the report is

dated five months later than the validation study report.) As the introduction points out,

> "While the purpose of the validation study was to determine if groups of Kentuckians agreed with existing content standards, the results of this standard setting study will be used to reset standards for 1995-96 arts and humanities and practical living/vocational studies" (p. 3).

This resetting of standards (from the original 1993 standards) was apparently necessary due to changes made in the assessment. This may be true, but we conclude that the point seems, at least on the surface, inconsistent with why the KDE did NOT reset, but rather "validated" the standards in the other areas. Recall that the argument for validating rather than resetting was that the standards had acquired meaning and therefore should not be changed. The distinction, according to the report, is that there had been no "promulgation" of the arts & humanities and practical living/vocational studies standards. This is a relevant distinction if, indeed, the educators in the state were not aware of the 1993 standards.

The report of this standard setting is deficient in describing how the content definitions were established (just as was the Standards Validation Study). As with the validation study, it "was not the role of each judge to establish their [sic] own standards criteria. Their role was only to consistently apply the performance level definitions to student work being reviewed." (p. 7).

Conclusions regarding Recommendation 6

We agree with the substance of the criticisms of the original standard setting raised by the OEA panel. There were inadequacies in the original standard setting process and in the documentation of the process and the qualifications of the participants. It was these inadequacies that led to their recommendation to reset the standards.

However, there was some logic to the NTWG recommendation that the standards, although arbitrary and neither set nor documented as well as would be hoped, had acquired some meaning through usage and it might be well to first attempt to validate the existing standards rather than to reset them. Thus, the decision to do a validation study, although not following the specific recommendation, was at least arguably responsive to the recommendation.

The data from the validation study were arguably consonant with the interpretation that the validation committee found the original standards to be appropriate. However, the new committee did not define new standards, they only placed existing papers into categories based on a slight rewrite of the original definitions. Thus, it is debatable whether the standards were validated, or the ability of the committee to place papers into the pre-defined categories was validated.

The validation study was reasonably well documented, but there were some inadequacies in its documentation -- especially with respect to the (very minor) rewriting of the definitions which was accomplished prior to the validation committee doing its task.

Some additional comments related to standard setting

Strong and Sexton study: A study by Strong and Sexton (1997) presents evidence that is somewhat disturbing. They found, for example, that twenty-seven percent of the seniors in their study that were judged as Novice on the Kentucky assessment had ACT mathematics scores which ranged from 18 to 36. It is unknown why this sort of discrepancy between KIRIS and ACT scores exists but several possibilities come to mind: (1) non-valid KIRIS tests, (2) differential student motivation for the KIRIS and ACT, and (3) inappropriately set cut scores for the KIRIS mathematics examination.

Judgmental method for equating performance events: As has been mentioned, there was no judgmental standard setting performed on the initial performance events. Rather, it was assumed that the proportion of students performing at each classification level on the performance events was the same as the proportion at each level on the open-ended items. Equating (discussed elsewhere) across years was done through a common person equating model with students outside Kentucky. Based on problems with that method, it was decided that there should be an equating study using a judgmental method -- which was completed in January 1996. The NTWG concluded that this procedure "appeared too sensitive to changes in task difficulty to be used with this set of performance events." (March 1-2 1996 minutes, p. 2). Thus, performance events were not used in the accountability index for the 1994-95 school year.

While one can consider this primarily a difficulty in the equating process, it should be pointed out that the OEA panel did indicate the lack of any judgmental standard setting of the performance events and criticized the non-scientific approach used. This criticism was not responded to in the "validation" study -- which did not address the performance events standard setting.

Section 7. 1997 KIRIS Scores in Perspective

In mid December 1997, we added to our scope of work for this review analyses of the just-released spring 1997 KIRIS test scores for KY elementary, middle, and high schools. We were especially interested in what the test score changes for 1997 and their means of reporting by KDE mean in the context of our technical review. There turn out to be several important connections to topics and issues raised in this report.

We first present selected overviews of the 1997 scores:

Score increases across the board. Table 1, on the following page, shows percentage increases in academic KIRIS index scores by level of school and by subject area. The baseline is the Accountability Cycle 3 base year average -- 1995 and 1996 scores averaged together; the 1997 gains are expressed as percentage increases over these baseline figures.

A first result displayed in Table 1 is that all KIRIS scores were up between the Cycle 3 baseline and the Cycle 3 midpoint in 1997. A second result is that some subject index scores are up by very substantial margins -- elementary and middle school science scores are up by 27.9 and 26.9 percent respectively, and middle school and high school practical living/ vocational studies scores are up by 45.8 percent and 28.9 percent respectively. A third result is that at least one score increase appears to be extraordinarily large, at least when scanning the annals of academic achievement indicators known to the measurement profession -- namely the 60.7 percent increase in the average school reading score. We have more to say about these increases and the reading score gain in particular below. Finally, only elementary school social studies (with a 1 percent gain) and high school arts and humanities (with a 3 percent gain) fail to register double-digit increases.

Effects of Grade shift of some tests? Table 1 also points out those elementary and middle school subject areas that were tested at new grade levels for the first time this year. The underscored figures in the elementary column reflect those tests that were given in grade 5 rather than grade 4. Recall that an important goal of the equating agenda for 1997 was to estimate and implement adjustments so that the expectedly higher scores of 5th graders, in mathematics for example, would not show up as increases in subject area or school performance in the 1997 index scores. All of the underlined 5th grade scores appear to be below the middle of the general spread of score increases for the elementary school tests -- these four tests averaged 11.3 percent gains and the remaining three averaged 19.5. The fifth graders produced smaller gains than the fourth graders for 1997.

The middle school grade shift story is similar. The three middle school subject matter tests shifted from grade 8 to grade 7 for 1997 were reading, science, and writing. These three tests averaged gains of about 14.5 percent, while the other four tests at this level showed increased scores averaging about 21 percent -- so smaller middle school gains were recorded in tests shifted down to grade 7. Note that increases in various subject levels are highly variable at all levels.

School Progress Toward Goals. Table 2 on the next page shows another perspective on 1997 KIRIS scores. The most global characterization contained in this table is that more than 78 percent of Kentucky schools progressed to some degree beyond their Cycle 3 baseline scores by 1997 -- and only 21 percent declined. These

proportions are similar at all school levels. This pattern of increases versus declines is a marked turnabout in comparison to school index score changes between 1995 and 1996, where 64.5 percent of schools lost ground and only 35.5 percent of schools reported increases in their overall index scores (the 1995 and 1996 figures are shown as part of Table 3). Recall that 1996 was the second year of the accountability pair of years in Accountability Cycle 2, and thus was a year in which results had comparatively high stakes.

Also shown in Table 2 is the fact that more than one-third of the schools in Kentucky fully met their Cycle 3 achievement goal in the first of the two accountability years of Cycle 3, with nearly 38 percent of middle schools earning this distinction. And nearly 7 percent of Kentucky schools overshot their Cycle 3 goal by at least 100 percent!

A less dramatic (and possible misleading) reading of Table 2 indicates that 56.7 percent of KY schools are halfway to their 1997-98 goal as of 1997, another 22 percent of schools or so have made some progress, and another 21 percent lost ground. We say misleading because of the following: the actual goals for schools are comprised of average performances in the two accountability years. Other things being equal, a school that gets half way to its goal in the first year, e.g. 1997, must exceed that target by 50 percent in the following year if the average is ultimately to settle on the goal. In a sense, getting halfway home in KIRIS gets a school only one third of the way to its accountability goal. Perhaps mathematics test item designers should consider streaming KIRIS test score problems in to the open response items!

Additional Score Change Dynamics. Table 3 on the next page shows how KIRIS school index score changes went for selected different "types" of schools. The first pair of rows compares schools in the highest versus lowest quintile of school index scores as of the 1996 KIRIS test. High performing schools had index numbers greater than 50 in 1996; the low-performing fifth of all schools scored below about 38.3.

On average, schools in both groups increased, but the odds were better if a school began in the low quintile – with an 83 % chance of increasing as opposed to 61 percent for the schools which started out with higher scores. About the same percentage of both groups (about 37 percent) met their Cycle 3 accountability targets in the first year.

The second pair of rows compares schools who increased between 1995 and 1996 versus those whose KIRIS scores declined during the same years. About 2/3 of all schools had declined between 1995 and 1996. Among decliners, 75 percent reported progress on their Cycle 3 base score by 1997. Among 95-96 score increasers, the percent reporting progress on their Cycle 3 base score by 1997 was an even higher 85.3 percent.

The final section of Table 3 shows another view of the data. Among all schools that declined between their Cycle 3 baseline score and the midpoint score in 1997, the average 1996 index score was about 46. Among those which increased, the average 1996 index score was 43.7.

High Poverty Schools. Table 4 below shows school district level accountability index scores for the past three years, by incidence of student qualification for free lunches -- a commonly used indicator of the prevalence of poverty among school families. In one sense, the results are what we would expect. As of 1995, there is a simple relationship between KIRIS scores and poverty levels -- the less poverty, the

higher the scores. One reaction to this from the measurement community is that there must be some validity to the scores -- every other standardized test we have seen would be expected to bear the same relationship.

Table 4

Average District Accountability Index Scores and Gain Scores, Cycle 3 to Midpoint, By Quintile Percent of Students Qualifying for Free Lunch (a poverty indicator)

	95 mean	96 mean	97 mean	Mean Gain, 95 to 97
Quintile 1	49.67	47.42	53.04	3.37
Quintile 2	47.38	44.37	49.87	2.49
Quintile 3	44.93	43.36	47.92	2.99
Quintile 4	43.85	43.25	47.03	3.18
Quintile 5	41.04	40.27	43.43	2.39

(Quintile 1 is lowest percentage "poverty" and so on. Annual overall score differences and Quintile 5 vs. other quintile score differences all significant (ANOVA; Tukey HSD)

Source: Analysis provided by Kentucky Office of Educational Accountability, Dr. Kenneth Henry

---

Table 4 also indicates that Kentucky's poorest school districts, on average, made the least progress in their index scores between 1995 and 1997 -- about 2.3 index points on average. The lowest poverty school districts gains 3.37 points in comparison, or about 41 percent more.

Quintile 4 -- the districts nearing the top of the poverty distribution -- shows a substantial come-back in KIRIS scores between 1995 and 1997 -- a gain exceeding all but that of the least impoverished districts.

We note that in contrast to these data shown for districts and collapsed into 5 poverty classifications, KIRIS accountability index scores for schools show a tremendous range of levels as well as gain and decline scores. It thus appears that these quintile groupings (with their reasonably close, though systematically differing KIRIS scores) are surely hiding a great deal of variation across the schools within each quintile that would be worth exploring in future policy studies related to KIRIS.

Student Performance Change vs. KIRIS School Accountability Index Change Scores

Finally, we undertook two simulations to help illustrate particular dynamics of the accountability index scores and what changes in their levels from year to year or across accountability cycles might mean.

We began this exercise with heightened curiosity concerning the very large reported score gain for **high school reading** for 1997 -- about 61 percent when compared to the 1995-96 Cycle 3 baseline average scores (see Table 1). This is an enormous sounding statistic -- one that does not tend to find its way into discussions of systemic reading achievement gains from year to year in other standardized testing programs.

We spoke with Neal Kingston, of ASME, about this score gain; his response was simply, "We don't understand it." Dr. Kingston described checks and rechecks of data prior to the release of these figures as well as re-sampling of papers to provide yet one additional affirmation that the 1997 tests were properly equated. We obtained and reviewed an internal memo describing the additional equating simulation for grade 12 reading and were convinced that ASME and KDE had taken the large score increase and related questions seriously.

We then turned to the student distributions across performance levels in high school reading -- these are shown in Table 5 on the next page. These distributions give readers a picture of what changes from year to year are necessary to produce an index score change. In the case of 1997 reading, in amassing the 61 percent index score increase, there were:

16 percent (as opposed to 31 percent) classified as novices,
52 versus 59 percent as apprentices,
28 versus 9 percent as proficient, and
4 versus 0 percent classified as distinguished when 1997 is compared to 1996.

A fundamental question confronting these distributional shifts is how much does learning or student performance have to increase to produce them?

For the following simulation, we ignore other issues of central interest to Hambleton et al., to this review, and other technically-interested observers -- issues such as test-preparation versus learning, various sources of measurement error, or cohort differences. Here is how our mental experiment goes, as illustrated in Table 5:

Extreme scenario: Here we assume that in 1996, most Kentucky 11th graders within each of the novice, apprentice, and proficient categories for **reading** are crowded up against, but not quite qualified to cross, the cut-off points for the next higher classification. If these scores were done on the basis of one open-response item, we might say that instead of having been assigned 1s, 2s, and 3s, these students really were performing at levels corresponding to 1.9, 2.9, and 3.9. -- close, but not quite close enough for the next level. And the 1996 scores are thus produced.

As outlined in the section of Table 3 labeled "Hypothetical Advances Across Cut Scores," we see that in order to generate the 1997 scores, 15 percent of the 1996 novices, 22 percent of the 1996 apprentices, and 4 percent of the 1996 proficient readers would have to move up a classification. We assume in the extreme scenario, that this movement was achieved by replacing sufficient numbers of students respectively crowded at the top levels of each classification by students now in the very lowest reaches of the next higher classification. That is, our needed movement across classifications was achieved by sufficient numbers of 1997 students scoring well enough to be "proficient," even though their actual performance levels on our scale were about 3.1. (This in contrast to the 2.9s registered by last year's students). Analogous small achievement gains also account for sufficient movement between novice and apprentice and proficient and distinguished classifications.

Table 5 then shows the percentage increase in student score gains needed to make sufficient student classification changes happen -- based on the necessary numbers of students going from 1.9 to 2.1 and so on:

> The percentage student achievement score increase needed to produce the **61 percent** reading index score gain statewide is **4.66 percent.**

This is an extreme and unrealistic example, but it correctly points to crucial questions about the relationships between percentage gains in index scores and percentage gains in student achievement. Perhaps one of the least understood, and most consternation generating aspects of KIRIS is just the relationship between school score gains -- aggregate or by subject -- and measures of how much "more" students are learning. This shows up in questions about gains in and of themselves as well as in comparisons of KIRIS to NAEP, ACT, CTBS, and the like.

Table 6 repeats the Table 5 simulation, this time with an "average" scenario that does not rely on such narrow assumptions about students precipitously poised to move up the KIRIS ladder. In this simulation, it is assumed that all students moving from one classification to the next go from the mid-point of one classification to the next. (Recall, this is a simulated longitudinal study -- students do not move -- they are replaced by next year's students.) So instead of gaining the comparatively small amount required to go from 2.9 apprentices to 3.1 "proficients", students in this exercise are given credit for moving from 2.5 (the middle of the apprentice band) to 3.5 (the middle of the proficient band). This of course causes our estimates of how much learning has to change to yield the 61 percent index score to increase considerably in comparison to the extreme scenario above. But has all been set square? By no means. **Even under this average scenario, student achievement needs to increase only 23.3 percent to produce a 61 percent gain in the reading index score.**

Thus, KIRIS school score gains are fundamentally inflated to the extent that KY citizens think of KIRIS gains they way they might think of typical standardized test score gains. This inflation is independent of any other reasons suggesting that the gains might be inflated. (Potential cohort differences and potential equating problems come to mind.)

Section 8. Instructional Consequences of KIRIS

Hambleton, et al. claim in Recommendation 12:

> "There has been a shift toward process at the expense of content in the curricula and this shift needs to be reconsidered....This situation needs to be reviewed to be sure that the impact on instruction, while presumed by the Dept. of Education to be positive, is, in fact, positive. In addition, the implications on 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."

KDE responded to the issues addressed in this recommendation by the OEA Panel. For its own part, KDE moved in 1996 to produce a more detailed specification of subject matter content for KIRIS testing in a report called the Core Content for Instruction (KDE, 1996). An apparent goal of this document, itself based on previous content specification documents such as the "Valued Outcomes" which had evolved since the start of KIRIS, was to ensure that specific skills and content knowledge were being incorporated into KIRIS open response items. This would mean, for example, that science items would require some recall and reporting of specific scientific content, or that math items would require knowing specific formulas, whereas earlier tests seemed to be more concerned with quality of writing and presentation and less concerned with specific subject knowledge. KIRIS also began using multiple choice items as part of on-demand student assessment tasks in 1997. While these items have not been studied and are not part of the accountability index yet, they will probably contribute to the accountability index in Accountability Cycle 4 (1996 and '97 vs. 1998 and '99 score averages). These multiple choice items bring opportunities to broaden the range of content topics tested by KIRIS for a given student and in a given year across students.

While we do not favor the term "consequential validity," because validity has to do with the accuracy of inferences, not the consequences of actions, we do believe in the importance of considering the consequences of any assessment. Haertel and Wiley suggest in their response to the OEA Panel that "The real purpose of KIRIS is to improve the education of Kentucky's children–The Panel's "measurement aspects" are secondary." (p. 2). Cunningham responds that there are two problems with that philosophy:

> "First, it implies that it is possible to justify awarding large sums of money to some teachers and destroying the careers of others on the basis of a technically flawed test, as long as that test is useful instructionally. Such a claim is immoral if not illegal. Second, there is the assumption that teachers will alter their instructional methods in the desired direction, even when they realize that there is little connection between their teaching behaviors and the scores their students obtain. As teachers begin to realize that the test has no legitimacy and that it is too technically deficient to be influenced by how they teach, they will stop paying attention to it." (p. 2).

The draft KIRIS Accountability Cycle II Technical Manual (April 1997), contains a chapter on "Validity-related evidence" and another on the "Consequential Validity of KIRIS and the Kentucky Assessment Program." As the authors of the manual point out, "high consequential validity is probably not possible without high degrees of traditional reliability and validity..." (p. 14-2). This seems in agreement with Cunningham's point above.

The manual also points out that:

> "little formal research is available on the impact of the assessment program on classroom practice, teacher development, or support of educational reform in Kentucky. In addition, in the available research it is often difficult to separate effects of the assessment program from other aspects of educational reform." (p. 15-2).

However, the manual does report on the survey and interview-based RAND study by Koretz et al. (1996). The manual correctly quotes the Koretz et al. study to the effect that "about 40 percent of the teachers reported that the open-response items and portfolios have had a great deal of positive effect." (from p. x of Koretz, et al., 1996). However the technical manual did not provide the page reference. It is interesting to note that in the rest of the paragraph quoted, Koretz et al. went on to say that:

> "Performance events, however, were cited as having a great deal of positive effect only about half as frequently as open-response items or portfolios ... Finally, portfolios were cited as having had negative effects on instruction almost as often as having had positive effects." (p. xi).

We could review the Koretz et al. report in more detail and in fact offer additional perspectives from this report immediately below. However, in the interests of space, we first point out that this is an objective report that leaves us feeling that the positive consequences do not necessarily outweigh the negative ones. For example, they report that "large majorities of teachers reported making instructional changes consonant with the goals of the program ... However, portfolios also put pressure on the regular curriculum, and in response, teachers placed less emphasis on the mechanics of writing (in fourth grade) and on computation and algorithms (in eight-grade mathematics)." (p. xi). Koretz et al. report that the responses of teachers "suggest a fundamental tension between the individualization and flexibility that is desirable for instructional reform and the standardization that contributes to comparability of measurement across schools." (p. xii). "Educators also reported widespread efforts to align instruction with KIRIS." (p. xii). However, "almost 90 percent of teachers agreed ... KIRIS had caused teachers to 'de-emphasize or neglect' untested material." (p. xiii).

With respect to increased scores, Koretz et al. reported that "Principals and teachers also reported substantial reliance on 'direct test preparation'" and that "almost all teachers reported that students were given practice on the previous years' KIRIS items." (p. xiii). Further, the Koretz et al. report stated that:

> "Appreciable minorities of teachers reported questionable test-administration practices in their schools. About one-third reported that questions are at least occasionally rephrased during testing time, and roughly one in five reported that questions about content are answered during testing, that revisions are recommended during or after testing, or that hints are provided on correct answers." (p. xiii).

These findings, and others that could be interpreted as negative consequences were not reported in the Technical manual. The Technical Manual has headings stating that "Consequences: Is fair to schools," "Consequences: Is fair to students," and "Consequences: Does not have unintended, negative consequences." We believe the evidence for the first two is largely lacking and that there is evidence refuting the last one -- as reported in Koretz et al. -- a report the authors of the manual were aware of since they quoted from it.

Additional research on the Effects of KIRIS

Effects of KIRIS on Instruction. The KIRIS "high stakes" performance-based student assessment system is intended to assess student learning *and* to drive curriculum, instruction, and school administration to ensure all schools meet the "goals for the Commonwealth's schools" (KRS 158.6451). Studies have attempted to document the impact of KIRIS in several areas.

Target areas in search for KIRIS Effects:

1. Curriculum content and instructional focus
2. Instructional practices
3. Teachers' beliefs and attitudes
4. Classroom assessment practices
5. Student achievement and performance.

Most studies of KIRIS' impact on instruction have found that KIRIS has had a positive effect on instruction while also having some negative effects. The Koretz et al. study (1996) identified both positive and negative effects of KIRIS in terms similar to those brought out in findings in the annual KIER surveys of attitudes toward KERA. The most common positive comments from teachers reflect a belief that student writing and thinking skills have improved and the most common critical comments reflected the belief that portfolios are time-consuming and burdensome.

Positive perceptions. KIRIS' impact according to statewide surveys and recent studies extends into instructional strategies, classroom assessment, school expectations and teacher attitudes, and student performance. Over half of those surveyed believed

KERA has had positive effects: a statewide survey of attitudes toward KERA sponsored by KIER found that more than 50% of each group surveyed (educators, parents, members of the general public) agreed that: opportunities to learn have increased for students from disadvantaged homes and for students in special education; students' writing has improved in quality; schools have raised the expectations of all students; and, most schools are performing at higher levels.

Effects on Instruction. Recent studies are consistent with earlier conclusions (Appalachia Educational Laboratory, 1997) that KIRIS appears to be a driving force behind the instructional change in Kentucky. In the Koretz et al. study (1996), teachers report that their efforts to improve instruction and learning have increased. A "large majority of teachers agreed that the performance based components have had *more than a small positive effect* on instruction in their schools" (p. x). Virtually all teachers (93%) reported that they have focused at least a moderate amount on "improving instruction generally" in their efforts to improve scores on KIRIS -- 87% of teachers reported that they were encouraged to experiment in their teaching, and 43% responded that KIRIS had led to an increase in the degree to which they are encouraged to experiment.

Is KIRIS an effective agent of reform? Most principals and teachers, in the survey done by Koretz et al., were positive about KIRIS' value as an agent of reform. 77% said KIRIS has been useful for encouraging positive instructional change among teachers who are very resistant to making changes to their instruction; 24% said it has been very useful in this respect (p. 52). Over half (57%) agreed KIRIS has caused some teachers who are resistant to change to improve their instruction.

Curricular Focus. Reported changes in curriculum focus include primarily an increased emphasis on writing and the writing process, and also changes in time devoted to problem-solving activities. Also reported is a decrease in instructional time devoted to art, social studies, science, and reading. Some specific curriculum impact observations:

1. 90% of 4th-grade teachers and 87% of 8th-grade mathematics teacher reported focusing a "moderate amount" or "a great deal" on improving the match between the content of their instruction and the content of KIRIS (Koretz et al., p. 23). 69% of teachers reported that there was content they emphasized more because of KIRIS.

2. Relatively few teachers agreed that KIRIS has led teachers to focus more on: real life applications (6% and 24% of fourth- and eighth-grade teachers, respectively); hands-on activities (4% and 11%, respectively); and cooperative learning (10% and 3%, respectively) (Koretz et al., 1996, p. 28).

3. While only a minority of school principals reported a de-emphasis on untested materials, most teachers did report a de-emphasis on untested materials (Koretz et al., p. 24). For example, 88 percent of fourth-grade teachers agreed that "KIRIS has caused some teachers to de-emphasize or neglect untested subject areas," and 40% strongly agreed with that statement. About 86% of mathematics teachers agreed that KIRIS has caused some mathematics teachers to de-emphasize or neglect untested mathematics topics, and 43% strongly agreed this change had occurred.

   Teachers appear to take time away from a wide variety of content areas as well as classroom activities to accommodate KIRIS; just what teachers took time away from to accommodate the test is not clearly dictated by the test (Koretz et al., p.25). The subject areas for which most teachers indicated a decrease in instructional time since

KIRIS were art, social studies, science, and reading. 89% of the teachers indicated these changes were largely due to KIRIS (P.25).

Some areas of concern do not seem large in the Koretz et al. results. For example, of fourth-grade teachers, only 13% expressed concern about the time taken away from instruction to prepare for KIRIS. And only 9% reported attention to content had been reduced because of time involved in portfolios. And only 11% believed that content related to basic skills had decreased because of KIRIS (Koretz et al., p.28). Fewer than a fourth of eighth-grade mathematics teachers reported having reduced the content covered in order to do portfolios, and fewer than one teacher in ten at this level reported that math content instruction had been reduced because of time required for writing.

4. Koretz et al. (1996) stated that four-fifths or more of teachers reported increasing the emphasis or time in instruction devoted to problem-solving, communicating mathematics, and writing. Within language arts, most 4th-grade teachers reported an increase in emphasis on writing and a decrease in emphasis on spelling, punctuation, and grammar (p.25). Within mathematics, most 4th grade teachers reported they have increased the emphasis on mathematics communication and meaningful problem-solving. The only area of mathematics in which a significant percentage of fourth-grade teachers indicated a decrease in emphasis was number facts and computation.

5. Other effects of KIRIS on elementary activities included a decreased amount of time spent, in order to accommodate KIRIS, on recess (50%), organized play (43%), student performances (43%), student choice time (e.g., games and computer work - 43%), and class trips (39%) (Koretz et al., p.26).

6. **Expectations.** In the Koretz et al. study, about two-thirds of teachers reported that expectations for students have changed because of KIRIS (p.53). This appears to apply more to high-achieving students than to low achievers -- 24% of teachers reported that expectations had increased greatly for high-achieving students, compared to 16% reporting expectation increases for low-achieving students and 12% increasing expectations for special-education students.

7. **Assessment within the curriculum.** A majority of teachers report increasing their use of assessment formats other than multiple choice as part of their classroom assessment activities. Teachers report that portfolios led them to be more innovative in planning and instruction.

   Pankratz (1997) found KIRIS is having a major impact on the use of performance assessments in the classrooms of teachers at selected schools. The use of performance assessments varies significantly, even among teachers within the same school, (Matthews, 1995; cited in Pankratz, 1997) and is primarily in preparation for KIRIS tests rather than as an integral part of their daily instruction. Sixty five percent of KY's schools purchase "continuous assessments" (KIRIS-like tests that can be used for practice in the non-accountability grades). New teachers report higher use of performance assessment in instruction than more experienced teachers. (About 95% of teachers with 1-5 years experience report using performance events within units of study; 70% with 6-10 years of experience report such use; 50% of teachers with more than 10 years of teaching experience report such use.)

Additional Studies of KIRIS Instructional Impact

8. A 1995 study (Clifford, University of Louisville) surveyed 400 teachers in 1994 and found that 73% of respondents reported an increase or large increase in using problem-solving activities as an instructional strategy. About 76% of respondents reported an increase or large increase in using cooperative learning/group work as a teaching technique.

9. A 1993 University of KY dissertation (Vitali) summarized by Guskey, (1994; cited in Pankratz, 1997) was less positive. Vitali found that, "the performance-based assessment program ... resulted in only modest changes in teacher instructional practices." Guskey 'contended this reflected teacher beliefs that they knew how to teach to the new assessments and that at this early point in reform, they had not had either the time or opportunity to receive training in new instructional approaches. Guskey's conclusion is that new assessment systems, to have an impact on instruction, must be accompanied by high-quality professional development opportunities.

10. A 1995 Columbia University dissertation (Keene; cited in Pankratz, 1997) concluded that positive changes in instructional practices resulted from performance assessment. These included more writing by students, improved writing, more cooperative learning, more use of hands-on assignments, and less emphasis on textbooks.

11. In the KIER-sponsored study (The Evaluation Center, 1995), the Western Michigan University authors concluded that "portfolios have great instructional potential" (p.7) especially for improving student writing, but concluded that they are less reliable than other assessment formats.

Effects on Students

12. Students report that their education has changed since the passage of KERA. Students believed the greater emphasis on student writing occurred because of writing portfolios. They were positive about these changes, but some worried they might not be adequately prepared for college and their future careers (Gregg, 1994, p.1 (cited in Pankratz)).

13. The 1995 KIER-sponsored study (The Evaluation Center, 1995), reported that teachers, district assessment coordinators (DACs), and superintendents all reported that student writing had improved in the state. Teachers surveyed by Koretz et al. (1996) commented that student's writing and communication skills had improved (p.27).

    Minorities of teachers report increases in emphasis or achievement in thinking skills. In the Koretz et al. study, about 18% of fourth-grade teachers commented that emphasis on thinking skills had increased or that students' thinking skills had improved. Just under a fourth of eighth-grade mathematics teachers noted increased emphasis or increased student achievement related to thinking skills, and 14% in this group noted increases involving problem solving skills (p.28).

Effects Of Individual KIRIS Cognitive Components

The impact on instruction of particular KIRIS test components -- multiple choice items, open-response items, performance events, and portfolios -- has been of continuing interest. Open-response items, which entail a brief written response and are arguably the least "performance-based" of the three performance components, were cited by the greatest number of teachers in the Koretz et al. study (80%) as having had more than a small positive effect (*"moderate amount"* or *"a great deal"* of positive effect) on instruction in their schools.

14. **Multiple-choice items** were not as credited for having had a positive effect on instruction as the other cognitive components. 41% of teachers reported more than small positive effects of multiple-choice items.

15. **Open response items** were cited by 80% of teachers as having more than a small positive effect (*"moderate amount"* or *"a great deal"* of positive effect) on instruction. Open-response items received much stronger support in comparison to portfolios: 43% said they had a great deal of positive effects, while only 6% said they had negative effects.

16. **Portfolios and performance events** were cited by 2/3 of teachers as having had more than small positive effects on instruction in their schools. Portfolios were perceived by 40% of teachers to have "a great deal" of positive effects. But 30% of teachers stated that portfolios had a great deal of negative effects.

    Surveys have shown almost all educators in the state believe that portfolios have had the greatest impact on classroom activities of any portion of the assessment program.

Instructional effects of test components. The KIRIS components most frequently cited as having a *great deal* of positive effect are open-response items and portfolios. Similarly, open-response items and portfolios were cited as providing the most useful information for improving instruction in their classes. Many teachers reported (in Koretz et al., 1996) that the portfolio program had led them to be more innovative in planning and teaching (p. 52). In fourth grade, portfolios were cited by about half of the teachers as having had a great deal of positive impact. However, as Koretz et al. noted, a smaller majority of teachers also reported that the portfolio assessment had had more than a small negative effect on instruction, and about a quarter of fourth grade teachers reported that the portfolio assessment had had a great deal of negative impact.

Negative Instructional and Curricular Outcomes of KIRIS

Statewide surveys of attitudes toward KERA sponsored by KIER (Wilkerson & Associates, 1996; Pankratz (1997)) report a number of negative instructional and curricular consequences of KIRIS. These studies found that more than 50% of teachers agree that:

17. KIRIS emphasis on writing has reduced instruction in grammar, spelling, and punctuation.

18. Preparing writing portfolios takes too much teaching and learning time from other subjects.

19. The emphasis on the application of knowledge has reduced instruction on important content.

20. Teachers are spending too much time on students practicing for KIRIS test questions, (even though they recognize that such practice may serve also as a positive learning experience).

21. Also among the reported negative effects of KIRIS are stresses placed on teachers and principals. About half of the principals and more than half the teachers felt schools and students are subjected to "undue pressure" because of KIRIS. Teachers in the accountability grades reported the added pressure of portfolios caused changes such as "less emphasis on the mechanics of writing (in 4th grade) and (less emphasis) on computation and algorithms (8th grade) (Koretz et al., 1996).

Summary Appraisal: Ongoing study of the effects of KIRIS seem absolutely important. At many turns, the sponsors of KIRIS – KDE, ASME, the National Technical Working Group, offer reactions to technical concerns with KIRIS tests that effectively claim that the results are worth the limitations. The OEA Panel, and to some degree this 1998 review of KIRIS find little difficulty pointing to important technical limitations. If curricular and instructional effect is to be the response, the burden is on the responder to document such effects.

Section 9. Scorer Interviews

We undertook an exploratory set of interviews with KIRIS scorers as an added part of this review. The procedures used to score KIRIS tests are vital contributors to the integrity of the whole process – in fact, in the absence of accurate and consistent scoring of student tests, the rest of the technical considerations surrounding KIRIS become of little concern.

While the general procedures surrounding scoring seem almost second nature to participants in the KIRIS process, there is little actual documentation of the scoring procedures and processes that in fact took place in any given year that allows for external review. The issues and processes of central interest include the qualifications and training of scorers, the monitoring of scoring to see that scorers remain "on-standard" in classifying student work, and the monitoring of scoring to see that scorers are scoring papers in ways matching the scoring of the previous year's tests.

We decided to conduct a limited set of interviews with 1997 scorers as a limited probe of scoring processes. We confined this probe to scorers under the present or recent employ of ASME; this meant we focused on 1997 scorers of Grade 11 tests as well as past scorers who may have scored at any of the levels and subjects. The scoring sub-contractor scoring 1997 tests for grades 4-5 and 7-8, Data Recognition Corporation of Minnetonka, MN, declined to have its employees take part in our interviews. In the process of trying to elicit DRC participation, an official of DRC described their procedures in great detail, and the description was on target with expectations for the scoring process.

The ASME interview sample had two sources. One was based on 20 scorers nominated by the review team from a stratified random sampling process – we drew 10 scorers from a list of all Grade 11 reading scorers provided by ASME, selecting only those who worked at least 90 hours on reading tests in 1997. We drew another 10 scorers meeting the 90 hour minimum from among all other 1997 ASME scorers. ASME proceeded to get permission for interviews from about 14 of these 20 individuals and we interviewed 12.

A second interview pool was generated from a list of 1992-93 ASME scorers provided by the LRC. We eliminated names from this list of present employees and randomly selected 12 for interviews, with about 24 backup selections because we realized many from this list might either (1) have moved and be hard to reach, or (2) refuse to participate in the interviews based on our phone call request. As things turned out, only two individuals refused to participate.

Our basic interview guide included questions about scorer background, how long they worked for ASME and on KIRIS, what tests they scored, how they were trained, how they were monitored, and what they might suggest to guarantee accuracy of KIRIS scoring. Here is a summary of the interviews:

Background. Scorer backgrounds are highly variable. Many are teachers working in the summer. Some are teachers looking for work. Some are college and graduate students. All seemed to believe it takes a 2-year college degree at minimum to become an ASME scorer for KIRIS. Respondents felt that fellow scorers tended to be qualified and well trained, and that the occasional individual who does not perform is re-trained or weeded out. Scorers typically had worked for ASME for 3 years and if a present scorer

intended to keep at the job. They typically scored at least 2 subject areas and typically maintained a main focus on one subject area.

Training. Scorers in both groups described extensive training and checking of trainees to see that they could meet standards, such as 8 out of 10 trial papers exactly matching correct scores. Many described re-training if actual scoring problems emerge in quality checks.

Quality control while scoring. All scorers were aware of back-reading and had some, though varying, experiences with back readers who questioned their work. One respondent was sure that his/her work was only questioned by a back-reader if it was 2 points off, and not if it differed by 1 point. This was not echoed by other respondents.

A readers in one subjects, Grade 11 science, described the fact that some student papers were presented to them on computer screen for the first time in 1997 -- they had been scanned in. This respondent believed the papers chosen for scanning were the more legible papers. We had not seen any reference to scanning as part of scoring, and no references to discussions about whether this might have any impact on scores. The process does in fact present two distinct "stacks" of papers to readers -- the legible, and all others. There could be a global effect on scoring. KDE and ASME, in a follow up discussion, believe this respondent was confusing scoring papers from another state for ASME.

Many respondents expressed awareness of an "equating box", from which some the papers they scored were periodically drawn. The presumed that their work on these papers would enter the checking processes to compare 1996 with 1997 scoring standards.

Ability to classify papers. Respondents seemed comfortable with their ability to assign scores of 1, 2, 3, or 4 to papers. This was more difficult for some subjects and for some grade levels. Several respondents complained of "fuzzy" scoring guides. Two respondents, including a senior scorer (and back-reader) in science, thought that newly developed items produced not by ASME but by a sub-contractor had unusually "fuzzy" and "ambiguous" scoring guides. It was not clear to them, nor to us, that this fuzziness would contribute simply noise in the data, or bias scores up or down. (More on scorer pleas for guideline clarity is reported below.)

Scorers reported receiving uniformly high levels of support from table leaders in cases where they could not decide that a paper deserved one score versus another.

Did standards appear about the same this year as last, or over the years? In general, multiple year scorers believed that standards had remained the same. Two notable exceptions were recorded:

1) The senior, science scorer and table leader claimed that a criterion blocking science papers from receiving a score of 4 (Distinguished) in the past was relaxed for the first time in 1997. This is what he/she described as the "no-fault clause." This was a stipulation that even an outstandingly distinguished paper may not receive a 4 if it contains a single error of fact. As of 1997, in the words of this respondent, such a paper now COULD receive a 4. Incidentally, the percentages of Grade 11 science papers receiving 4s was 0 in 1996 and 4 in 1997.

2) Another respondent had scored KIRIS papers for 6 years, including the inaugural year. This individual was quite certain that student papers during 1991-92 showed such limited performance levels that scorers once in awhile bent over backwards to assign a 3 or a 4 just to be sure there would be a few of these scores represented. Reflecting further, this scorer was sure that some of the scores getting 3s and 4s back in the first year or two would NOT get the same scores this year.

The upshot of this, admittedly singular, long view of KIRIS is that the scorer was making a case that scoring standards had in fact crept toward being more stringent over time. This pattern bears an interesting relationship with claims of inflation of KIRIS scores over time -- it suggests that the scores might have gone up even further under more consistent scoring over the full 6 years.

This observation suggests to us an interesting study that KDE should sponsor, provided papers remain on file and can be suitably sampled. A good retrospective test of the consistency of scoring and the growth of student performance would be to sample substantial numbers of student papers in a subject, or subjects, for each of 6 years and to engage in a re-scoring analysis. The results could be held up against reported annual index scores on this test; typical "3" papers from 1993 could be held up against typical "3" papers from 1997 to convey consistency or lack thereof, and so on.

Scorer work load: Scorers were very consistent on our question of workload. Most had acceptable recall of this and claimed to score somewhere between 15 and 25 papers an hour -- no more unless they happened to be in the midst of a run of 4th grade papers where students wrote very little.

Other observations. Several scorers noted something to the effect that "too many" of the papers they scored were limited to an off topic remark or no response. This is an indicator that some students do not care how well they do on KIRIS -- a problem experienced to some extent with all tests, and an issue about which we have no data to compare KIRIS with other tests.

Suggestions for Improving Quality of KIRIS scores. We received a uniformly consistent suggestion in response to this question. This was the plea for ever-clearer scoring guidelines or rubrics.

End of Report

(References and Bibliography of Reviewed Documents Follow)

References

Advanced Systems for Measurement and Evaluation, (May, 1996). KIRIS Standards Validation Study: Mathematics, Reading, Science & Social Studies.

Advanced Systems for Measurement and Evaluation, (October, 1996). KIRIS Standards Setting Study: Arts & Humanities, Practical Living/ Vocational Studies.

Awbrey, (1996). An Advanced Systems Appraisal. will complete.

Boyson, T. C., Kentucky Department of Education, (February, 1994). KIRIS 1992-93 Technical Report *Appendices.*

Boyson, T. C., Kentucky Department of Education, (August, 1994). KIRIS 1992-93 Technical Report..

Boyson, T. C., Kentucky Department of Education, (January, 1995). KIRIS Biennium 1 Technical Manual:
Based on data analysis from the 1991-92 through 1993-94 School Years.

Condon, W., & Hamp-Lyon, L. (1991). Introducing a portfolio-based writing assessment: Progress through problems. In P. Belanof & M. Dickerson (Eds.), *Portfolios: Process and product* (pp. 231-247). Portsmouth, NH: Heinemann.

Cunningham G. K., (March, 1997). KIRIS: A Critical Analysis.

Cunningham, G. K., (1997). The Reliability and Validity of KIRIS: A report prepared for the Task Force on Public Education Assessment and Accountability Issue Group.

Gearhart, M. & Herman, J. L., (1995). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability (CSE Evaluation Comment). Los Angeles: University of California, Center for Research on Evaluation, Standards and Student Testing.

Haertel, E. H., & Wiley, D. E., (June, 1995). Response to the OEA Panel Report "Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991- 1994."

Haertel, E. H., (August, 1995). Personal Communication.

Hambleton, R. K. et al., (June, 1995). Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991 - 1994. Final Report.

Harp, L., (August, 1997). Three Districts Sue Calling Testing System Unfair. Lexington Herald.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1* (3), 201-224.

Hill, R., (November, 1991 - October, 1994). File containing memos, letters and various documents.

Kentucky Department of Education, (August, 1995). Equating Plan: Accountability Cycle II. Open response tests in Mathematics, Reading, Science, and Social Studies.

Kentucky Department of Education, (September, 1995). KDE Response to June 1995 Resolution of the LRC Subcommittee on Assessment and Accountability.

Kentucky Department of Education, (1996). Core Content for Assessment Version 1.0.

Kentucky Department of Education, (1996). World Class Standards for World Class Kids. Performance Level Guides for Open-Response Common Items, Grades 4, 8 & 12.

Kentucky Department of Education, (April, 1997). KIRIS Accountability Cycle II Technical Manual: An analysis of data from 1992-93 through 1995-96 school years.

Kentucky Department of Education, (Revised September, 1997). Noncognitive Guidelines: Transition to Adult Life Success Rate, (high school graduates). Dropout Rate (grades 7 - 12).

Kingston, N. and Dings, J. (1995). Estimating the Accuracy of Complex School Accountability Decisions. Frankfort, KY: Kentucky Department of Education.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont Portfolio Assessment Program* (CSE Tech. Rep. 355). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Koretz, McCaffrey, Klein, Bell & Stecher, (1993). will complete.

Koretz, Barron, Mitchell & Stecher, (1996). (Survey of KY teachers on KIRIS related practices; title page missing from our copy).

National Technical Working Group, (March, 1997). Meeting Minutes.

Nitko, A. J., (May, 1997). Equating Study: (KIRIS -- CTBS, CAT, CTBS5) A Description and Comparison of the Comprehensive Tests of Basic Skills, The California Achievement Tests, the Terranova (CTBS5), and the Kentucky Instructional Results Information System Assessments.

Nitko, A. J., et al., (1997). How Well are the Kentucky Academic Expectations Matched to the KIRIS 96 Assessments, CTBS and CAT?

Office of Education Accountability, (1997). Accountability Index of Documents.

Office of Education Accountability, (February, 1997). Performance Events Chronology.

Pankratz, R., (1997). Research Related to KIRIS's Impact on Instruction: Excerpts from Selected Studies and Reviews on Research.

Petrosko, J. M., (May, 1997). Assessment and Accountability.

Stecher, B. M., & Hamilton, E. G. (1994, April). *Portfolio assessment in Vermont, 1992-93: The teachers' perspective on implementation and impact.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Strong S., & Sexton, L., (1995-97). Series of Research Articles and studies regarding Performance Assessment.

Webb, N. M. (1994) *Group collaboration in assessment: Competing objectives, processes, and outcomes* (CSE Tech. Rep. 386). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Wilkerson and Associates, (1995). A Review of KIRIS for Kentucky Institute for Educational Research.

Wise., L. L., (November, 1997). KIRIS 1997 Grade Shift Adjustments.

Additional Reviewed Documents

Alkin, M., (November, 1997). Review of eight OEA Panel recommendations and suggested documentation pertaining to each recommendation.

Advanced Systems for Measurement and Evaluation, (May, 1996). KIRIS Standards Validation Study: Mathematics, Reading, Science & Social Studies.

Advanced Systems for Measurement and Evaluation, (October, 1996). KIRIS Standards Setting Study: Arts & Humanities, Practical Living/ Vocational Studies.

Advanced Systems for Measurement and Evaluation , (Jan, 1997). Response to Dr. Penney Sanders Performance Event Component of KIRIS.

Boyson, T. C., (February, 1994). KIRIS Supporting Technical Information for 1992-93 Accountability Reports.

Boyson, T. C., (February, 1994). KIRIS 1992-93 Technical Report *Appendices.*

Catterall, J. S., (1997). KIRIS tech review analysis, Sketch Plan. Eight OEA Report Recommendations to be assessed.

Cody, W. S., Kentucky Department of Education, (September, 1995). 1995-96 District Assessment Coordinator Implementation Guide. (Also 1996-97 Guide).

Cody, W. S., Kentucky Department of Education ,(September, 1997). 1997-98 District Assessment Coordinator Implementation Guide.

Coogle, F., (1995-96). Tabular and Graphical Data.

Coopers & Lybrand (September, 1997). Outline of ASME Contracts 1991-95, (detailed outline of documentation KDE should have received).

Coopers & Lybrand (September, 1997). Outline of ASME Contracts 1996-97, (detailed outline of documentation KDE should have received).

Coopers & Lybrand, (November, 1997). Kentucky subcontract agreement.

Data Recognition Corporation, (July, 1997). 1996 and 1996 DAC/Principal questionnaires and results -- re logistics and administration of tests.

Dings. J., (November, 1997). Six Equating-related documents.

Edwards, B., (December, 1995). Summary of KIRIS Standards Validation: Process.

The Evaluation Center Western Michigan University, (January, 1995). An Independent Evaluation of KIRIS.

Gong, B., (October, 1997). Memorandum to Regina Wenz.

Gong, B., (November, 1997). Transmittal letter, 11/26

Gong. B., (December, 1997). KIRIS National Technical Working Group Contacts.

Haertel, E. H., & Wiley, D. E., (June, 1995). Response to the OEA Panel Report "Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991- 1994."

Haertel, E. H., (August, 1995). Personal Communication.

Hambleton, R. K. et al., (June, 1995). Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991 - 1994. Final Report.

Harp, L. (August, 1997). Three Districts Sue Calling Testing System Unfair. Lexington Herald.

Hill, R., (November, 1991). Update on NAEP's plans for math assessment; brief discussion of comparability issues.

Hill, R., (December, 1991). Issues related to the drawing of samples to create norms for the scrimmage tests.

Hill, R., (January, 1992). Issues related to students being absent from performance events testing.

Hill, R., (February, 1992). Issues related to KIRIS reports: defining "World Class Standards"; operationally defining the four performance levels; outline of proposed reports and value judgments that must be made before reports can be produced; plans for aggregating data to the school level.

Hill, R., (February, 1992). Plans for administration of performance events tests.

Hill, R., (July, 1992). Issues to be resolved before Accountability Index can be computed.

Hill, R., (July, 1992). Table summarizing performance of blacks and whites on 1991-92 grade 12 results.

Hill, R., (August, 1992). Answers to questions about how we planned to handle the data for the schools whose answer booklets burned in the Federal Express truck fire.

Hill, R., (September, 1992). Using proposed procedures for computing baselines and thresholds, examples of how a hypothetical school might be successful.

Hill, R., (November, 1992). Arguments for counting distinguished performance as a score of 120, not 140.

Hill, R., (November, 1992). Cover letter for Threshold's paper; first mention of procedures for recasting data (estimating statewide data from non-representative samples).

Hill, R., (November, 1992). Plans for the adjustment of portfolio scores, based on audit.

Hill, R., (November, 1992). Rationale to include portfolio data in Accountability Index.

Hill, R., (December, 1992). Response to Tom's concerns re: article in Education Week about Rand's criticism of Vermont portfolio's.

Hill, R., (December, 1992). Proposed schools to audit.

Hill, R., (December, 1992). Suggestions from Ed Haertel about equating performance events.

Hill, R., (January, 1993). Estimates on the reliability of thresholds.

Hill, R., (February, 1993). Motivations of Grade 12 students.

Hill, R., (March, 1993). Vision of where the assessment will be in 5 years.

Hill, R., (March, 1993). Schedule for implementation of KIRIS (need to slow down implementation of portfolios).

Hill, R., (May, 1993). Justification for choosing the 20th day of enrollment for accountability purposes, and alternatives.

Hill, R., (September, 1993). Explanation of use of "0" and "1" in 1992-93 scoring guides .

Hill, R., (September, 1993). Percentages of students leaving questions blank.

Hill, R., (September, 1993). Data and thoughts about the summer re-scoring of 1992-93 writing portfolios; arguments against including on-demand writing scores in the accountability index.

Hill, R., (October, 1993). Data from summer re-scoring and auditing of portfolios.

Hill, R., (October, 1993). Tables comparing how schools would be classified if on-demand writing were included in computation of index vs. how they would be classified if on-demand writing were not used; first findings that selection of schools to be audited was going to be successful.

Hill, R., (October, 1993). Why we excluded the matrix-sampled questions from scrimmage reports.

Hill, R., (November, 1993). Suggestion to score portfolios explicitly for quality, effort and improvement.

Hill, R., (November, 1993). Data on comparing New Hampshire audit scores to Writing Advisory Committee; data on scoring copies vs. originals.

Hill, R., (November, 1993). Letter of transmittal to accompany tables showing correlation between 1992 and 1993 scores.

Hill, R., (December, 1993). Changes in scores for items used in common in 1992 and 1993.

Hill, R., (December, 1993). Numbers of forms on which a score of "Distinguished" was possible.

Hill, R., (December, 1993). Impact on results of making a maximum score "Distinguished."

Hill, R., (December, 1993). Follow-up and correction to memo of 12/17.

Hill, R., (March, 1994). Response to complaints to Tom Saterfiel from Greenwood High School that their audited portfolio scores were unfair.

Hill, R., (March, 1994). Data on a proposed procedure to estimate the reliability of writing portfolios.

Hill, R., (March, 1994). Arguments why auditing of portfolios should be done in New Hampshire.

Hill, R., (March, 1994). Coefficient alpha for multiple choice (m.c.) and open-ended (o.e.) 1993 KIRIS tests; correlations of m. c., with o.e., mean scores on m.c. for each performance level (determined by scores on o.e.).

Hill, R., (March, 1994). Information on the reliability of change scores.

Hill, R., (March, 1994). Response to concerns of one superintendent on the lack of correlation between KIRIS on-demand writing scores and Educational Record Bureau writing assessment.

Hill, R., (March, 1994). Contributions of each component in the accountability index to error variance.

Hill, R., (April, 1994). Listing of major KIRIS accomplishments and vulnerabilities.

Hill, R., (April, 1994). Alternative means of computing two-year averages (to handle missing data on performance events).

Hill, R., (May, 1994). Initial thoughts about equating grade 11 tests to grade 12.

Hill, R., (June, 1994). Issues related to the year-to-year re-scoring of open-response questions (to ensure equality of scoring across years).

Hill, R., (July, 1994). Possible ways of changing writing assessment so that on-demand writing can be reported to parents.

Hill, R., (July, 1994). Procedures for equating matrix-sampled questions.

Hill, R., (July, 1994). Reliabilities of tests at the student-level when m.c. and o.e. data are combined.

Hill, R., (July, 1994). Thoughts about including the m/s test questions in a student's score.

Hill, R., (July, 1994). Problems we will encounter in the auditing of 1994 writing portfolios.

Hill, R., (July, 1994). The relationship of KIRIS performance levels to classroom grades.

Hill, R., (August, 1994). Letter of transmittal to accompany two papers.

Hill, R., (August, 1994). Preliminary results for 1994, with cautions about interpretation.

Hill, R., (August, 1994). Updated data for Tom Saterfiel memo.

Hill, R., (August, 1994). Response to Tom Saterfiel memo about our procedures for computing reliability.

Hill, R., (August, 1994). Generalizability indices when items are considered random.

Hill, R., (August, 1994). Comparison of three years' data of original vs. re-scoring writing portfolio data; improved accuracy of schools audited in 1993.

Hill, R., (August, 1994). Update on re-scored writing portfolio data; analysis of accuracy of scoring by re-scoring method chosen.

Hill, R., (August, 1994). Update on re-scored writing portfolio data.

Hill, R., (Oct, 1994). Data related to the low reliability of grade 4 science data.

Hoffman, G., (September, 1996). Initial Draft Research Plan, Deliverable 3.1.9.1 of ASME Contracts 1996-97.

Innes, R., (1997). Various Reports, Studies, and data tables regarding KIRIS.

Jorgensen, M., (June, 97). ETS Paper: "High Stakes Assessment Undermines School Reform.

Kentucky Board of Education, (April 1997). Findings of Fact, Conclusions of Law, Recommended Order and notice of Exception and Appeal Rights.

Kentucky Citizen Digest, (December, 97). Comparison of 1990-97 test scores.

Kentucky Education System, (October., 1997). Code of Ethics for Appropriate Testing Practices for School and District Personnel.

Kentucky Department of Education, (No date). Grading Teachers, Grading Schools (Kingston, Reidy Chapter)

Kentucky Department of Education, (1991-97). Contracts between Kentucky and ASME.

Kentucky Department of Education, (June, 1993). Report of the visiting Committee on KIRIS

Kentucky Department of Education, (1994-5). Grade 4 Data for KIRIS.

Kentucky Department of Education, (1994-5). Grade 8 Data for KIRIS.

Kentucky Department of Education, (1994-5). Grade 11 Data for KIRIS.

Kentucky Department of Education, (1995). Letter to Superintendents.

Kentucky Department of Education, (August, 1995). Equating Plan: Accountability Cycle II. Open response tests in Mathematics, Reading, Science, and Social Studies.

Kentucky Department of Education, (April, 1995). Estimating the Accuracy of Complex School Accountability Decisions.

Kentucky Department of Education, (November, 1995). National Technical Advisory Committee.

Kentucky Department of Education, (November, 1995). 1994-95 Mathematics Portfolio Scoring Analysis Final Report .

Kentucky Department of Education, (1996). Core Content for Assessment Version 1.0.

Kentucky Department of Education, (1996). Kentucky Writing Portfolio Project & School Visit Summary Report , a KDE/AEL Collaborative study on writing portfolio improvement.

Kentucky Department of Education, (April, 1996). Effects of Students and Tasks on Gain scores Used in Complex School Accountability Decisions.

Kentucky Department of Education, (May, 1996). Livingston County Court Case.

Kentucky Department of Education, (September, 1996). 1996 Writing Portfolio Audit: Final Report.

Kentucky Department of Education, (September, 1996). Kentucky 1996-96 Grade 4 & 8 Re-scoring and Equating Analysis.

Kentucky Department of Education, (February, 1997). Summer 1994 Mathematics Portfolio Review Study Final report.

Kentucky Department of Education, (April, 1997). Changes in Spelling, Capitalization, Punctuation, and Subject Verb Agreement Skills Under the Kentucky Education Reform Act.

Kentucky Department of Education, (May, 1997). Comparing Student Performance on KIRIS with Other Indicators.

Kentucky Department of Education, (June, 1997). Minutes from the National Technical Working Group Meeting

Kentucky Department of Education, (June, 1997). ASME Invoice.

Kentucky Department of Education, (July, 1997). Alternate Portfolio: Scoring Consistency, Accuracy, and Recommended Scoring Operational Changes

Kentucky Department of Education, (July, 1997). Performance of Students with Disabilities on the KIRIS On-Demand Assessment: 1 992-93 Through 1995-96

Kentucky Department of Education, (Revised September, 1997). Noncognitive Guidelines: Transition to Adult Life Success Rate, (high school graduates). Dropout Rate (grades 7 - 12).

Kentucky Department of Education, (September, 1997). Validity Research Plan for the Final Component of KIRIS.

Kentucky Department of Education, (September, 1997). The relationship between School Gains in 8th Grade KIRIS Scores and Instructional Practices In Mathematics, Science and Social Studies.

Kentucky Department of Education, (November, 1997). 2-sentence scorer-agreement statement.

Kentucky Department of Education, (November, 1997). Kentucky Mathematics Portfolio Pilot Scoring Facets Analysis.

Kentucky Department of Education, (November, 1997). Accountability Index standard error source comment. (refers to Kingston and Kingston Dings 1996 papers).

Kentucky Department of Education, (December, 1997). Kentucky School and District Accountability Results. Accountability cycle 3 midpoint report (1994-95 to 1996-97).

Kentucky Department of Education, (December, 1997). Kentucky School and District Accountability Results. Accountability cycle 3 midpoint report (1994-95 to 1996-97). This is the briefing packet.

Kentucky Department of Education, (1997). KIRIS Validation.

Kentucky Department of Education, (1997). KIRIS Standards.

Kentucky Department of Education, (1997). KIRIS Appendices.

Kentucky Department of Education, (1997). Sharpen you Child's Writing Skills.

Kentucky Department of Education, (1997). Major Changes for 1996-97 with Regard to Assessment and Accountability.

Kentucky Herald Leader, (December, 97). Report on KIRIS scores from 93-97.

Kingston, N., (1995). Memorandum and other documents regarding KIRIS.

Lindle, J. C., Petrosko, J. M., & Pankratz, R. S., (May, 1997). 1996 Review of Research on the Kentucky Education Reform Act.

Millman, J., (1997). Grading Teachers, Grading Schools: Is Student Achievement a Valid Education Measure.

National Technical Working Group, (March, 1997). Meeting Minutes.

Newspaper Reports, (October, 1997). Wall Street Journal & USA today.

Nitko, A. J., (May, 1997). Equating Study: (KIRIS -- CTBS, CAT, CTBS5) A Description and Comparison of the Comprehensive Tests of Basic Skills, The California Achievement Tests, the Terranova (CTBS5), and the Kentucky Instructional Results Information System Assessments.

Nitko, A. J., et al., (1997). How Well are the Kentucky Academic Expectations Matched to the KIRIS 96 Assessments, CTBS and CAT?

Office of Education Accountability, (1997). Accountability Index of Documents.

Office of Education Accountability, (February, 1997). Performance Events Chronology.

Pankratz, R., (1997). Research Related to KIRIS's Impact on Instruction: Excerpts from Selected Studies and Reviews on Research.

Petrosko, J. M., (May, 1997). Assessment and Accountability.

Picus, L. 0., (February, 1996). Estimating the Costs of Student Assessment in North Carolina and Kentucky : A State-Level Analysis.

Picus, L. 0., (March, 1997). Estimating the Costs of Alternative Assessment Programs: Case Studies in Kentucky and Vermont.

Public Education Task Force, (March, 1997). Assessment and Accountability Issue Group. Meeting No. 1.

Public Education Task Force, (April, 1997). Assessment and Accountability Issue Group, Meeting No. 2.

Public Education Task Force, (April, 1997). Assessment and Accountability Issue Group, Meeting No. 3.

Public Education Task Force, (May, 1997). Assessment and Accountability Issue Group, Meeting No. 4.

Public Education Task Force, (May, 1997). Assessment and Accountability Issue Group, Meeting No. 5.

Public Education Task Force, (June, 1997). Assessment and Accountability Issue Group, Meeting No. 6.

Public Education Task Force, (July, 1997). Assessment and Accountability Issue Group, Meeting No. 7.

Public Education Task Force, (August, 1997). Assessment and Accountability Issue Group, Meeting No. 8.

Rand, (1996). Impact of KERA and KIRIS on Education in Kentucky.

Sanders, K. P., (November, 1997). E-Mail to Richard Innes

Sexton, L., & Strong, S., (December, 97). A Validity Study of KIRIS. This fax was sent by Doug Terry/Ken Henry in the Office of Educational Accountability.

Smith, D. C., (Oct, 1997) Assessing Race and Gender Subgroup Performance In KIRIS Accountability Cycle 2.

Strong S., & Sexton, L., (1995-97). Series of Research Articles regarding Performance Assessment.

Supovitz, J., & Brennan, R. T., (1997). Equitability of Portfolio Assessment Relative to Standardized Tests.

Various Authors, (1996 - 1997). Critical Standards, Validity and Equating Documents.

Wise, L. L., (June, 1997). Merging ASVAB and KIRIS On-demand Scores: Report of Preliminary Results.