Office of Education Accountability

Legislative Research Commission

Kentucky State Testing for Education Accountability: An Examination of Security-related Threats to Making Valid Inferences and Suggested Best Practices

Research Report No. 385

Prepared by

Marcia Ford Seiler, Director; Keith White, PhD; Ken Chilton, PhD;
Albert Alexander; Brenda Landy; Deborah Nelson, PhD; Sabrina Olds; and Pam Young

Kentucky State Testing for Education Accountability: An Examination of Security-related Threats to Making Valid Inferences and Suggested Best Practices

Project Staff

Marcia Ford Seiler, Director
Keith White, PhD
Ken Chilton, PhD
Albert Alexander
Brenda Landy
Deborah Nelson, PhD
Sabrina Olds
Pam Young

Research Report No. 385

Legislative Research Commission
Frankfort, Kentucky
lrc.ky.gov

Accepted November 15, 2011, by the
Education Assessment and Accountability Review Subcommittee

Paid for with state funds. Available in alternative format by request.

Foreword

In December 2010, the Education Assessment and Accountability Review Subcommittee directed the Office of Education Accountability to examine Kentucky's state testing data validation processes. This study, with guidance and input from the National Technical Advisory Panel on Assessment and Accountability (NTAPAA), provides a description of Kentucky's data validation processes at the development, implementation, and post-testing levels, for the Kentucky Core Content Test and ACT's Educational Planning and Assessment System instruments. Staff extends gratitude to all who assisted in this study's development and completion and send special thanks to NTAPAA for its wisdom, guidance, direction, and support.

Robert Sherman
Director

Legislative Research Commission
Frankfort, Kentucky
November 2011

Contents

List of Tables

List of Figures

Summary

Education tests and assessments are an integral part of the teaching and learning process. Results from education tests and assessments often are the sole pieces of information used for making judgments about how well students are learning and how effectively teachers are teaching. In addition, resource-related decisions hinge on performance as measured by these instruments at the federal, state, and local levels. It is essential that the information garnered from education tests and assessments is as accurate as possible. As Kentucky transitions to a new accountability model with greater reliance on education tests and more rigorous standards, the likelihood of potential inappropriate or unethical testing practices may increase. Instances of cheating and similarly inappropriate practices may appear as they have in Georgia, Pennsylvania, and Washington, DC. All stakeholders must have confidence in the inferences they make from education test and assessment results, and breaches in test security pose a threat that cannot be ignored.

Threats to test reliability and validity exist at every level of the testing and assessment process—from development to the final score report. Testing instruments should align with professionally developed standards, and test items must be tested to make sure the items can consistently measure what they were designed to measure. Additionally, educators must understand how to administer education tests and assessments including before-, during-, and after-test protocols. Finally, appropriate follow-up activities such as a review of results and item analyses aimed at score accuracy should occur.

One major threat to making valid inferences based on education test and assessment data occurs when test scores are the result of cheating or other inappropriate practice that is not the result of student learning. Specifically, test score gains that exist because of potentially bad or inappropriate practice on the part of education professionals pose a tremendous threat to making valid inferences. Test scores may be artificially increased by blatant answer sheet manipulation and pre-, during-, and post-test improprieties.

The Council of Chief State School Officers (CCSSO) and the Association of Test Publishers (ATP) released their *Operational Best Practices for Statewide, Large-Scale Assessment Programs* in 2010. Among their 10 key best practice guidelines were

- outside audits of test security practices,
- clear focus on test administration parameters,
- never-ending protection of content,
- employment of strong test user agreements,
- need for plans and systems to vigorously pursue rule violators,
- plans and systems for forensic analyses of test- and assessment-related data, and
- development of comprehensive security breach plans of action.

The common themes in each recommendation are forethought and strategy. Proactive states need to develop comprehensive and clear policies regarding how they plan to manage test and assessment integrity issues at each step. While it is impossible to eliminate all inappropriate practice, cheating, or other threats to making valid inferences, an approach based on strategy over reactivity is crucial.

The Kentucky Department of Education (KDE) has a relatively comprehensive system for ensuring the reliability and validity of all instruments that are parts of its education accountability model. The department addresses reliability and validity threats at development and during and after phases of testing including appropriate alignment and reliability and validity studies. In addition, KDE conducts, in conjunction with test vendors, the appropriate professional development activities. Further, KDE has in place a report and review mechanism for when instances of inappropriate education test and assessment practice occur.

The major weakness in Kentucky's education testing system is that testing irregularities or instances of inappropriate practice rely on whistle-blowers and coincidental discovery of irregular or outlier test scores. KDE has in place some of the components recommended by CCSSO and ATP. Lacking in Kentucky are outside audits of test security practices, plans and systems designed to vigorously pursue rule violators, plans and systems for forensic analyses, and overall plans for what to do when security breaches occur. While most testing experts agree that the CCSSO and ATP recommendations are valid, there is concern about the return on investment of enhanced security measures. Nevertheless, there are some inexpensive strategies that may be employed to address cheating-related threats to making valid inferences.

Test security experts agree that one of the least expensive ways to reduce inappropriate practices, such as cheating, is to clearly communicate with education professionals and students what is expected of them and that there is a routine process in place to identify and follow up on questionable cases. In short, the theory is that education professionals are less likely to exhibit unwanted behavior if they know said behavior will be monitored and questioned—much like the threat of an Internal Revenue Service audit to reduce inappropriate income tax practices.

Recommendation 3.1
The Kentucky Department of Education should continue to augment training for district assessment coordinators, building assessment coordinators, and proctors to include clear communication of post-test analyses aimed at identifying outliers along with the consequences of being investigated and found guilty of unethical or inappropriate testing practices.

Further, a strong monitoring system should include a mechanism for identifying and tracking specific testing conditions and individuals responsible for those conditions.

Recommendation 3.2
The Kentucky Department of Education should track district assessment coordinators, building assessment coordinators, and assessment proctors via a unique identifier to link those education professionals with specific testing events.

A system aimed at identifying and addressing inappropriate practice should have at its core a proactive, strategic, evidence-based framework. KDE does not regularly conduct an analysis of education test and assessment scores aimed at identifying conspicuous or improbable gains and increases (outliers). An Office of Education Accountability (OEA) analysis using a common outlier methodology yielded 25 schools that exhibited exceptional growth from school year 2009 to school year 2010 when Kentucky Core Content Test scores were examined. While it is not expected or implied that these schools achieved gains via inappropriate practice, statistically improbable growth from one year to the next warrants further attention in the form of follow-up analyses and more comprehensive fact-finding endeavors.

Recommendation 3.3
The Kentucky Department of Education should build into the education test and assessment data plan, through third-party vendors/agencies, internally, or through augmented existing contracts or study plans, analyses commensurate with available funding aimed at identifying outliers and other data irregularities.

While Kentucky has many components necessary to produce valid and reliable education tests and addresses potential threats to making valid inferences, an agency-specific plan of action detailing steps to take in the instance of a wide-scale education testing security breach is needed. According to an OEA survey of state testing leaders from across the US, approximately 75 percent of respondents have a system in place to manage and investigate allegations of widespread inappropriate testing practices.

Recommendation 3.4
The Kentucky Department of Education should augment the current formal process for attending to test violation allegations and complaints with comprehensive plans commensurate with available funding for addressing wide-scale test security allegations with specific state agency roles defined.

The first step in developing a comprehensive plan that would allow Kentucky to align itself with best practices in terms of education test security may be to seek the guidance of a third-party test security firm. As Kentucky moves into a new test-based accountability model, the state would likely benefit from an external audit of its current testing system.

Recommendation 3.5
The Kentucky Department of Education should continue plans to contract with an education test and assessment security company to obtain an audit of the state's current and proposed education testing system.

There is no single method that will eliminate inappropriate education testing practices. Challenges to making valid inferences are an inherent part of the education testing process. Nevertheless, proactive, strategic recognition of the potential challenges along with comprehensive and specific plans of action that are clearly communicated to all stakeholders may be the key to successfully managing said threats.

Chapter 1

Introduction

This study focuses on validation of data from the state's education tests after results are reported to the Kentucky Department of Education (KDE) and proposes an outlier methodology for identifying extremely high and low score gains or losses.

In December 2010, the Education Assessment and Accountability Review Subcommittee directed the Office of Education Accountability (OEA) to examine Kentucky's state testing data validation processes for the state's summative, state-mandated assessments required to gauge student mastery of core content knowledge. This study focuses on test and assessment data validation processes that occur after results for each testing period are reported to the Kentucky Department of Education (KDE) and proposes an outlier analyses methodology to identify potential testing irregularities. Kentucky does not have a system in place to review education test and assessment data for the presence of anomalous test data or data patterns that may indicate inappropriate testing practices.

Description of this Study

This study, with guidance and input from the National Technical Advisory Panel on Assessment and Accountability (NTAPAA), provides

- a description of Kentucky's data validation processes at the development, implementation, and post-testing levels for the Kentucky Core Content Test (KCCT) and ACT's Educational Planning and Assessment System (EPAS);
- an overview of Kentucky's strengths and weaknesses as defined by best practices;
- a review of literature for secure testing and specific methodologies for detection and follow-up on detected outliers including item-level and forensic tools;[1]
- a model or models for future data validation efforts based on outlier analysis methodologies, including possible fiscal costs;
- results based on analyses of school year 2010 KCCT and school years 2010 and 2011 EPAS test results; and
- recommendations related to best practices and missing pieces such as assignment of proctor identifiers and professional development recommendations.

[1] Forensic tools include a range of scientifically derived evidence to be used as part of a larger investigation. Common tools are fingerprint and handwriting analysis.

Organization of the Report

The remainder of this chapter provides an overview of Kentucky's current data validation processes. Chapter 2 includes a review of literature and provides a context for secure testing and specific methodologies for detection and follow-up on detected outliers, including item-level and forensic tools. Chapter 3 provides a model for future data validation efforts based on outlier analysis methodologies. In addition, Chapter 3 provides results based on school year 2010 KCCT and school years 2010 and 2011 EPAS analyses.

Contextual Overview

An education test or assessment must be reliable and valid to be of use to educators and policy makers.

A test or assessment designed to measure student learning is useful only if it is reliable, valid, and actionable. The test or assessment must yield consistent results, measure what it was designed to measure, and provide information in an efficient enough fashion that students, educators, administrators, and other interested parties may use the results to adjust learning, teaching, and administrative strategies. Reliability and validity are statistical properties that can be measured in many ways. Each comes with its own statistical computation. Reliability refers to consistency in assessment results, while validity is the degree to which an instrument measures what it was intended to measure (Sternberg).

This report identifies potential threats to making valid inferences and provides a systemic approach for managing those threats.

The purpose of this report is to identify potential threats to making valid inferences based on education assessment and test results. The report also provides a systematic approach within which Kentucky's eduction test and assessment results may be better understood and used to increase student learning and to improve teaching.

Threats to making valid inferences based on education tests and measures exist at every step of the testing process. This report deals primarily with potential unethical testing practices.

While this report focuses on making valid inferences based on education test and assessment data—specifically the impact that accountability-based pressure may have on test results in the form of blatantly or potentially unethical testing practices—it is important to understand that there are threats to reliability and validity at every step of education testing. Improprieties can occur in the creation and adoption of standards; when score reports are issued; and at the classroom level where some of the most difficult threats to monitor, detect, and remedy are found. These threats are discussed only briefly in this report because they are of sufficient complexity and importance to warrant their own study.

This report focuses on instances of unethical testing practice that may or may not be definitive.

This report focuses on unethical testing practices, both definitive and nondefinitive.

Examples of definitive activities include blatant and strategic efforts to increase test scores through methods such as coordinating student answer sheet changes. Nondefinitive activities may include practices such as preparing students for test-specific material via tips and tricks.

Kentucky's State Testing Processes

Education test and assessment construction is a complex and lengthy process beginning with the definition or identification of what is to be measured. Most often, education instruments aim to measure student performance at a variety of hierarchical levels ranging from general recall to students' evaluation of testing content.

Three instrument clusters make up the state's testing/assessment cadre: the Kentucky Core Content Test (KCCT), ACT's Educational Planning and Assessment System (EPAS), and the Iowa Test of Basic Skills (ITBS).

In Kentucky, three instruments make up the education testing/assessment cadre: the KCCT, EPAS, and the Iowa Test of Basic Skills (ITBS). In school year 2009, KDE oversaw the distribution, completion, return of, and report on approximately 1.5 million test instruments, including each KCCT subject test, each test in the EPAS system, and the ITBS.

School year 2012 will bring a new set of tests and assessments, including ACT's EPAS, end-of-course exams, and a hybrid criterion- and norm-referenced exam based on the new common core standards.

In addition, Kentucky plans to implement a set of assessments to comply with Senate Bill 1 2009 requirements. School year 2012 will see a combined norm- and criterion-referenced test that will be a hybrid of the current KCCT with a norm-referenced component similar to the ITBS, and the addition of end-of-course exams at the high school level. A norm-referenced test measures how a student performs compared to all other test takers, and a criterion-referenced test measures student performance based on mastery of specific content (Sternberg). Each of these instruments will play a primary role in the state's new education accountability model aimed at student learning and that the hybrid instrument will be based on the new common core standards, which are said to be more rigorous than previous math and reading standards. Table 1.1 compares Kentucky's current (interim) test and assessment structure and the state's proposed assessment system that will align with requirements of SB 1 from the 2009 Regular Session.

Table 1.1
Kentucky Education Assessment Systems 2009-2011 Compared to 2011-2012

Assessment	Interim 2009-2011	Proposed 2011-2012
KCCT: Reading	Yes	No
KCCT: Mathematics	Yes	No
KCCT: Science	Yes	No
KCCT: Social studies	Yes	No
KCCT: Writing on demand	Yes	No
Combined norm- and criterion-referenced test: reading	No	Yes
Combined norm- and criterion-referenced test: mathematics	No	Yes
Combined norm- and criterion-referenced test: science	No	Yes
Combined norm- and criterion-referenced test: social studies	No	Yes
Combined norm- and criterion-referenced test: writing on demand	No	Yes
Writing editing and mechanics	No	Yes
Writing program (including instructional portfolio)	No	Yes
Arts/humanities	No	Yes
Practical living/career studies	No	Yes
Norm-referenced test (Iowa Test of Basic Skills)	Yes	No
ACT: EXPLORE	Yes	Yes
ACT: PLAN	Yes	Yes
ACT	Yes	Yes
Alternate assessments	Yes	Yes
ACCESS* for English language learners	Yes	Yes
End-of-course exams	No	Yes

*ACCESS refers to Assessing Comprehension and Communication in English State-to-State.
Source: Compilation of data from the Kentucky Department of Education.

Each instrument listed has its own set of administrative rules, guidelines, and proctor training modules.

Each of the assessment instruments listed in Table 1.1 has its own set of administrative rules, guidelines, and proctor training modules. Though KDE makes every effort to standardize process-related and logistic components of the assessment system, different tests and assessments may require different return, report, or customer service processes. KDE, with assistance from each test vendor, provides training to district assessment coordinators, building assessment coordinators, test proctors, and any other staff member who may be involved in test administration. Basic EPAS and KCCT parameters are described in Appendix A.

Threats to Making Valid Inferences

Threats to validity may occur at the development, administration, and report stages.

Cognitive tests such as the KCCT and those in the EPAS system cannot include infinite items, measurements, and scorers, which results in a variety of potential threats to making valid inferences. These threats may occur at the time the test is developed, when the test is given, or after the test is taken. Threats include misaligned or diffused instrument design (as related to standards, curriculum, and instruction), an inadequate sampling of items per content type, inappropriate weighting of item types, and test taker/proctor incompetency or dishonesty (Wert).

Most education tests are relatively predictable in terms of content and performance expectations.

An additional threat to an instrument's validity is the reality that most tests are relatively predictable (Koretz. *Validation*). A selected response or multiple-choice test, depending on the number of potential answers options offered, provides even unprepared test takers with a fairly good chance of selecting a correct answer. For example, if a multiple-choice question has four potential answers, there is a 25 percent chance of answering correctly without reading the question. Further, if a test taker is able to eliminate one answer choice, the odds of selecting a correct answer with no or limited content knowledge improve accordingly.

Test score increases have been found to not be generalizable to other tests or assessments designed to measure similar content.

Studies based on Kentucky's education tests and assessments have indicated that increases in scores over time were not generalizable to other tests or assessments with similar content. The studies also indicated that pressure to increase scores led teachers to view score increases as their goal rather than student learning, causing teachers to focus only on "tested" content (Koretz. *Evidence*).

A host of things in the test conception and construction phase, during testing, and after testing could interfere with making valid inferences. Table 1.2 offers a summary of Kentucky's education assessment cycle and associated action items at each phase in the cycle. It also provides a brief description of how Kentucky addresses validity at each phase. Each step listed requires an enormous amount of work, and many of the steps are complex, time consuming, and costly. In short, each step is worthy of its own descriptive chapter or report.

Table 1.2
Phases in Kentucky's Education Testing and Assessment Cycle

Phase 1: Address development-related threats to instrumentation and eventually making valid inferences.

- Standards are developed by trained professionals.
- KDE partners with high-quality education testing vendors.
- Tests and assessments are designed to align with standards, and items are field tested to establish reliability and validity.
- Alignment and validity studies are ongoing as required by KRS 158.6453(17).*

Phase 2: Address administration-related threats to making valid inferences.

- District and building assessment coordinators and proctors receive training on how to administer tests and assessments and other processes related to test administration, submission, and reporting.
- All test administrators must be trained on the assessment code of ethics and sign it.
- Education professionals, students, parents, and other stakeholders may report suspected testing code violations to KDE's testing allegation coordinator.

Phase 3: Address report-related threats to making valid inferences.

- Draft score reports are sent to KDE and schools in an effort to correct scoring and submission errors.
- Final reports are issued to schools and the public.

*Appendix B explains the Kentucky Department of Education's Biennial Plan for Validation Studies.
Source: OEA staff compilation of Kentucky Department of Education information.

Regulatory Processes

To increase reliability and validity, KDE promulgated 703 KAR 5:080, which explicitly states the rationale for appropriate testing practices. KDE oversees annual reliability and validity studies, works with well-qualified test vendors, makes vendor support and expertise available to district and building staff, builds into vendor contracts periods for report review and revision, gets guidance from the National Technical Advisory Panel on Assessment and Accountability (NTAPAA), and is part of national test consortia.

To put into place a student assessment system that is reliable and valid as prescribed in SB 1 from 2009, 703 KAR 5:080 states the rationale for appropriate testing practice, defines appropriate assessment practices, describes how violations of the administration code may be handled, explains how requests for test material review may be handled, and describes the proper reporting of nonacademic indicators. Appendix B further explains the Kentucky Department of Education's Biennial Plan for Validation Studies. To bolster reliability and validity, KDE performs the following tasks:

- Contracts with the Human Resources Research Organization to complete annual reliability and validity studies.
- Works with well-qualified test vendors to develop reliable and valid instruments and professional development activities (e.g. Kentucky Content Leadership Networks, district assessment coordinator training, proctor training) designed to enhance reliability and validity for each respective test type.
- Makes available test vendor customer support specialists who are available to address test administration questions that may arise during testing periods.

- Builds into each vendor contract a period for score review and revision.
- Contracts with an external psychometric consultant.
- Gets guidance from NTAPAA.
- Participates in both national common core assessment consortia.[2]

KDE provides professional development and training for district and building testing staff.

KDE also provides professional development and training requirements for district and building assessment coordinators and test proctors to control for external threats to validity. KDE's system allows for field-level collection and reporting of testing allegations ranging from inappropriate assistance and intervention by proctoring staff to what to do in the case of missing testing materials.

KDE has a system in place for reporting, investigating, and managing testing allegations.

Specifically, KDE receives formal complaints or testing misconduct allegations, and those complaints and allegations are handled by KDE's testing allegation coordinator in the Office of Guiding Support Services. The testing allegation coordinator, with the assistance of legal counsel and support staff, follows up on complaints and allegations. Some complaints and allegations are easily handled because they arose as the result of a miscommunication or other similar cause. Other complaints and allegations require a full investigation that includes site visits, stakeholder interviews, and document reviews.

Consequences for committing a testing offense range from a reminder of the rules to a referral to the Education Professional Standards Board.

In instances of confirmed inappropriate action, consequences range from a reminder of the rules and a related professional development refresher course to reduction or elimination of scores and a possible report to Kentucky's Education Professional Standards Board (EPSB) if the violation was determined to be intentional. EPSB also opens cases when district officials report that they believe cheating has occurred in their districts. Since 2000, EPSB opened 91 testing violation-based cases and closed 67 of them; 35 closed cases resulted in dismissal. Thirty-two of the remaining cases were actionable, with 9 of those cases resulting in suspension and 1 resulting in license revocation. Reported incidents, responses, and potential consequences are reviewed each month by the Testing Board of Review.

[2] The SMARTER Balanced Assessment Consortium and the Partnership for Assessment of Readiness of College and Careers were formed to develop common core assessments. The work of those consortia will seek to ensure that future assessments designed to measure students' performance given the new common core standards are reliable and valid.

Over the past 13 years, 2,407 testing allegations were reported, and 461 of those cases resulted in violations that resulted in score reductions.

Table 1.3 presents reported testing allegation data from school years 1996 through 2009. In addition, it provides a summary of examples of inappropriate testing practices. Over the past 13 years, 2,407 testing allegations were reported, and 461 of those cases resulted in score reductions; this is a relatively small number of score reductions considering the number of instruments, students, and locations involved.

Table 1.3
Reported Testing Allegations From School Years 1996 to 2009

Type of Allegation	Number of Allegations
Inappropriate assistance/intervention by staff. Allegations include inappropriate comments evaluating student work, any actions or comments that assist students in answering questions, and providing materials that are not part of the state-required assessment.	356
Test security breaches by staff. Allegations include leaving test materials unattended, allowing students to leave the room with test materials or to have test materials in their possession without supervision, and allowing others who have not received administration code training or permission to have access to or view secured test items.	314
Test security breaches by student. Allegations include looking ahead to other parts of the test and taking test booklets from the testing area without permission or supervision.	199
Staff testing out of order. Allegation is that the test administrator or proctor did not follow the order of test indicated in the testing manual and test schedule.	423
Student testing out of order. Allegation is that the student worked ahead of the test schedule.	180
Special education staff breach in accommodations protocol. Allegation is that staff provided inappropriate accommodations or did not provide accommodations listed in the individual education plan while administering the test.	290
Staff breach of portfolio protocol. Allegations include inappropriate assistance.	123
Inappropriate action on the part of the student or students. Allegations include sharing answers during the test, using hidden notes to assist in answering questions, and using dictionaries or calculators when instructed not to do so.	164
Missing materials. Allegations involve test materials that cannot be accounted for after testing.	37
Other general staff allegations. Allegations include providing feedback to students on good-faith efforts prior to completion of test, allowing students to change or complete answers after the test session has ended, and not providing dictionaries or calculators when indicated.	321
Total	2,407
Total Resulting in Score Reductions	461

Source: Data from Commonwealth. Department. *Reported.*

Kentucky does not have a system to routinely analyze test results for outliers.

Kentucky does not have a system that routinely analyzes test results for outliers and anomalies that may require further investigation. Nor is there an explicit process to investigate when outliers are identified by analytical efforts or formal complaints.

Chapter 2

Review of Literature

Introduction

Internationally, testing and assessment integrity issues are becoming a primary concern of education professionals. Those international concerns and issues are prevalent in the US as well. There is a national trend toward increasing enforcement efforts, including development and implementation of policies that address what action to take when inappropriate testing practices are suspected, along with appropriate action steps following a confirmation of unethical activity (Sorensen. *2008*).

Test security concerns include content protection, detection of test and assessment proxies, and coding errors. The Council of Chief State School Officers (CCSSO) and the Association of Test Publishers (ATP) recommend that states have in place a proactive and comprehensive system for addressing test security breaches.

Challenges related to test and assessment integrity include content protection, detection and management of test and assessment proxies, and postsubmission coding errors. According to the Council of Chief State School Officers and the Association of Test Publishers, states should have in place a proactive and comprehensive testing and assessment plan that includes strategies for overcoming and avoiding test and assessment security breaches. Likewise, the US Secretary of Education issued a brief to chief state school officers urging chief state school officers to review their state's processes for protecting the validity of accountability data through review of test security systems (Duncan). However, a single proactive strategy for detecting and dealing with test and assessment integrity breaches has not been developed.

Making Valid Inferences

A test has validity if it measures what it was designed to measure. There are many statistically valid education tests and assessments on the market.

According to education measurement researchers, "a test has *validity* if it measures what it purports to measure" (Allen 95). There are many education tests and assessments on the market that produce consistent results and that measure what they are supposed to measure. Their reliability and validity statistics are acceptable, and most provide reports that assist learners and educators in their efforts to understand the results in a way that facilitates learning and teaching.

Education tests and assessments are only as valid as the inferences made using their scores.

Reliability and validity, however, are more than statistical computations. Education tests and assessments are only as valid as the inferences made using their scores. The qualitative reliability and validity of the inferences made about education test and assessment results are as important as statistical reliability and validity. Making a valid inference based on test data is crucial for the accountability process to be useful.

Testing Pressures

Pressures to inflate test scores at all costs pose a serious threat to making valid inferences as those pressures lead to artificially inflated scores.

The increasing push to raise students' assessment and test scores creates a system in which accountability plays a primary role. Granting rewards for good results and sanctions for bad test results may cause educators to push their districts, schools, and classrooms to a higher level, with unintended consequences. One researcher warned:

> Scores on tests used for accountability have become inflated, badly overstating real gains in student performance. Some of the reported gains are entirely illusory, and others are real but grossly exaggerated. The seriousness of this problem is hard to overstate. When scores are inflated, many of the most important conclusions people base on them will be wrong, and students—and sometimes teachers—will suffer as a result (Koretz. *Measuring* 235).

Extreme pressure to increase test scores in Georgia, Pennsylvania, Indiana, Maryland, and Washington, DC, have led to multiple allegations of corruption. An ACT investigation of Kentucky scores produced findings that indicated the commonwealth is not immune to these national trends.

In addition to general score inflation, an intense focus on cognitive measures, such as student learning, can corrupt the specific measure. This seemed to be the case in Georgia, where more than 50 Atlanta schools were flagged for cheating in 2009. Schools in the District of Columbia, Indiana, Maryland, and Pennsylvania also have been identified for cheating. Likewise, an ACT investigation of Kentucky PLAN and ACT scores from school years 2009 and 2010 led to cancellation of tests scores in one district because testing staff did not adhere to ethical testing practices (Innes).

While the common assumption is that changes in test scores equal changes in student learning, that is sometimes not the case. This dynamic is an important one to understand as Kentucky moves from one testing system to a more high-stakes environment that imposes more rigorous education standards.

A high-stakes testing environment increases the likelihood of unethical behavior in the form of inappropriate testing practices.

A high-stakes testing environment increases the temptation of educators to focus on test preparation over student learning. Educators and students are more likely to exhibit inappropriate testing practices when faced with the pressure to increase test and assessment scores. Given these pressures, it is important to recognize and understand the many ways education professionals may undermine making valid inferences based on education test and assessment scores.

The use of test scores as a major determining factor in funding, resources, and favorable public opinion can be an incentive for dishonest and unfair testing practices.

The use of test scores as a major determining factor in funding, resources, and favorable public opinion can be an incentive for dishonest and unfair testing practices. Some administrators and teachers may feel threatened by what they perceive to be unreachable test performance goals. Therefore, they may be motivated to gain public approval by exhibiting large gains in test and assessment results. These pressures can influence district leaders, school administrators, and classroom teachers to use inappropriate testing strategies. These strategies include supplying students with answer keys, allowing more or augmented time to complete tests, and recoding student answer sheets after the tests are turned in.

Testing experts describe two ways to corrupt a measurement. To artificially change a measure, the sample used to obtain the measure can be distorted (Koretz. *Measuring* 240). This blatant cheating includes changing answer sheets after the fact and extreme coaching that focuses only on the material thought or known to be included on a test.

Results of a national test security survey indicated that education professionals expressed concern about inappropriate pressure to perform on education tests and about coaching in preparation for a test.

In a national survey considering test security, respondents were asked to rate the test security threat caused by education administrators putting inappropriate pressure on teachers to increase test scores. Results on a scale of 1 to 5, with 5 being the highest, indicated that respondents rated inappropriate pressure as a 3.6, second only behind concerns over lost or stolen test booklets. That same survey yielded an average rating of 3.2 for teacher coaching prior to tests and 2.9 for teaching/coaching during the test. The possibility of teachers or administrators changing student responses after testing earned an average rating of 2.7 (Sorensen. *2008*).

Allegations of inappropriate education testing behavior on a large scale are not new in Kentucky.

While most administrators or educators likely do not participate in unethical or inappropriate testing practices, those who do tend to fall into one or more of the broad survey categories listed above. Kentucky has been at the forefront of statewide student assessment and accountability since 1990. There have been examples of inappropriate testing practices. For example, a 1996 investigation by KDE of Bell County School District’s scores uncovered approximately 80 testing violations committed by a high school principal who encouraged teachers to increase test scores by whatever means necessary (Nichols). Further, a 2010 investigation of education test scores in the Perry County School District concluded that student answer sheets had been altered by educators or proctors who provided answers during the test (Ritchie).

The pressure to increase test scores trickles down from federal, to state, to district, to school, and finally to classroom mandates. Students are the final inheritors of the extreme pressure to perform. While extreme pressure from high-stakes testing conditions may be a relatively new phenomenon within and among education professionals, preventing and detecting student cheating has always been a challenge. Cheating is generally defined as any clandestine assistance while taking an education test or assessment or gaining unpermitted access to testing material in advance of the test or assessment (Sorensen. *50*).

Best Practices for Statewide Assessment Programs

CCSSO and ATP recommend that all education testing programs include external audits, a focus on test administration, content protection, strong test user agreements, plans and systems to vigorously pursue rule violators and for forensic analyses, and comprehensive plans of action.

To recognize the potential threats posed by statewide education testing and assessment programs, the Council of Chief State School Officers (CCSSO) and the Association of Test Publishers (ATP) released their *Operational Best Practices for Statewide, Large-Scale Assessment Programs* in 2010. Among their 10 key best practice guidelines are

- outside audits of test security practices;
- clear focus on test administration parameters;
- never-ending protection of content;
- employment of strong test user agreements;
- development of plans and systems to vigorously pursue rule violators;
- plans and systems for forensic analyses of test- and assessment-related data; and
- development of comprehensive security breach plans of action.

The US Secretary of Education echoed the CCSSO and ATP guidelines in a letter to chief state school officers (Duncan). OEA staff found that the common themes in each recommendation offered by the CCSSO and ATP documents are forethought and strategy. Test and assessment integrity issues will only increase with the importance placed on performance results. States need to develop comprehensive and clear policies regarding how they plan to manage test and assessment integrity issues at each phase and level. While complete elimination of all invalid score shifts is not possible, a strategic approach based on transparency and clear expectations and consequences is crucial, especially in light of current testing integrity trends.

KDE has in place many of the components recommended by the CCSSO and ATP.

KDE has in place many of the components recommended by the CCSSO and ATP. Specifically, KDE has a clear focus on test administration parameters, and it communicates this focus through training and professional development activities. Likewise, KDE and its test vendors have high-level security measures to protect test content. Test users are required to receive training on ethical testing practices and to sign user agreements to document receiving the training and understanding appropriate testing practices.

Kentucky's system lacks outside audits of test security practices, systems to vigorously pursue rule violators, systems for forensic analyses, and overall action plans for security breaches.

Lacking in Kentucky are outside audits of test security practices, plans and systems designed to vigorously pursue rule violators, plans and systems for forensic analyses, and overall plans for managing security breaches. While most analysts agree with the CCSSO and ATP recommendations, there is some concern about the return on investment of enhanced security measures.

An external education test security consultant can review a state's testing system, collect evidence, and provide follow-up action.

Frequently, the security measures implemented by a state require the expertise of external consultants. Test security companies generally approach the security needs of a testing system with a three-step process.

First, test security vendors conduct a comprehensive review of the state's education testing and assessment system including all test and assessment policies, instruments, and systems.

The second step is the collection of forensic evidence related to the testing period in question. Typical forensic investigations gather evidence of cheating or copying, using advanced knowledge to answer questions or answering in unexpected ways, and stealing tests (Caveon. *Sample*).

The third phase is a follow-up investigation in which any outstanding investigation-related issues are addressed. The degree to which a test security company undertakes some or all of the investigative components listed above is case dependent. Costs for each component vary.

Fiscal Considerations

KDE paid approximately $17 million to test vendors and test and assessment consultants in fiscal year 2008.

A recent audit of KDE's oversight of state assessment contracts found that the department paid nearly $17 million to test vendors and test and assessment consultants in fiscal year 2008

(Commonwealth. Auditor). The total expenditures for the department's learning and results services for assessment and accountability were $24.9 million for FY 2010, and $33.6 million and $35.3 million were requested for FY 2011 and FY 2012, respectively (Commonwealth. Office).

Typical test security consultant expenses range from a low-end basic review for less than $100,000 to a high-end security package of nearly $450,000.

There are additional costs associated with carrying out the enhanced security components prescribed by the CCSSO and ATP recommendations. One of the nation's leading test security consulting firms provided OEA with a general cost estimate based on the parameters of a basic detection, investigation, and report package. While the needs of Kentucky may vary from the services provided under this estimate, the estimate provides some general guidelines for consideration. Typical contracts range from a low-end basic review for less than $100,000 to a high-end security package of nearly $450,000 (Caveon. *Cost*). These costs are based on a one-time audit, and resulting costs are highly variable and dependent on specific audit findings. Annual costs would increase in tandem with the complexity and scale of specific audit recommendations. Kentucky may be able to build similar analyses and processes into current and upcoming vendor contracts at a reduced cost.

Many test and assessment integrity issues would be easy to remedy and may be resolved by clear communication of expectations, appeals to ethics, and specifying the process of detecting cheaters and the consequences.

Security experts agree that many test and assessment integrity issues would be easy to remedy. One test security firm executive reported that many challenges may be overcome with clear communication about what is expected of each key player in the testing process—from item writer to parent. In the training process, appeals to ethics and honesty go a long way to dissuade potential cheaters. Likewise, specifying the process involved with detecting cheaters and the consequences associated with cheating is essential to discouraging it. While these efforts are not as sophisticated as complete third-party security packages, they are more feasible than adding to testing costs. The fact that schools and districts that exhibit statistically improbable gains or losses will be reviewed is an inexpensive and highly effective factor for dissuading unethical behavior, according to test security experts (Fremer).

US Responses to Test Security Survey

States are concerned with test security, and test security efforts are on the rise.

In 2008, a major test security company surveyed all 50 states to obtain information about how each state perceived, detected, and handled education test security issues. The survey results indicated that

- states were very concerned with security as it related to their state assessment programs,

- test security enforcement efforts were on the rise,
- many states had plans in place to address inappropriate testing practices and associated complaints, and
- many states had action plans upon confirmation of inappropriate practice (Sorensen. *2008*).

To obtain comparable information and to determine the extent to which other states are implementing security measures around their education test and assessment data, OEA staff surveyed test and assessment coordinators across the country, receiving responses from 33 of 51 potential respondents.[1]

The Office of Education Accountability surveyed US state testing leaders and learned that approximately 58 percent of respondents identify potential threats to making valid inferences based on test data into vendor contracts and that 75 percent of respondents have systems in place to address allegations of widespread testing allegations.

Respondents held positions similar to that of KDE's associate commissioner in the Office of Assessment and Accountability. The survey asked participants to indicate approaches their state uses to identify potential threats to making valid inferences based on education test and assessment results. In addition, the survey asked if states have a system in place should widespread testing allegations occur. Most (57.6 percent) respondents indicated that they attempted to identify potential threats to making valid inferences based on education test and assessment results through building into test vendor contracts such analyses and routinely conducting in-house analyses. Table 2.1 presents specific results for the first question. Most respondents (75 percent) indicated that their state had a system or systems in place to address allegations of widespread testing allegations.

Table 2.1
Survey Responses for Survey Question 1

Please indicate which of the following approaches your state uses to identify potential threats to valid inference making based on education test and assessment results? Select all that apply.	
We do not conduct analyses aimed at identifying potential threats to valid inference making	9.1%
We routinely conduct in-house analyses.	36.4%
We build these analyses into test vendor contracts.	57.6%
We contract with an external consultant.	9.1%
We contract with an external test security company.	9.1%
Other	21.2%

Source: Results from OEA State Test Validation Survey.

[1] One respondent for each of the remaining 49 states plus the District of Columbia and an extra potential respondent from North Carolina, which listed co-directors.

Common open-ended response themes from the OEA survey included mention of erasure analyses and regular education data review with an eye toward identifying irregularities.

The survey also solicited open-ended responses aimed at providing more detail about each state's selected responses. Common themes in the open-ended responses included mention of erasure analyses and regular review of education data with the express purpose of detecting irregularities; these reviews are conducted either in-house by state assessment staff or by the testing contractor. Responses providing more detail about what happens when testing allegations occur were highly variable with no single investigative entity being mentioned most often. Some states handle testing allegations through their education departments while others use agencies similar to EPSB and OEA.

Kentucky has no comprehensive system that defines key players and their roles, should a cheating allegation occur.

Kentucky contracts with an external psychometric consultant[2] for some validity-related review; a certain degree of review is built into existing test vendor contracts, though that review is limited and not necessarily a routine part of the testing and report process. While many of the essential components are in place, there is no comprehensive system that identifies key players and defines their roles should a cheating allegation occur.

Test Security Procedure

When scores are identified as being notably larger or smaller than those in the majority of the distribution, a more in-depth or detailed examination is warranted.

In addition to assessment and ethics training, which are components of the Kentucky testing system, states have implemented other measures to ensure score validity through a comprehensive review of test results. When scores are identified as being notably larger or smaller than those present in the majority of the distribution, a more in-depth or detailed examination of those instances is warranted. A large score growth or decline indicates neither that any unethical testing practices have occurred nor that a given school has done anything groundbreaking to change test scores. Identification means only that the scores in question require a closer look. To determine whether improprieties occurred, analyses in greater depth are necessary.

Inquiries that follow identification of extremely large score jumps or drops range from simple state efforts to more complex vendor-level analyses.

Inquiries that follow identification of extremely large score increases or decreases range from simple state efforts to more complex vendor-level analyses. The next section reports descriptions of approaches used by vendors and states to identify and investigate schools with anomalous or outlier scores.

[2] A psychometric consultant guides work related to the theories and methodologies involved with measuring psychological constructs such as content knowledge.

Initial State Efforts

Upon identification of outlier scores, an efficient option is to perform more descriptive analyses.

Upon identification of outlier scores, an efficient option for moving forward is to perform more descriptive analyses, such as computing means and comparing different populations within the identified school and, when available, examining student-level answer patterns and performance trends on different tests and assessments covering similar content. For example, a student within a certain ACT mathematics performance range is more likely to be in a similar performance range on other mathematics tests and assessments. Questions such as the following could guide initial follow-up analyses:

- How did students in the test pool perform on different tests and assessments covering similar content?
- Are there notable or statistically significant differences between regular and special education students within the identified school?
- Is the school from a size classification or configuration that warrants extreme variability?
- Are there any extreme subtest fluctuations that may have contributed to the composite score differences?
- Have the school's population parameters changed drastically from one year to the next?
- Have there been extreme school level efforts aimed at specific score increases?
- Did any historical effects such as natural disasters or national events affect the school during the testing period?

After an external review of testing practices based on a widespread testing allegation, one Kentucky school's PLAN scores dropped 3.8 points, indicating a huge drop in performance after changes to testing administration in the school, including identifying the staff eligible to participate in the testing process.

When OEA staff examined some results for Kentucky, one school within the medium school size category was identified as being a lower-limit PLAN outlier. This means that of all the medium-sized schools that took the PLAN test during the most recent testing period, one school's scores dropped significantly. A review of the data used to make that determination revealed the composite PLAN score for that school dropped 3.8 points from one year to the next. Further review revealed that the school in question was part of an external test vendor investigation. After investigation, the vendor found that test scores from prior years for that school were the product of unethical testing practices on the part of the staff and faculty, and the most recent scores came from exams given under review and observation by KDE and ACT. As a result of the test vendor investigation, the school was required to make changes to testing administration in the school, including identifying the staff eligible to participate in the testing process. In this case, the drastic drop likely reflects the investigation's impact.

Schools with outlier scores are not necessarily participating in unethical testing practices, but more in-depth analyses are appropriate.

The increased or decreased scores may be the result of less nefarious causes. For example, accommodated students may be raising or lowering scores significantly. A state-level investigation may determine that a particular school, during the testing years in question, received an external grant aimed at facilitating drastic but legitimate increases. Some schools are just doing the right things to increase student learning and are using education test and assessment data to improve; however, without review, it cannot be determined which changes in scores are valid and which are the result of improper instructional or testing practices.

A quick state-level analysis is not sufficient to answer questions about why a particular school's score rose or dropped significantly. In those cases, more-sophisticated efforts may be necessary.

Detailed State and Vendor Analysis

In most cases, a quick state-level analysis is not sufficient to answer questions about why a particular school's score rose or dropped significantly. In those cases, more-sophisticated efforts may be necessary. Among the most common investigative techniques are reviews of the actual mechanical processes used to scan in score sheets, stakeholder interviews, answer sheet analysis including answer pattern and forensic analyses, and controlled retesting. Generally, this level of analysis is performed by the test vendor or other outside contractor.

In the case of anomalous scores or score differences, a good first step is an examination of the mechanical processes by which test answer sheets were scanned.

The first step could be a request that the test vendor examine the actual mechanical processes by which the score sheets for the test in question were processed. For example, almost 140,000 Kentucky students take either the EXPLORE, PLAN, or ACT exam per year. Each student uses a physical answer sheet that must be fed into a scanning machine, processed, added to an aggregate report with multiple data points, and filed. Each form can result in a process error. Mechanical error includes multiple scans of a single score sheet, miscoded score sheets, and spoiled score sheet batches.

Stakeholder interviews are a necessary step in gathering evidence in the case of testing allegations.

Another effective follow-up activity is stakeholder interviews with the building and district assessment coordinators and test proctors about the potential causes for the extreme scores. Such interviews can help eliminate the need for a formal, costly, full-blown investigation. Some potential pieces of evidence include proctor recall of any test-day anomalies in addition to allegation-relevant pieces of evidence. For example, if it is suspected that certain students shared information during a test, a seating chart from the day of testing may be helpful. In addition, student, parent, teacher, and administrator interviews can be conducted to provide allegation-related information.

Forensic analyses include examination of physical evidence, biometric evidence such as fingerprint analyses, and statistical evidence.

Another investigative technique used by test vendors and contractors is a forensic analysis. Forensic analyses of test results include physical evidence, biometric evidence fingerprint analyses, and statistical evidence. Forensic analysis of physical evidence includes examining actual test booklets for notes or similarities, seating charts, room configuration reports, proctor-specific answer patterns, calculation patterns, or word use patterns.

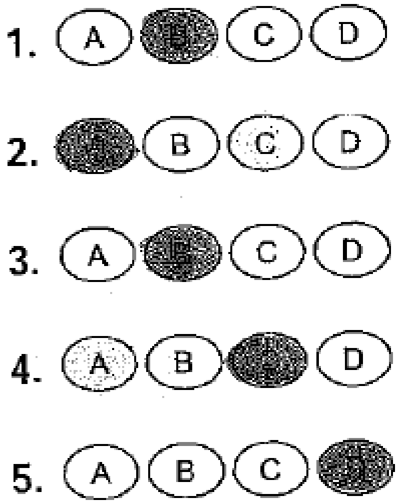
Item response patterns and erasures are relatively easy pieces of information to collect and examine.

Relatively easy pieces of physical evidence to collect and examine are item response patterns and erasures. These require reviewing how students responded to each item. Items on a test have individual sets of statistical properties. If all or most of the students miss a relatively easy item, or if high numbers of students get a relatively difficult item correct, then that information could be used to determine suspected testing irregularities. If an item that typically does a good job of differentiating between high and low performers yields no difference in a given classroom, that information is telling as well.

An erasure is an instance in which a student fills in an answer, erases that answer, and fills in a final answer.

An erasure is an instance in which a student fills in an answer, erases that answer, and fills in a final answer. Each time an answer is erased, it leaves a shaded trace, and those smudges are countable. Figure 2.A illustrates a typical erasure.

Figure 2.A
Samples of Erasure Smudges



Source: OEA staff synthesis with a Prentice-Hall Free Use Bubble Sheet.

A close look at items 2 and 4 in Figure 2.A reveals slight traces on answer options C and A, respectively. While there are general rules about acceptable number of erasures, each testing pool has its own expected erasure benchmarks.

One Washington, DC, school averaged 12.7 incorrect-to-correct erasures per test taker. The average incorrect-to-correct erasure rate for all other 7th-grade students in Washington was 1.

Crosby S. Noyes Education Campus in Washington, DC, offers an excellent example of how erasure analysis may be used to investigate suspected impropriety. Over the past few years, Noyes has been at the center of controversy around drastic performance increases. While the school was being praised for high growth, it was also being flagged by the testing vendor for having extraordinarily high wrong-to-right erasures. One Noyes classroom averaged 12.7 incorrect-to-correct erasures per test taker. The average incorrect-to-correct erasure rate for all other 7th-grade students in Washington was 1.

Similar instances of high incorrect-to-correct erasure rates have occurred in Georgia over the past 2 years. While Washington authorities and administrators chose to conduct discreet, private investigations to determine which, if any, allegations of wrongdoing were valid, Georgia is conducting a full-blown criminal investigation that will likely lead to prosecutions (Galloway).

While high incorrect-to-correct erasure counts warrant further investigation, those high counts alone are not necessarily evidence of any unethical activity.

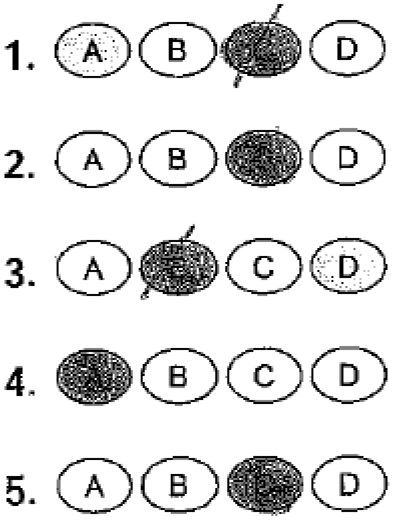
While high incorrect-to-correct erasure counts warrant further investigation, those high counts alone are not necessarily evidence of any unethical activity. High erasure counts are improbable but not impossible, and some educational best practices may increase incorrect-to-correct erasure counts. For example, reminding students to double-check their work or to take time to review their answers may yield higher counts of incorrect-to-correct answers. Additionally, students may make legitimate, wide-scale mistakes, such as mistakenly filling in answers on the wrong line.

Erasure counts are, however, useful when combined with other pieces of information, such as answer patterns and other answer sheet marks.

Erasure counts are, however, useful when combined with other pieces of information, such as answer patterns and other answer sheet marks. Other considerations include additional answer sheet evidence, such as multiple authors, and evidence indicating that answers may have been changed after students submitted their work.

Figure 2.B offers an example of an erasure combined with another answer sheet mark. In this case, the mark indicated a correct answer. Items 1 and 3 have erasure marks and lines through the final student responses.

Figure 2.B
Samples of Erasure Smudges Combined With Other Marks



Source: OEA staff synthesis with a Prentice-Hall Free Use Bubble Sheet.

A high number of erasures combined with other marks relating to chosen or abandoned evidence may necessitate additional analyses, including fingerprint or palm print analysis.

If an answer sheet has a high number of erasures and includes other marks, especially marks relating to chosen or abandoned answers, a follow-up analysis may include a fingerprint or palm print analysis.

An answer sheet may have multiple prints on certain areas, while other areas should have only test-taker prints. For example, if a proctor has to replace an answer sheet, it would not be uncommon for an additional fingerprint to be along the sides or edges of the document. It would be much less likely that the proctor's fingerprints would be in locations and positions similar to those of the test taker.

Handwriting analyses assist in determining whether more than one person placed answers on an answer sheet.

Finally, forensic test analysis may include a handwriting analysis. Pencil pressure, mark density, writing instrument, and similar properties help to determine whether multiple writers placed answers on the answer sheets. Again, there are situations in which anomalies in any one of the conditions discussed above may legitimately exist, and it is only when multiple pieces of evidence align that the likelihood of an anomaly occurring by pure chance is reduced or eliminated.

Response to Investigative Findings

A supported claim of compromised test scores requires a state to negate test scores, take disciplinary actions, and mandate districts ensure test administration compliance during the next test.

A state must determine the steps to take in response to investigative findings. When the preponderance of evidence supports a claim that test scores have been compromised, disciplinary actions may include negating test scores, proceeding in disciplinary actions against responsible parties, and mandating district actions to ensure test administration compliance during the next test. The state could also determine that a strict third-party monitor of the next assessment or a formal retesting of the population is required.

If the scores for an entire test are found to be invalid, retesting under appropriate testing conditions may be the best way to determine the magnitude of inappropriate test practices on student scores.

The nature of the retesting conditions depends on the cause of invalid score increases.

While the varied ways retesting could occur are beyond the scope of this report, it is important to note that the nature of the retesting conditions is dependent on the cause of invalid score increases. For example, if test scores increased because students were provided with content-specific strategies aimed at a particular test type, then a retest would need to include equivalent content on a different test, as readministering the same test would likely produce the same invalid scores while introducing test familiarity bias.

Routine, sample-based testing under highly secure and monitored conditions is another possible retesting strategy.

Another viable strategy, related to retesting, is routine, sample-based testing under highly secure and monitored conditions. The state's current wide-scale testing plan requires a great deal of trust in and collaboration with local agency testing staff. While this is certainly cost efficient and not, by default, unethical or inappropriate, situations that warrant a closer look may benefit from a high-security test administration. Likewise, each year, some assessments could be randomly assigned to randomly selected schools, and those assessments would be managed and proctored by KDE or an appropriate third party. Results from such an administration would likely provide a more valid indication of content mastery free from many of the biases present when using known instruments that are proctored by familiar proctors in regular classrooms.

Not all investigations stemming from testing allegations yield actionable results.

Not all investigations stemming from testing allegations yield actionable results.

Chapter 3

Outlier Analyses

Introduction

Kentucky does not conduct a regular review of education test results aimed at identifying outliers-- scores that are beyond normal expectations.

One missing element of the current system, and potential gateway to identifying and addressing inappropriate testing practices, is a regular review of education test results aimed at identifying scores that are beyond normal expectations, or outliers. Such reviews are frequently conducted by the test vendor as part of the state assessment contract or by an outside contractor. Testing and assessment literature indicates that many data integrity challenges may be remedied with a transparent system that prescribes what is expected of key players in the testing process, what will be done to detect potential violations of testing regulations, and what will happen if potential violations are detected (Fremer). KDE has in place many of the components necessary to facilitate reliable and valid education test results, and processes are in place to make clear what is expected of key players along the way. Regular outlier analyses would augment the work it already has in place.

Regular outlier analyses are a relatively inexpensive and efficient way to identify schools that yield education test and assessment scores that were notably different from the scores of most other schools.

Regular outlier analyses are a relatively inexpensive and efficient way to identify schools that yield education test and assessment scores that were notably different from the scores of most other schools. A basic outlier analysis offers a simple, descriptive look at how scores were distributed within and between different schools. Such analysis also serves as a gatekeeper for more in-depth inferential, investigative, and forensic analyses such as those described in Chapter 2. Additionally, the education testing and assessment literature indicates that if stakeholders know that data will be routinely reviewed, potential threats to the integrity of the data will be reduced or eliminated (Fremer).

Overview

Kentucky does not regularly review test data to identify outliers or other test score anomalies. In order for such analysis to be useful, data review is conducted prior to public release of results. In Kentucky, no external entity has access to education test results with the express purpose of identifying irregularities until final data reports have been issued. The following section describes the

KCCT and EPAS instruments and outlines a potential outlier analysis methodology.

Outlier Identification Methodology

OEA conducted an outlier analysis on differences in scores from one year to another taking school size into account.

To compare the most accurate data, OEA staff collected data for each test. Schools in which no students tested or for which no scores were available were excluded from the research pool. For each school, score differences were calculated by subtracting the previous year's score from the current year's score. To obtain score differences for KCCT performance, OEA staff used KCCT raw scores. Appendix D explains OEA's outlier identification methodology. The schools were ranked into four categories by school size. Size groups were determined by establishing each school's percentile rank based on the number of students tested. Schools defined as tiny fell at or below the 25th percentile; small schools fell between the 26th and 50th percentiles; medium schools were those between the 51st and 75th percentiles; and large schools were at or above the 76th percentile based on number of students tested.

OEA staff determined expected score distributions for each school size category. From those calculations, staff could determine what scores would be considered extreme, either by significant score increases or by significant score decreases. These upper- and lower-outlier limits would be subject to additional review.

The concern with upper-limit outliers is that some action at the school has influenced student performance from one test year to the next.

Upper-limit Outliers. The primary concern with upper-limit outliers is that some action at the school has influenced student performance outcomes from one testing period to the next. It is statistically improbable to increase scores so much in such a short period of time. However, it is possible that a genuine increase occurred due to a radical shift in the testing cohort, a data entry error, or potentially the introduction of a new learning program.

Potential reasons for improbable drops in scores from one year to the next could be data entry error or a major change in the student population.

Lower-limit Outliers. Statistically improbable test score decreases can also occur. Potential reasons for such improbable drops could be data entry error or a major change in the student population. For example, the relocation of a large employer away from a school district could have a negative impact on school performance. An improbable performance decrease could also be evidence of past improper testing practices.

Kentucky Core Content Test

The Kentucky Core Content Test (KCCT) is designed to assess student mastery core content in addition to higher-order thinking and communication skills and is made up of open-response and multiple-choice questions in reading, mathematics, science, social studies, and writing.

The Kentucky Core Content Test is produced, administered, scored, and reported on by Measured Progress, a large, nonprofit testing company that serves 17 other states. The KCCT is designed to assess student mastery of Kentucky's core content in addition to higher-order thinking and communication skills. The test is made up of open-response and multiple-choice questions in reading, mathematics, science, social studies, and writing. Overall, in school years 2009, 2010, and 2011, KCCT testing required 7 to 10 hours of completion time for 91 to 122 questions, depending on the grade level. Specific testing time and grade-level content information is outlined in Appendix C.

Originally the chief mechanism for generating academic indices for schools, the KCCT results discussed in this report are from the state's interim assessment period, defined by SB 1 as school years 2009, 2010, and 2011. This interim period allows for the development of new instruments mandated by SB 1 and enables a transition period from the old core content to the new core content standards and respective assessments. Instead of an academic index, the interim period uses novice, apprentice, proficient, and distinguished categories to align with the No Child Left Behind Act, and the raw scores from the KCCT selected-response sections were used in this report (Commonwealth. Department. *Kentucky Core*).

In 2010, approximately 367,000 Kentucky students from 648,297 students in the state took the KCCT. The typical difference score between 2009 and 2010 was approximately 0.03, indicating that scores rose slightly from one year to the next.

In 2010, approximately 367,000 Kentucky students out of 648,297 students in the state, from 228 elementary, middle, and high schools, took the KCCT. Raw scores can range from 0.00 (the lowest score) to 1.00 (the highest score). The actual score range for schools ran from 0.47 to 0.95. The typical difference score between 2009 and 2010 was approximately 0.03, indicating that scores rose slightly from one year to the next.

The typical difference score between 2009 and 2010 for elementary schools was 0.02.

The typical difference score between 2009 and 2010 for elementary schools was 0.02, and schools with difference scores at or less than -0.04 for lower-limit outliers or at or greater than 0.09 for upper-limit outliers were identified. Twenty-three schools met these conditions.

The typical difference score between 2009 and 2010 for middle schools was 0.04.

The typical difference score between 2009 and 2010 for middle schools was 0.04, and schools with difference scores at or less than -0.02 for lower-limit outliers or at or greater than 0.08 for upper-limit outliers were identified. Seven schools met these conditions.

The typical difference score between 2009 and 2010 for high schools was 0.03.

The typical difference score between 2009 and 2010 was 0.03 at the high school level, and schools with difference scores at or less than -0.03for lower-limit outlier or at or greater than 0.09 for upper-limit outliers were identified. Six schools met these conditions.

The grade level and school size category with the greatest number of upper-limit outliers was the medium-sized elementary school category, with 7.

Table 3.1 presents KCCT difference score outlier results for each grade level and school size. The grade level and school size category with the greatest number of upper-limit outliers was the medium-sized elementary school category with seven schools. The high variability associated with performance at the elementary level combined with the large number of schools at the elementary level may be potential reasons for the large number of outliers in this category.

Table 3.1
KCCT Difference Score Outliers Based on 2009 and 2010 Scores

Grade Level	School Size	Median	Lower-limit Difference Score	Number of Scores at or Below Lower-limit Score	Upper-limit Difference Score	Number of Scores at or Above Upper-limit Score
Elementary	Tiny	0.03	-0.07	1	0.13	2
	Small	0.02	-0.04	3	0.09	4
	Medium	0.02	-0.02	6	0.07	7
	Large	0.02	-0.04	0	0.09	0
Middle	Tiny	0.05	-0.06	0	0.15	2
	Small	0.03	-0.03	0	0.10	1
	Medium	0.03	-0.01	0	0.07	2
	Large	0.02	-0.01	0	0.06	2
High	Tiny	0.04	-0.06	0	0.15	1
	Small	0.03	-0.04	0	0.09	2
	Medium	0.03	-0.03	0	0.10	1
	Large	0.03	-0.03	1	0.09	1

Note: KCCT is the Kentucky Core Content Test.
Source: OEA staff compilation of Kentucky Department of Education data.

EPAS

Kentucky uses EXPLORE, PLAN, and ACT tests to assess next-level core content readiness.

KRS 158.6453 mandates that all Kentucky public school 8th-, 10th-, and 11th-grade students participate in a testing system that includes components for assessing next-level readiness in English, reading, mathematics, and science. Kentucky uses ACT's EPAS to address these guidelines.[1] EPAS is made up of the EXPLORE, PLAN, and ACT assessments. The details of each test and results of analysis are provided below.

EXPLORE

EXPLORE was designed for use at grades 8 and 9 and has a possible score range of 1 to 25.

EXPLORE, the first test in ACT's EPAS sequence, was designed for use at grades 8 and 9 and has a possible score range of 1 to 25. All 8th-grade students in Kentucky are required to take the test. It consists of 128 questions over four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 30 minutes for each test section. Students earn scores for each content area as well as a composite score. In the 2011 school year, 47,791 students from 324 schools participated in the exam.

The outlier methodology identified nine schools that scored notably lower than expected and five schools that scored notably higher than expected.

Overall, Kentucky's average median difference score was 0.075 meaning that, in general, scores were expected to increase by under 1 point per year. Using the outlier identification methodology, nine schools scored notably lower than expected. Five schools scored notably higher than similar schools. Table 3.2 presents EXPLORE outlier results.

Table 3.2
EXPLORE Difference Score Outliers Based on 2010 and 2011 Scores

School Size	Median	Lower-limit Score	Number of Scores at or Below Lower-limit Score	Upper-limit Score	Number of Scores at or Above Upper-limit Score
Tiny	0.1	-1.75	3	1.85	3
Small	0.2	-1.25	2	1.55	1
Medium	0	-1.43	1	1.58	0
Large	0	-0.70	3	0.90	1

Source: OEA staff compilation of Kentucky Department of Education data.

PLAN

PLAN was designed for use at grade 10 and has a possible score range of 1 to 32.

PLAN, the second test in ACT's EPAS sequence, was designed for use at grade 10 and has a possible score range of 1 to 32. PLAN consists of 145 total questions over four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 30 minutes for the English test section, 40 minutes for the mathematics section, 20 minutes for the reading

[1] Next-level readiness for 8th-grade students refers to readiness for high school, while next-level readiness for 10th- and 11th-grade students refers to readiness for postsecondary education or training.

section, and 25 minutes for the science section. Students earn scores for each content area as well as a composite score. In the 2011 school year, 48,477 students from 230 schools participated in the PLAN exam.

The outlier methodology identified three schools that scored notably lower than expected.

Overall, Kentucky's average median difference score was 0.13. Using the outlier identification methodology, three schools scored notably lower than expected. Two schools scored notably higher than similar schools. Table 3.3 presents PLAN outlier results.

Table 3.3
PLAN Difference Score Outliers Based on 2010 and 2011 Scores

School Size	Median	Lower-limit Score	Number of Scores at or Below Lower-limit Score	Upper-limit Score	Number of Scores at or Above Upper-limit Score
Tiny	0.1	-1.55	2	2.05	1
Small	0.2	-1.40	0	1.80	0
Medium	0.1	-1.10	1	1.30	1
Large	0.1	-1.08	0	1.13	0

Source: OEA staff compilation of Kentucky Department of Education data.

ACT

ACT was designed for use at grades 11 and 12 and has a possible score range of 1 to 36.

ACT, the final component in the EPAS sequence, was designed for use at grades 11 and 12 and has a possible score range of 1 to 36. ACT consists of 215 total questions over four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 45 minutes for the English test section, 60 minutes for the mathematics section, 35 minutes for the reading section, and 35 minutes for the science section. Students earn scores for each content area as well as a composite score. In the 2010 school year, the most recent year for which ACT test data is available, 43,148 juniors from 228 schools participated in the ACT exam.

The outlier methodology found that one school scored notably lower than expected.

Overall, Kentucky's average median difference score was 0.29. Using the outlier identification methodology, one school scored notably lower than expected. No schools scored notably higher than similar schools. Table 3.4 presents ACT outlier results.

Table 3.4
ACT Difference Score Outliers Based on 2009 and 2010 Scores

School Size	Median	Lower-limit Score	Number of Scores at or Below Lower-limit Score	Upper-limit Score	Number of Scores at or Above Upper-limit Score
Tiny	0.35	-2.45	1	2.75	0
Small	0.2	-1.55	0	2.05	0
Medium	0.3	-1.08	0	1.93	0
Large	0.3	-1.05	0	1.75	0

Source: OEA staff compilation of Kentucky Department of Education data.

Considerations

Though the outlier methodology used in this report is not biased by extreme scores as are some methodologies, the method described here only detects drastic changes from one year to the next and does not take into account maintenance of effort over several years.

The outlier procedure described is anchored around a median and quartiles rather than an average score and variance, so it is less likely to be biased by the very outliers it is in place to detect. One weakness of the specific methodology outlined in this report is that it does not take into account maintenance of effort. Another potential limitation is that the outlier analysis described here is only based on 1 year and does not take into account previous years' increases or decreases. Computing a difference score does help identify cases in which large increases or decreases occurred, but it does not identify those cases in which scores may have, at one point or another, increased significantly and remained at a peak plateau over time. Achieving a high performance level and maintaining that level is considered maintenance of effort. In some cases the "effort" is legitimate, while in others it is one or more inappropriate testing practices.

To identify maintenance of effort, regular, descriptive outlier checks on composite, total, and subtest scores would need to be in place.

To identify maintenance of effort, regular, descriptive outlier checks on composite, total, and subtest scores would need to be in place. These checks could be conducted as difference score checks and would be based on single-year performance rather than on comparisons between years. Scores that are identified as high and low outliers based on composite and total scores should be examined as a matter of course because they hold valuable process-related information. Schools with upper outlier scores may be employing strategies that could assist other schools in raising performance levels, or they may be employing inappropriate testing practices. Likewise, schools with composite or total scores in the lower outlier range may be in need of assistance. Either way, follow-up efforts are in order to determine when the spikes or drops occurred and what historic factors may have influenced the change.

Conclusion and Recommendations

A clear, comprehensive, and strategic plan is the essential first step to reducing or eliminating test integrity issues.

A clear, comprehensive, and strategic plan is the essential first step to reducing or eliminating test integrity issues. The system should specifically outline stakeholder expectations at each level of the testing system and should explicitly define integrity breaches, including cheating. Moreover, the system should make clear the processes for detecting potentially unethical practice and explicitly list the consequences of cheating for each level of the model.

If all stakeholders understand that a series of checks and balances exists, then unethical behavior will likely decrease.

If all stakeholders understand that a series of checks and balances exists, then unethical behavior will likely decrease. Potential systematic components may include outlier analyses based on within- and between-group comparisons, erasure analyses, and appropriate forensic analyses.[2]

An acceptable system of test security practices would provide a multistep process that would include protocols for identifying and following up on suspicious activity.

A comprehensive system of test result security practices, such as those suggested by the CCSSO, would provide a multistep process. Such a process would include protocols for identifying and following up on suspicious activity or testing-related activities reported as suspicious or questionable. The system should be established to identify and inspect scores that appear inconsistent with past performance or are in excess of generally acceptable ranges of growth or decline.

A robust testing security system aimed at identifying potentially inappropriate education test practices should include checkpoint analyses aimed at identifying outliers based on typical or expected performance trends, a third-party review of the testing system, security procedures and scores, protocols for following up with evidence collection on suspicious testing activity, a process for how appropriate agencies will be assigned to the investigation of questionable scores, and a clear stakeholder communication plan.

The following are parts of a robust system. The more technically advanced steps, such as the forensic analysis of test forms, come with high costs. However, a number of the steps can be conducted at the state level or through the test vendor with little additional cost.

- There should be at least one automatic checkpoint at which all aggregate responses are available for review. A good example of a potential checkpoint would be the first postscoring report. If a batch of score sheets from a particular district, school, or classroom yields results that are highly similar, then those results may be eligible for further review.
- There should be at least one automatic checkpoint at which school-level responses are aggregated and compared to that school's scores for the 2 or 3 testing years prior to the current testing year.[3] If a school exhibits unlikely growth or decline,

[2] Within-group comparisons are comparisons in which results from a specific pool are compared. Between-group comparisons are comparisons in which results from separate pools are compared or results from a single group are compared to scores from a similar group.
[3] Disaggregated analyses would be part of this process as well--for example, special education compared to regular education scores within different school sizes.

current and past responses from that school may be eligible for review.

- There should be a third-party review.
- There should be a protocol for how follow-up information will be collected in cases of suspected improper or dishonest testing practice.
- There should be a process for how appropriate agencies, both state and federal, will be assigned to the investigation of questionable scores with a listing of potential consequences per agency. The process should recognize that there are multiple uses of test scores, and there may need to be a multiagency team to deal with the most egregious cases.
- There should be a clear and comprehensive communication plan to inform all stakeholders and test users about best testing practices and associated consequences for violating state testing regulations.

Below are four recommendations designed to provide direction in the important undertaking of reducing inappropriate and unethical education testing practices.

Recommendation 3.1

Recommendation 3.1

The Kentucky Department of Education should continue to augment training for district assessment coordinators, building assessment coordinators, and proctors to include clear communication of post-test analyses aimed at identifying outliers along with the consequences of being investigated and found guilty of unethical or inappropriate testing practices.

Recommendation 3.2

Recommendation 3.2

The Kentucky Department of Education should track district assessment coordinators, building assessment coordinators, and assessment proctors via a unique identifier to link those education professionals with specific testing events.

Recommendation 3.3

Recommendation 3.3

The Kentucky Department of Education should build into the education test and assessment data plan, through third-party vendors/agencies, internally, or through augmented existing contracts or study plans, analyses commensurate with available funding aimed at identifying outliers and other data irregularities.

Recommendation 3.4

Recommendation 3.4

The Kentucky Department of Education should augment the current formal process for attending to test violation allegations and complaints with comprehensive plans commensurate with available funding for addressing wide-scale test security allegations with specific state agency roles defined.

Recommendation 3.5

Recommendation 3.5

The Kentucky Department of Education should continue plans to contract with an education test and assessment security company to obtain an audit of the state's current and proposed education testing system.

While it is impossible to eradicate cheating and less explicit but equally harmful ways of increasing test scores without increasing student learning, it is possible to put into place testing systems that recognize potential threats to making valid inferences and that provide guidance in cases when unethical or inappropriate practices do arise.

Emphasis on education tests and their role in federal and state accountability models increases the likelihood that education professionals and students may feel pressured to raise test scores by any means necessary. While it is impossible to eradicate cheating and less explicit but equally harmful ways of increasing student learning, it is possible to put into place testing systems that recognize potential threats to making valid inferences and that provide guidance in cases when unethical or inappropriate practices do arise. Proactivity, vigilance, and clear communication in the form of professional training are key elements of any system aimed at providing the most accurate education test and assessment data possible.

Works Cited

Allen, Mary J., and Wendy M. Yen. *Introduction to Measurement Theory*. Belmont: Wadsworth, 1979.

Berk, Kenneth N., and Patrick Carey. *Data Analysis with Microsoft Excel.* Pacific Grove: Duxbury, 2000.

Caveon Test Security. *Cost Overview.* Received via email from Ken Draut to Keith White. April 7, 2011.

---. *Sample Forensics Report.* <http://www.caveon.com/reports/df_sample_reports.pdf> (accessed March 8, 2011).

Commonwealth of Kentucky. Auditor of Public Accounts. *Performance Audit of the Kentucky Department of Education's Oversight of State Assessment Contracts.* <http://www.auditor.ky.gov/Public/Audit_Reports/Archive/2009EducationAssessmentContractsPerformanceaudit.pdf> (accessed March 27, 2012).

---. Department of Education. *Kentucky Assessment System: Validation and Research Agenda Biennial Plan for Validation Studies 2010 - 2012.* Email from Ken Draut to Keith White. July 1, 2010.

---. ---. *Kentucky Core Content Test (KCCT) Interim Performance Report Interpretive Guide 2010.* <http://www.education.ky.gov/NR/rdonlyres/4757A9BC-7782-4F8C-8AC6-DDC27C1B840D/0/2010_KCCT_Interpretive_Guide.pdf> (accessed Jan. 6, 2011).

---. ---. *Reported Allegations by Year.* <http://www.education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/District+Support/Assessment+Regulations+and+Training/Combined+Regulations+Assessment+Training.htm> (accessed Jan. 11, 2011).

---. Office of the State Budget Director. *2010–2012 Budget of the Commonwealth. Volume I, Part B.* <http://www.osbd.ky.gov/NR/rdonlyres/F826D66D-15AE-48C0-88AC-4F6430A1E451/0/1012BOCVolumeIB.pdf> (accessed March 2, 2011).

Council of Chief State School Officers and the Association of Test Publishers. *Operational Best Practices for Statewide, Large-Scale Assessment Programs*. 2010.

Duncan, Arne. *Key Policy Letters from the Education Secretary or Deputy Secretary*. Issued to Chief State School Officers June 24, 2011. <http://www2.ed.gov/print/policy/elsec/guid/secletter/110624.html> (accessed Aug. 30, 2011).

Fremer, John, and Steve Addicott. *Test Security Lessons from the European American Test Publishers Conference*. Webinar. Presented by Caveon. Oct. 21, 2010.

Galloway, Jim. "Good Cop, Bad Cop, and the Georgia Testing Scandal." <http: //blogs.ajc.com/ political-insider-jim-galloway/2010/02/13/good-cop-bad-cop-and-the-georgia-testing-scandal/> (accessed Jan. 6, 2011).

Heiman, Gary W. *Basic Statistics for the Behavioral Sciences*. 2nd Ed. Boston: Houghton Mifflin, 1996.

Innes, Richard. "Hazard-Herald confirms: Perry County educators under investigation for cheating on ACT." <http://bluegrasspolicy-blog.blogspot.com/2011/03/hazard-herald-confirms-perry-county_12.html> (accessed March 12, 2011).

Koretz, Daniel. *Evidence Pertaining to the Validity of Score Gains on the Kentucky Instructional Information System (KIRIS).* Paper. National Council of Measurement in Education annual meeting. April 16, 1998.

---. *Measuring Up.* Cambridge: Harvard Univ. Press, 2008.

---. *Validation Project Notes.* Memorandum. June 12, 2011.

Nichols, Sharon L., and David C. Berliner. *The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing.* Education Policy Studies Laboratory. March 2005.

Resmovits, Joy. "Educators Accused of Tampering With Students' Tests from D.C. to Pennsylvania."
<http://www.huffingtonpost.com/2011/07/11/educators-tamper-students-tests_n_895179.html> (accessed July 16, 2011).

Ritchie, Cris. "EPSB Moves to Open Inquiry Into ACT Score in Perry." *Hazard Herald*. March 8, 2011.

Shoemaker, Jack. SAS Paper 161. "Robust Outlier Identification Using SAS."
<http://www2.sas.com/proceedings/sugi24/Infovis/p161-24.pdf> (accessed Dec. 10, 2010).

Sorensen, Donald. *50 Ways Students Cheat on Tests.* Caveon Test Security Report.
<http://www.caveon.com/downloads/50_Ways_12-15-10.pdf> (accessed March 27, 2012).

---. *2008 State Education Test Security Survey Results.* Caveon Test Security Report.
<http://www.caveon.com/downloads/Caveon_2008_Education_Survey_Results.pdf> (accessed March 27, 2012).

Sternberg, Robert J., and Wendy M. Williams. *Educational Psychology*. Boston: Allyn and Bacon, 2002.

Tukey, John W. *Exploratory Data Analysis*. Boston, MA: Addison-Wesley, 1977.

Wert, James E., Charles O. Neidt, and J. Stanley Ahmann. *Statistical Methods in Educational and Psychological Research.* New York: Appleton Century Crofts, Inc., 1954.

Appendix A

Accountability Assessment Parameters

Kentucky Core Content Test (KCCT) Number of Test Items by Grade Level

Grade Level	Reading	Math	Science	Social Studies
3	3 OR 45 MC 210 minutes	5 OR 38 MC 210 minutes	Not tested	Not tested
4	4 OR 39 MC 240 minutes	5 OR 38 MC 210 minutes	4 OR 32 MC 160 minutes	Not tested
5	4 OR 39 MC 240 minutes	5 OR 38 MC 210 minutes	Not tested	4 OR 32 MC 160 minutes
6	4 OR 39 MC 240 minutes	5 OR 38 MC 210 minutes	Not tested	Not tested
7	4 OR 39 MC 240 minutes	5 OR 38 MC 210 minutes	4 OR 32 MC 160 minutes	Not tested
8	4 OR 39 MC 240 minutes	5 OR 38 MC 210 minutes	Not tested	4 OR 32 MC 160 minutes
10	4 OR 39 MC 240 minutes	Not tested	Not tested	Not tested
11	Not tested	5 OR 38 MC 210 minutes	4 OR 32 MC 160 minutes	4 OR 32 MC 160 minutes

Note: OR refers to open-response, and MC refers to multiple-choice questions.
Source: OEA staff compilation of information from the KCCT Spring 2010 District and Building Assessment Coordinators' Manual, and KCCT Interim Performance Report Interpretive Guide (2010).

Educational Planning and Assessment System

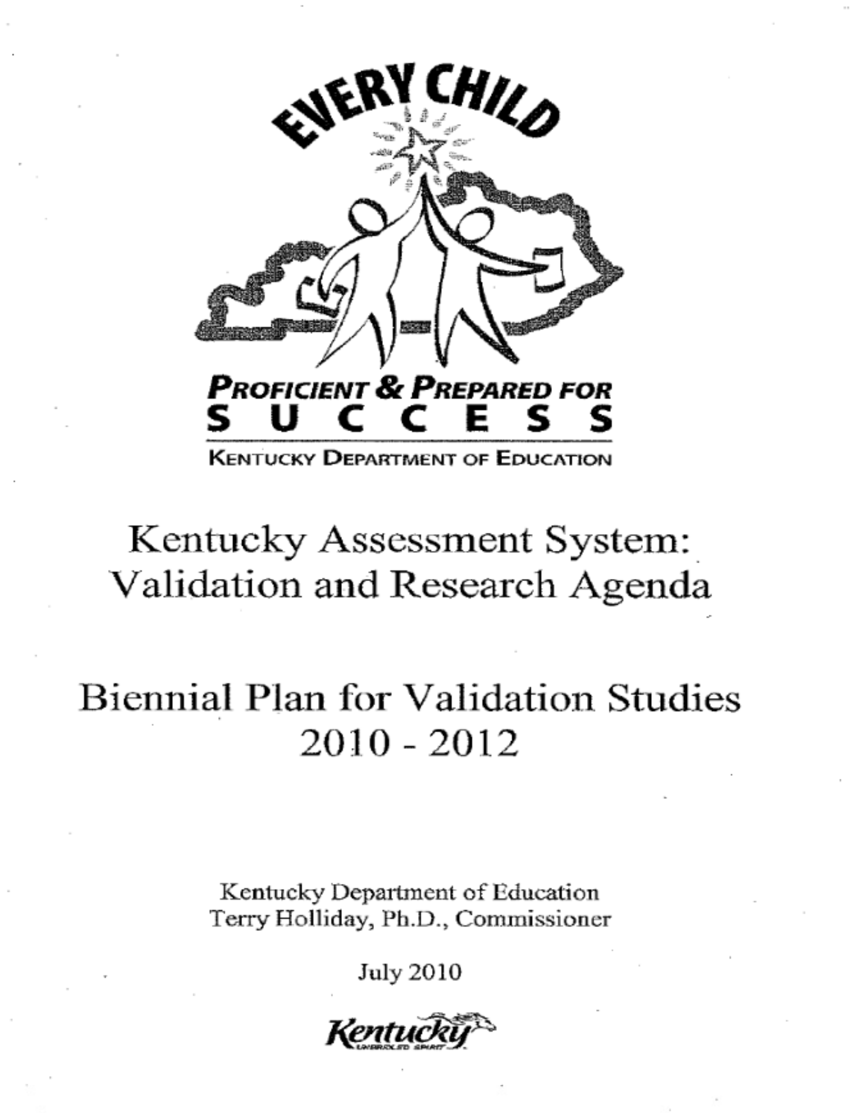
EXPLORE, the first test in ACT's EPAS sequence, was designed for use at grades 8 and 9 and has a possible score range of 1 to 25. EXPLORE consists of a total of 128 questions in four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 30 minutes for each test section. Students earn scores for each content area as well as a composite score.

PLAN, the second test in ACT's EPAS sequence, was designed for use at grade 10 and has a possible score range of 1 to 32. PLAN consists of a total of 145 questions in four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 30 minutes for the English section, 40 minutes for the mathematics section, 20 minutes for the reading section, and 25 minutes for the science section. Students earn scores for each content area as well as a composite score.

ACT, the final component in the EPAS sequence, was designed for use at grades 11 and 12 and has a possible score range of 1 to 36. ACT consists of a total of 215 questions in four test sections covering English, mathematics, reading, and science. Nonaccommodated students are allowed 45 minutes for the English section, 60 minutes for the mathematics section, 35 minutes for the reading section, and 35 minutes for the science section. Students earn scores for each content area as well as a composite score.

Appendix B

Kentucky Assessment System:
The Kentucky Department of Education's
Biennial Plan for Validation Studies 2010-2012



EVERY CHILD

PROFICIENT & PREPARED FOR SUCCESS

KENTUCKY DEPARTMENT OF EDUCATION

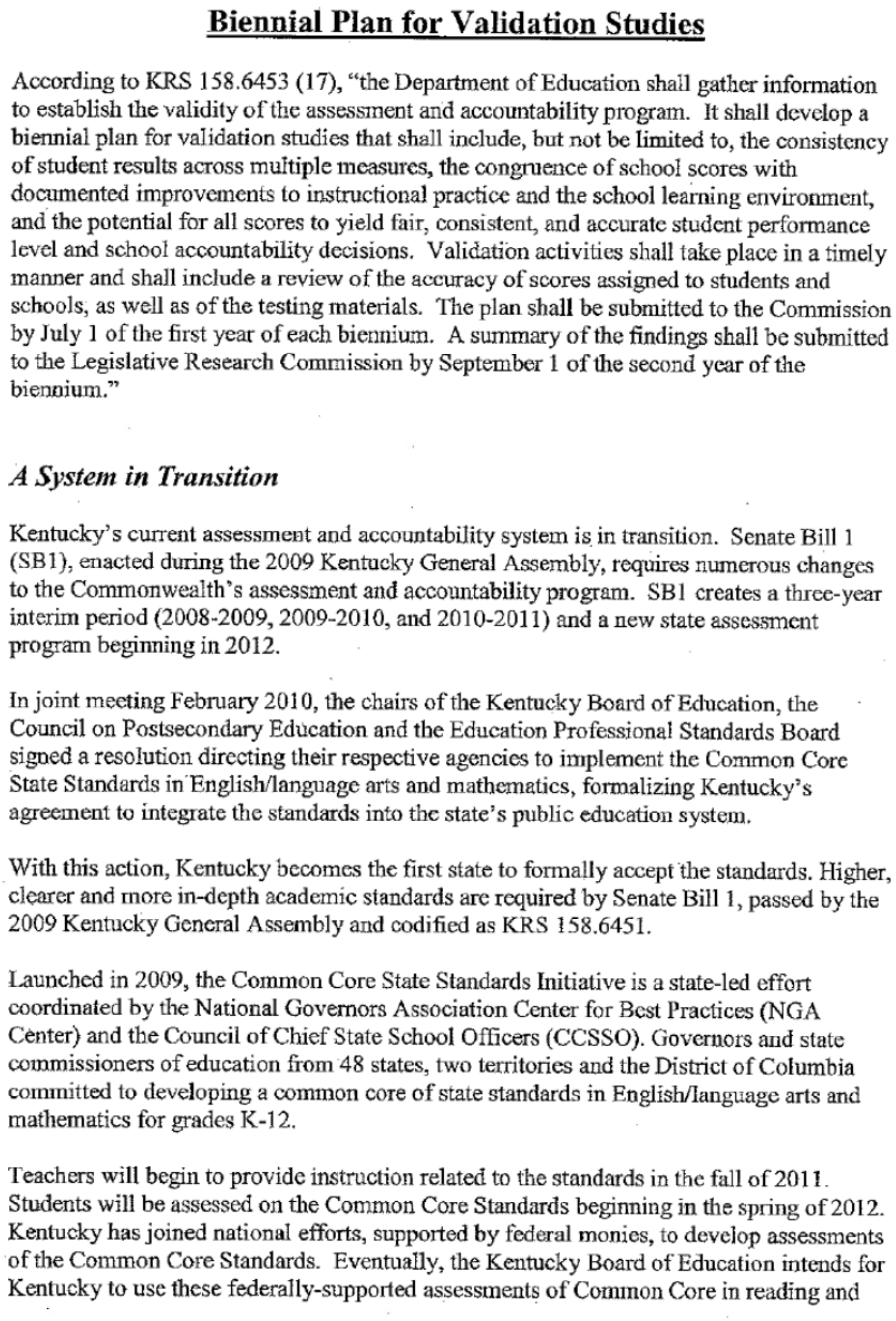
Kentucky Assessment System:
Validation and Research Agenda

Biennial Plan for Validation Studies
2010 - 2012

Kentucky Department of Education
Terry Holliday, Ph.D., Commissioner

July 2010

Kentucky



Biennial Plan for Validation Studies

According to KRS 158.6453 (17), "the Department of Education shall gather information to establish the validity of the assessment and accountability program. It shall develop a biennial plan for validation studies that shall include, but not be limited to, the consistency of student results across multiple measures, the congruence of school scores with documented improvements to instructional practice and the school learning environment, and the potential for all scores to yield fair, consistent, and accurate student performance level and school accountability decisions. Validation activities shall take place in a timely manner and shall include a review of the accuracy of scores assigned to students and schools, as well as of the testing materials. The plan shall be submitted to the Commission by July 1 of the first year of each biennium. A summary of the findings shall be submitted to the Legislative Research Commission by September 1 of the second year of the biennium."

A System in Transition

Kentucky's current assessment and accountability system is in transition. Senate Bill 1 (SB1), enacted during the 2009 Kentucky General Assembly, requires numerous changes to the Commonwealth's assessment and accountability program. SB1 creates a three-year interim period (2008-2009, 2009-2010, and 2010-2011) and a new state assessment program beginning in 2012.

In joint meeting February 2010, the chairs of the Kentucky Board of Education, the Council on Postsecondary Education and the Education Professional Standards Board signed a resolution directing their respective agencies to implement the Common Core State Standards in English/language arts and mathematics, formalizing Kentucky's agreement to integrate the standards into the state's public education system.

With this action, Kentucky becomes the first state to formally accept the standards. Higher, clearer and more in-depth academic standards are required by Senate Bill 1, passed by the 2009 Kentucky General Assembly and codified as KRS 158.6451.

Launched in 2009, the Common Core State Standards Initiative is a state-led effort coordinated by the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO). Governors and state commissioners of education from 48 states, two territories and the District of Columbia committed to developing a common core of state standards in English/language arts and mathematics for grades K-12.

Teachers will begin to provide instruction related to the standards in the fall of 2011. Students will be assessed on the Common Core Standards beginning in the spring of 2012. Kentucky has joined national efforts, supported by federal monies, to develop assessments of the Common Core Standards. Eventually, the Kentucky Board of Education intends for Kentucky to use these federally-supported assessments of Common Core in reading and mathematics. The federal grant requires the assessments to be available in 2014-2015. Between 2011-2012 and 2013-2014, Kentucky must produce its own assessment of the Common Core for reading and mathematics and Kentucky standards for science, social studies and writing.

Summary of Biennial Plan Studies

The evolving nature of Kentucky's assessment and accountability program shapes the validation and research agenda for the biennium. This document reflects both research contracted through Fiscal Year 2010-2011 for the Kentucky Core Content Test and future research connected to the new assessment system beginning in 2011-2012 as required in Senate Bill 1.

This section includes a brief description of each study currently planned for this biennium. A more detailed description of each study is present in the last section of this document.

1. Annual Third-Party Checking of KCCT Scaling and Equating

Following a psychometric processing error in the early 1990s, procedures were added to ensure a thorough, independent check of data before reporting. Psychometric processing continues to undergo parallel, independent, third-party analysis within the operational time frame to ensure the accuracy of the computations.

This project occurs in real time and mirrors the testing contractor's schedule (June-July typically, with a report delivered in August-September).

2. Alignment Study to Investigate the Utility of Kentucky's Item Pool for Measuring Common Core and State Standards

Until the availability of federal-supported assessments of the Common Core, Kentucky will comply with SB1 requirements by implementing an assessment in 2011-2012 of Common Core in reading and mathematics and Kentucky standards in the other content areas (science, social studies and writing) required by SB1.

As Kentucky shifts from assessing Kentucky's *Core Content for Assessment 4.1* to the Common Core, questions arise regarding the similarities and differences between the two sets of standards. If the content were entirely different, an entirely new assessment must be created to measure student academic achievement. On the other hand, if the two sets of standards were essentially the same, an assessment may be constructed from existing items. This study will determine the match between mathematics and reading items in Kentucky's item pool and Common Core standards. It will identify items that may be immediately used, items that require editing and new field testing and what content standards require new item development.

3. Research and Validity Studies in Support of Senate Bill 1

The Department plans to present to the National Technical Advisory Panel on Assessment and Accountability (NTAPAA) for committee advice on additional studies needed for successful implementation of Senate Bill 1. With the assistance of NTAPAA, the studies will be prioritized and completed within the guidelines of the current vendor contract(s).

The research activities cover a broad range and, among others, include the following possible studies:

A. Periodic Alignment Studies of Norm-referenced Tests
B. Annual Multiple Assessments/Convergent Validity Evidence
C. Annual Item Content, Item Difficulty, and Item Type Mapping for Assessment (2011-2012 through 2013-2014)
D. School Classification Accuracy Analyses
E. Student Classification Accuracy Analyses

Detailed Description of Biennial Plan Studies

A more detailed description for each study is presented in this section. Each study is described in terms of:

- Purpose (Why do the research?)
- Audience (Who will use the results of the research and how will they use it?)
- Methodology (How will the research be conducted?)
- Final Product (How will the results be packaged and distributed?)

1. Annual Third-Party Checking of KCCT Scaling and Equating

Purpose (Why do the research?)

Psychometric processing undergoes parallel, independent, third-party analysis within the operational time frame to ensure the accuracy of the computations.

Audience (Who will use the results of the research and how will they use it?)

All stakeholders interested in the processing accuracy for KCCT will be assured by this work. However, the primary audience is the technical staff of the primary contractor and those who might perform a technical review or audit of the Kentucky Core Content Test.

Psychometric information is required during the peer review process conducted by the United States Department of Education.

Methodology (How will the research be conducted?)

In parallel with the primary psychometric contractor (currently Measured Progress), the third-party contractor (currently HumRRO) will replicate all psychometric scaling and equating. Results will be shared between Measured Progress and HumRRO, with all discrepancies investigated by both parities until a common resolution is reached and applied.

Final Product (How will the results be packaged and distributed?)

The final product of this work is a brief technical report describing the third-party contractor's results, including resolution of initial discrepancies with the primary contractor. For this research, the procedures followed and the intermediate product yielded from the procedures are perhaps more valuable than the final report. This intermediate product is a summary table in which both Measured Progress and HumRRO indicate mile-post findings for their scaling and equating work. The table is shared back and forth via secure web site and e-mail notifications. It is updated continually until agreement is reached for all KCCT grade/subject combinations. Psychometric, data and contract managers with the Kentucky Department of Education monitor this iterant process and associated communications.

2. Alignment Study to Investigate the Utility of Kentucky's Item Pool for Measuring Common Core Standards

Purpose (Why do the research?)

Until the availability of federally-supported assessments of the Common Core, Kentucky will comply with SB1 requirements by implementing an assessment beginning in 2011-2012 of the Common Core in reading and mathematics and Kentucky standards in the other content areas (science, social studies and writing). The study will provide the foundational detail necessary for creating an assessment of the Common Core. It will identify exiting Kentucky test items that measure the Common Core, test items that need revision and standards requiring the development of new items.

Audience (Who will use the results of the research and how will they use it?)

The findings will be used to inform policy stakeholders (the Kentucky Board of Education, KDE, NTAPAA, and the Office of Education Accountability) of the link between KCCT reading and mathematics existing items and Common Core standards. The complete findings with secure test item information will be directed to KDE and NTAPAA. A summary appropriate for public discussion will be developed for other stakeholders.

Methodology (How will the research be conducted?)

A modified version of the Webb alignment method is used to complete this project. Briefly, test items are matched to test standards by item raters. The item raters also determine the cognitive complexity level of the items on a 1-4 Depth of Knowledge (DOK) scale. Each of the standards was also given a DOK rating. The numbers of items per each standard, as well as descriptive statistics for items-by-categories of standard are then calculated. Comparisons are made between item DOK level and standard DOK level. Taken together the results of this study provide a map for determining which existing items might be used for a Common Core Assessment, as well as indicating where item development should be focused to address gaps in content representation.

Final Product (How will the results be packaged and distributed?)

The vendor will provide written reports to the Department, one for middle school grades and one for elementary school grades in the summer 2010.

3. Research and Validity Studies in Support of Senate Bill 1

Purpose (Why do the research?)

The Kentucky Department of Education is charged with "Maintaining a vigorous ongoing program of research and documentation of the effects of the assessment and accountability system on Kentucky schools." KDE must also comply with specific research found in Kentucky Revised Statutes (KRS) and to ensure compliance with the federal No Child Left Behind Act of 2001, 20 U.S.C. secs. 6301 et seq., or its successor. The research activities cover a broad range and, among others, include the following possible studies. With the assistance of NTAPAA, the studies will be prioritized and completed within the guidelines of the current vendor contract(s).

A. Periodic Alignment Studies of Norm-referenced Tests

The KDE will conduct an alignment study of the norm-referenced test (NRT) used in conjunction with the criterion-referenced test (CRT) required in SB1. According to KRS 158.6453 (21) "The Kentucky Board of Education shall conduct periodic alignment studies that compare the norm-referenced tests required under subsection (5) of Section 2 of this Act with the standards in the different content areas to determine how well the norm-referenced test align and adequately measure the depth of knowledge and breadth of Kentucky's academic content standards. Based on its findings from the studies, the board may decrease the number of required criterion-referenced items required under subsection (5) of Section 2 of this Act."

B. Annual Multiple Assessments/Convergent Validity Evidence

Valid tests, by definition, produce test scores that behave in theoretically predictable ways. Therefore, an important method for establishing the validity of any given test is to systematically observe relations between scores that the test produces and various other indicators that are expected to be associated with the test scores. Readily available data include student demographic data, questionnaire data, student scores and data from the state-required assessments (KCCT, NRT, PLAN, EXPLORE and ACT). These data allow examination of convergent and discriminate validity relations as well as analyses of differences in performance related to gender, ethnicity and socio-economic status. Trends over time are also of interest.

C. Annual Item Content, Item Difficulty, and Item Type Mapping for Assessment (2011-2012 through 2013-2014)

Item mapping is simply displaying two-way distributions of item content, difficulty, and depth of knowledge (DOK) within and across the test forms administered for each assessed grade/subject combination. These maps summarize content validity and comparability of forms in terms of content, difficulty, and DOK requirements and are designed to answer the following questions:

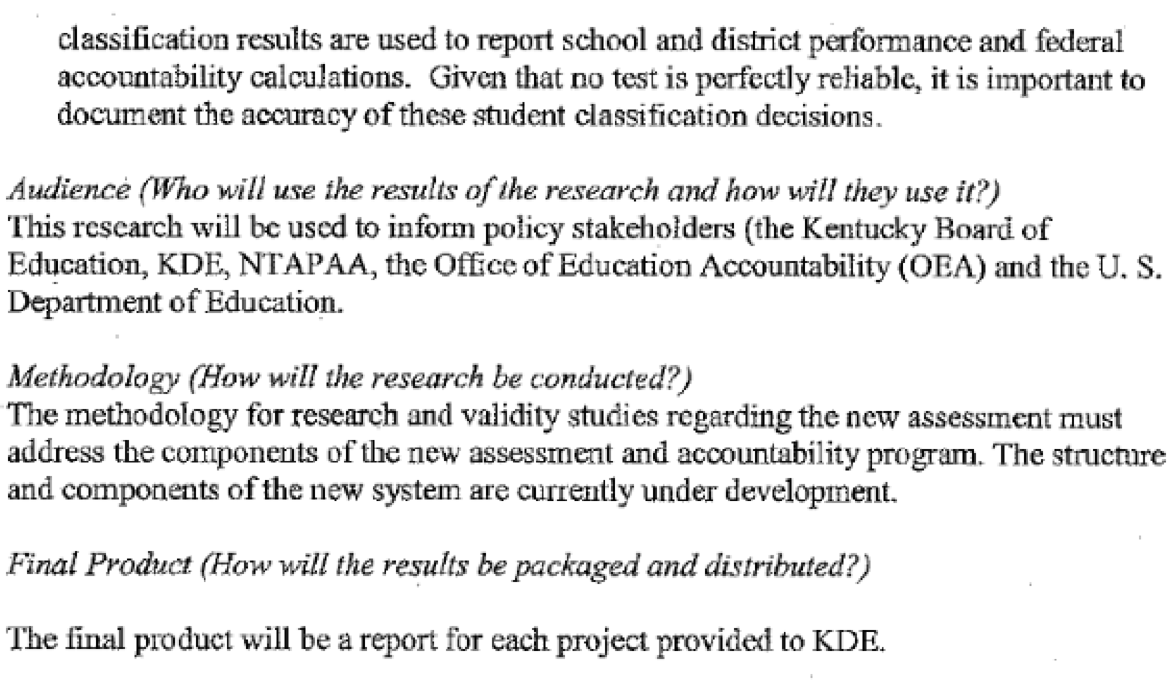
1. Is the content equally distributed across Novice, Apprentice, Proficient and Distinguished (NAPD) achievement levels?
2. How is the content distributed by type of item (multiple choice versus open response)?
3. What is the distribution of item types across NAPD scale?
4. How are the item-level DOK ratings distributed across NAPD achievement levels and type of item?

D. School Classification Accuracy Analyses

The Kentucky Board of Education will implement a new accountability system for classification of Kentucky public schools. The system will report multiple measures. Possible measures under discussion reflect growth, gap and achievement information. Measures will be based on required assessments (Kentucky's new CRT/NRT test in grades 3-8, End of Course tests at high school, readiness test (EXPLORE, PLAN and the ACT). An additional critical focus will monitor progress in college and career readiness. Data will also be available from program reviews for arts/humanities, practical living and career studies and writing. This array of data should provide a very stable base for making classification decisions; however, because no measurement system is perfect, it is important to specifically document this accuracy. Additional analyses related to NCLB or other federally-required classification will also be considered during this phase of the project.

E. Student Classification Accuracy Analyses

The new Kentucky assessment system will be administered annually in 21 different grade/subject combinations (e.g., Grade 4 Reading, Grade 8 Mathematics). Based on responses to test items, students are classified into one of four basic categories (Novice, Apprentice, Proficient and Distinguished, commonly referred to as NAPD) and the



classification results are used to report school and district performance and federal accountability calculations. Given that no test is perfectly reliable, it is important to document the accuracy of these student classification decisions.

Audience (Who will use the results of the research and how will they use it?)
This research will be used to inform policy stakeholders (the Kentucky Board of Education, KDE, NTAPAA, the Office of Education Accountability (OEA) and the U. S. Department of Education.

Methodology (How will the research be conducted?)
The methodology for research and validity studies regarding the new assessment must address the components of the new assessment and accountability program. The structure and components of the new system are currently under development.

Final Product (How will the results be packaged and distributed?)

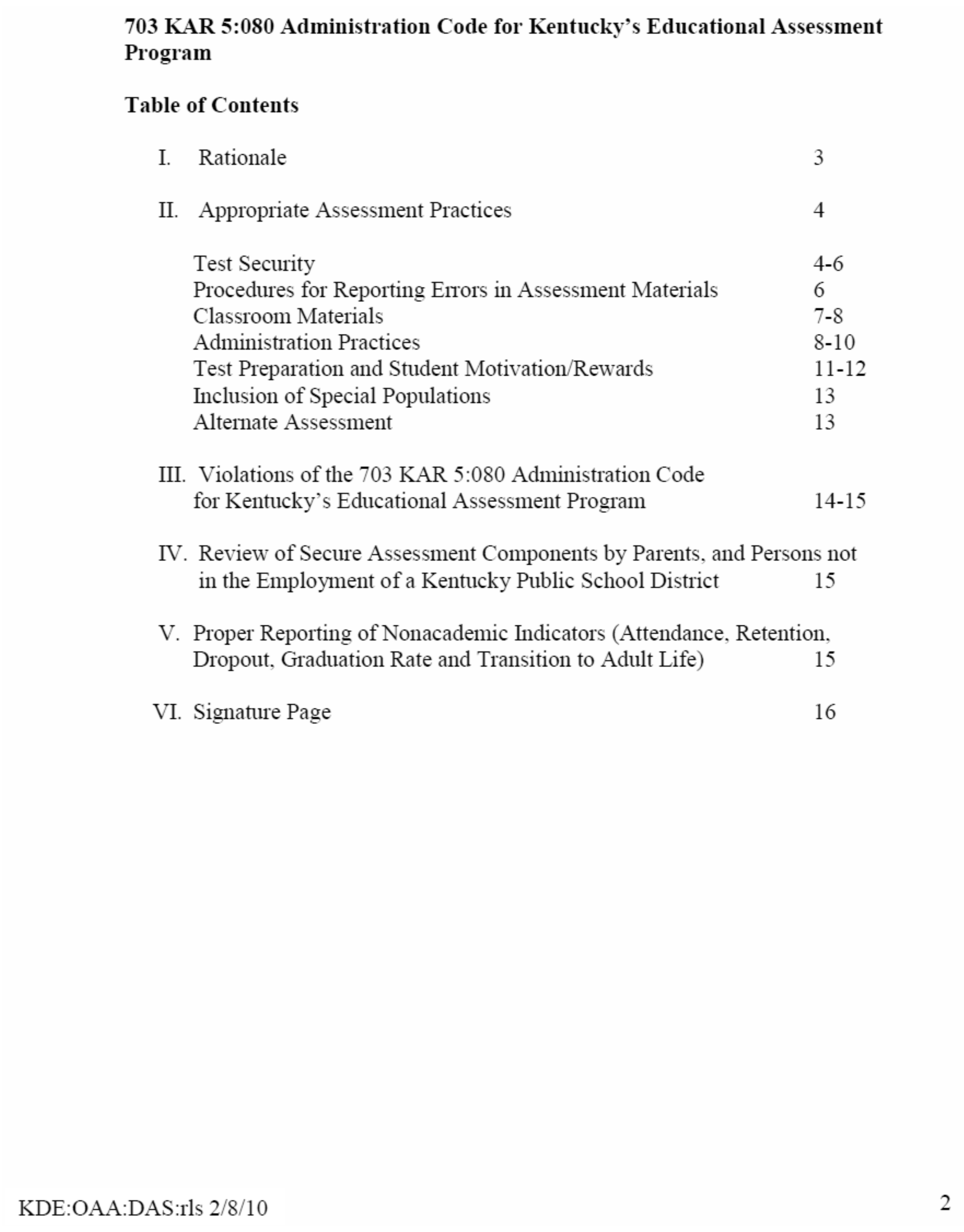
The final product will be a report for each project provided to KDE.

Appendix C

703 KAR 5:080

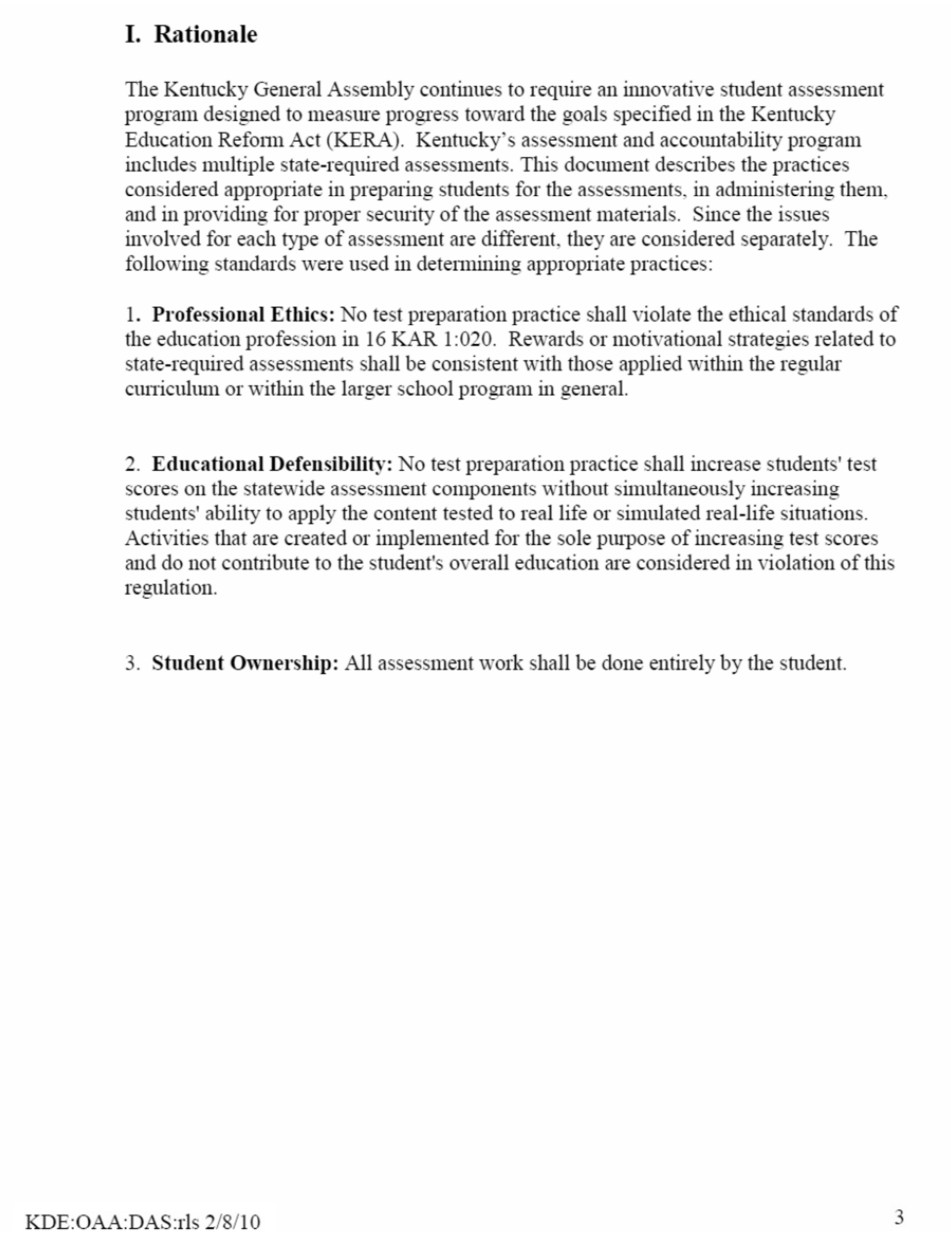
Administration Code for Kentucky's Educational Assessment Program

Kentucky Department of Education



703 KAR 5:080 Administration Code for Kentucky's Educational Assessment Program

Table of Contents



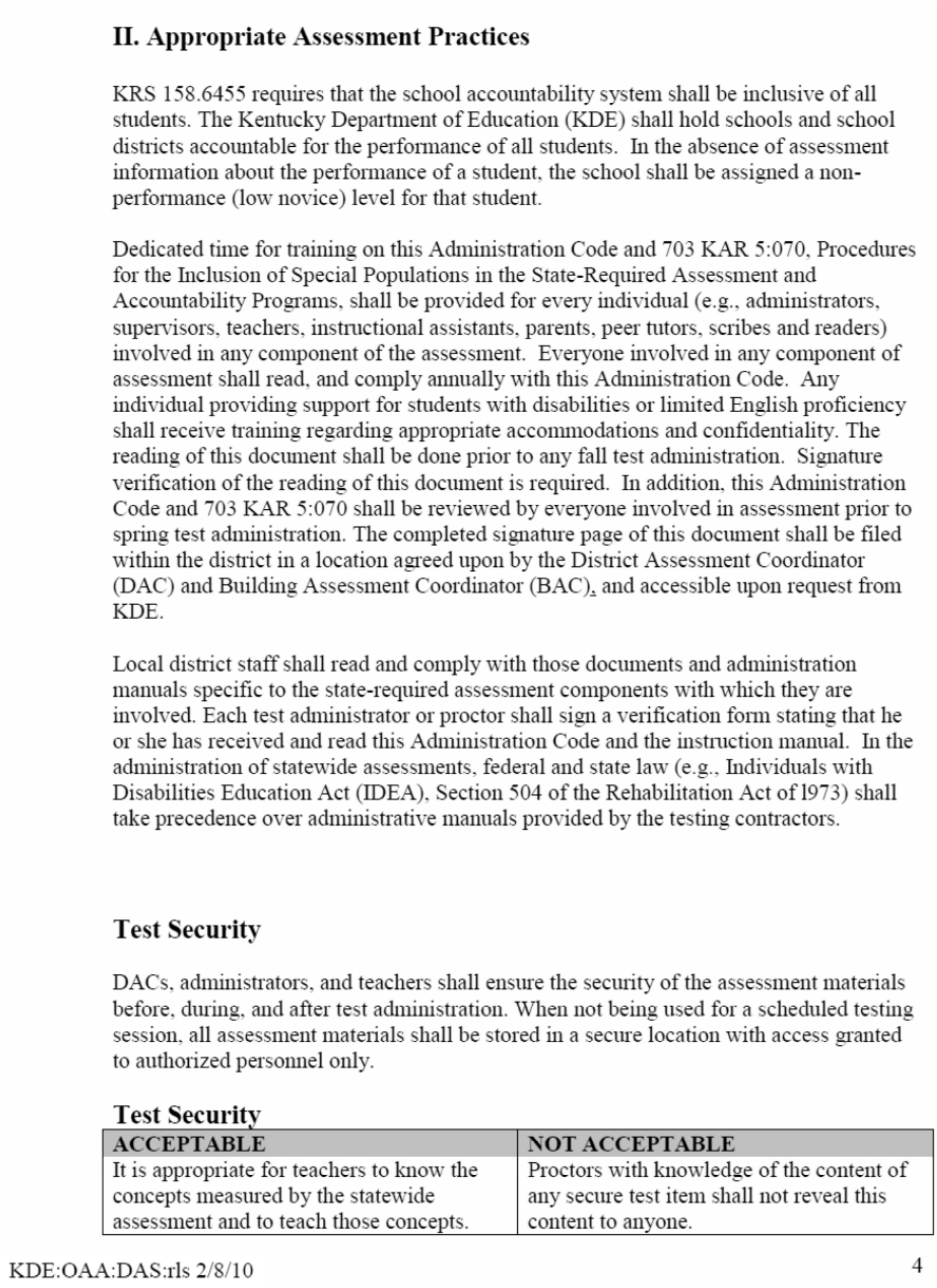
I. Rationale

The Kentucky General Assembly continues to require an innovative student assessment program designed to measure progress toward the goals specified in the Kentucky Education Reform Act (KERA). Kentucky's assessment and accountability program includes multiple state-required assessments. This document describes the practices considered appropriate in preparing students for the assessments, in administering them, and in providing for proper security of the assessment materials. Since the issues involved for each type of assessment are different, they are considered separately. The following standards were used in determining appropriate practices:

1. **Professional Ethics:** No test preparation practice shall violate the ethical standards of the education profession in 16 KAR 1:020. Rewards or motivational strategies related to state-required assessments shall be consistent with those applied within the regular curriculum or within the larger school program in general.

2. **Educational Defensibility:** No test preparation practice shall increase students' test scores on the statewide assessment components without simultaneously increasing students' ability to apply the content tested to real life or simulated real-life situations. Activities that are created or implemented for the sole purpose of increasing test scores and do not contribute to the student's overall education are considered in violation of this regulation.

3. **Student Ownership:** All assessment work shall be done entirely by the student.



II. Appropriate Assessment Practices

KRS 158.6455 requires that the school accountability system shall be inclusive of all students. The Kentucky Department of Education (KDE) shall hold schools and school districts accountable for the performance of all students. In the absence of assessment information about the performance of a student, the school shall be assigned a non-performance (low novice) level for that student.

Dedicated time for training on this Administration Code and 703 KAR 5:070, Procedures for the Inclusion of Special Populations in the State-Required Assessment and Accountability Programs, shall be provided for every individual (e.g., administrators, supervisors, teachers, instructional assistants, parents, peer tutors, scribes and readers) involved in any component of the assessment. Everyone involved in any component of assessment shall read, and comply annually with this Administration Code. Any individual providing support for students with disabilities or limited English proficiency shall receive training regarding appropriate accommodations and confidentiality. The reading of this document shall be done prior to any fall test administration. Signature verification of the reading of this document is required. In addition, this Administration Code and 703 KAR 5:070 shall be reviewed by everyone involved in assessment prior to spring test administration. The completed signature page of this document shall be filed within the district in a location agreed upon by the District Assessment Coordinator (DAC) and Building Assessment Coordinator (BAC), and accessible upon request from KDE.

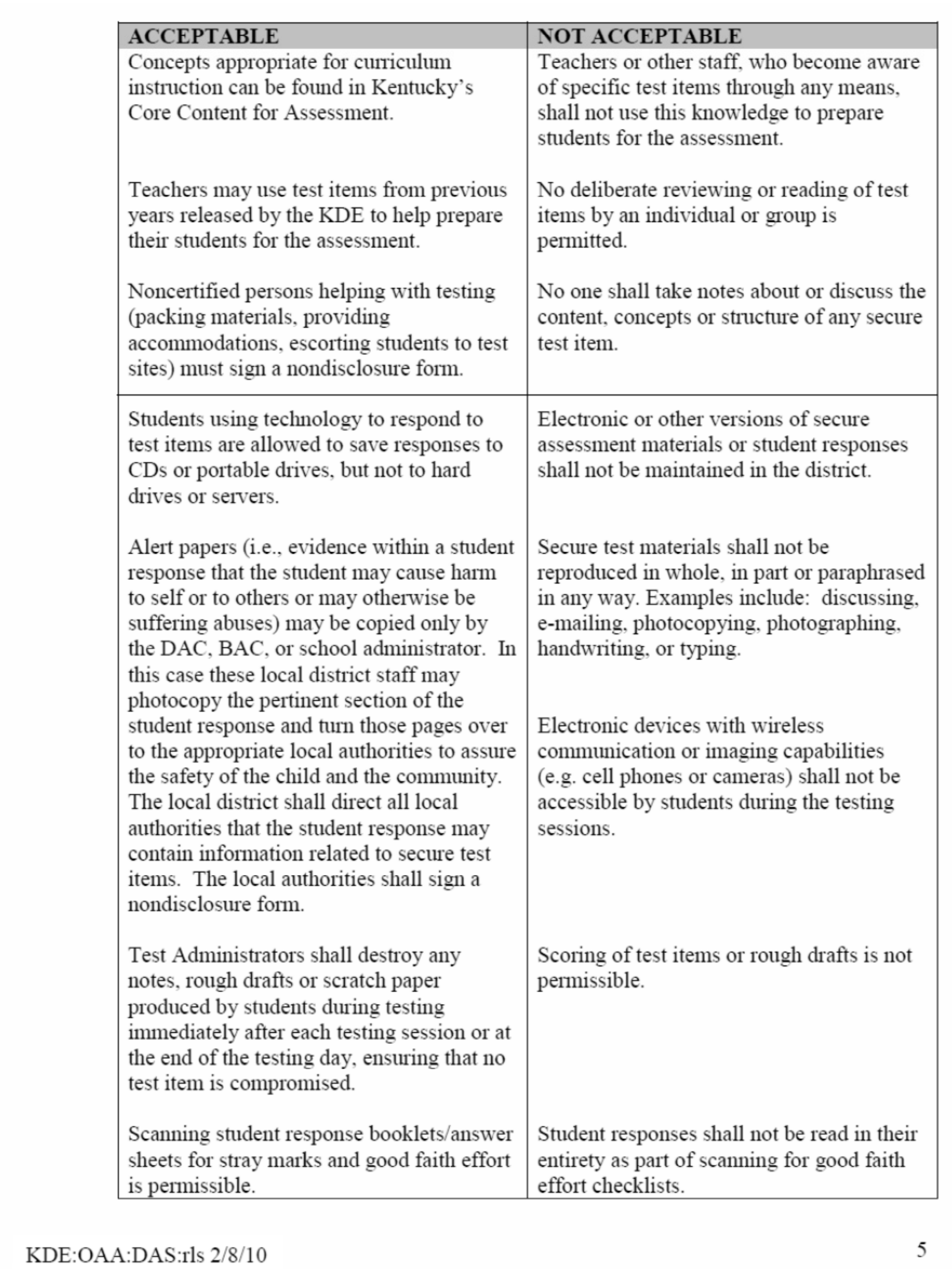
Local district staff shall read and comply with those documents and administration manuals specific to the state-required assessment components with which they are involved. Each test administrator or proctor shall sign a verification form stating that he or she has received and read this Administration Code and the instruction manual. In the administration of statewide assessments, federal and state law (e.g., Individuals with Disabilities Education Act (IDEA), Section 504 of the Rehabilitation Act of l973) shall take precedence over administrative manuals provided by the testing contractors.

Test Security

DACs, administrators, and teachers shall ensure the security of the assessment materials before, during, and after test administration. When not being used for a scheduled testing session, all assessment materials shall be stored in a secure location with access granted to authorized personnel only.

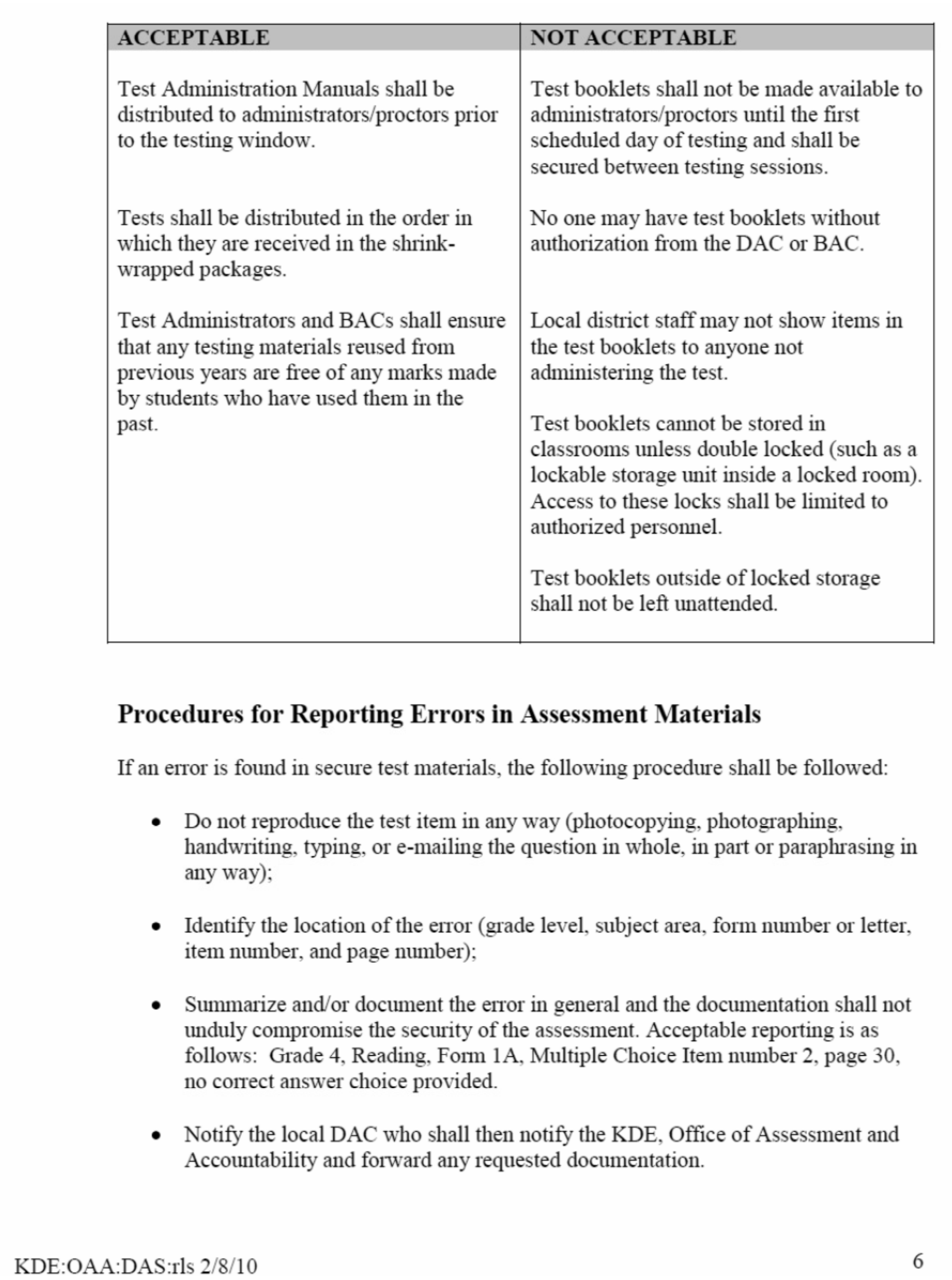
Test Security

ACCEPTABLE	NOT ACCEPTABLE
It is appropriate for teachers to know the concepts measured by the statewide assessment and to teach those concepts.	Proctors with knowledge of the content of any secure test item shall not reveal this content to anyone.



ACCEPTABLE	NOT ACCEPTABLE
Concepts appropriate for curriculum instruction can be found in Kentucky's Core Content for Assessment.	Teachers or other staff, who become aware of specific test items through any means, shall not use this knowledge to prepare students for the assessment.
Teachers may use test items from previous years released by the KDE to help prepare their students for the assessment.	No deliberate reviewing or reading of test items by an individual or group is permitted.
Noncertified persons helping with testing (packing materials, providing accommodations, escorting students to test sites) must sign a nondisclosure form.	No one shall take notes about or discuss the content, concepts or structure of any secure test item.
Students using technology to respond to test items are allowed to save responses to CDs or portable drives, but not to hard drives or servers.	Electronic or other versions of secure assessment materials or student responses shall not be maintained in the district.
Alert papers (i.e., evidence within a student response that the student may cause harm to self or to others or may otherwise be suffering abuses) may be copied only by the DAC, BAC, or school administrator. In this case these local district staff may photocopy the pertinent section of the student response and turn those pages over to the appropriate local authorities to assure the safety of the child and the community. The local district shall direct all local authorities that the student response may contain information related to secure test items. The local authorities shall sign a nondisclosure form.	Secure test materials shall not be reproduced in whole, in part or paraphrased in any way. Examples include: discussing, e-mailing, photocopying, photographing, handwriting, or typing. Electronic devices with wireless communication or imaging capabilities (e.g. cell phones or cameras) shall not be accessible by students during the testing sessions.
Test Administrators shall destroy any notes, rough drafts or scratch paper produced by students during testing immediately after each testing session or at the end of the testing day, ensuring that no test item is compromised.	Scoring of test items or rough drafts is not permissible.
Scanning student response booklets/answer sheets for stray marks and good faith effort is permissible.	Student responses shall not be read in their entirety as part of scanning for good faith effort checklists.

KDE:OAA:DAS:rls 2/8/10 5

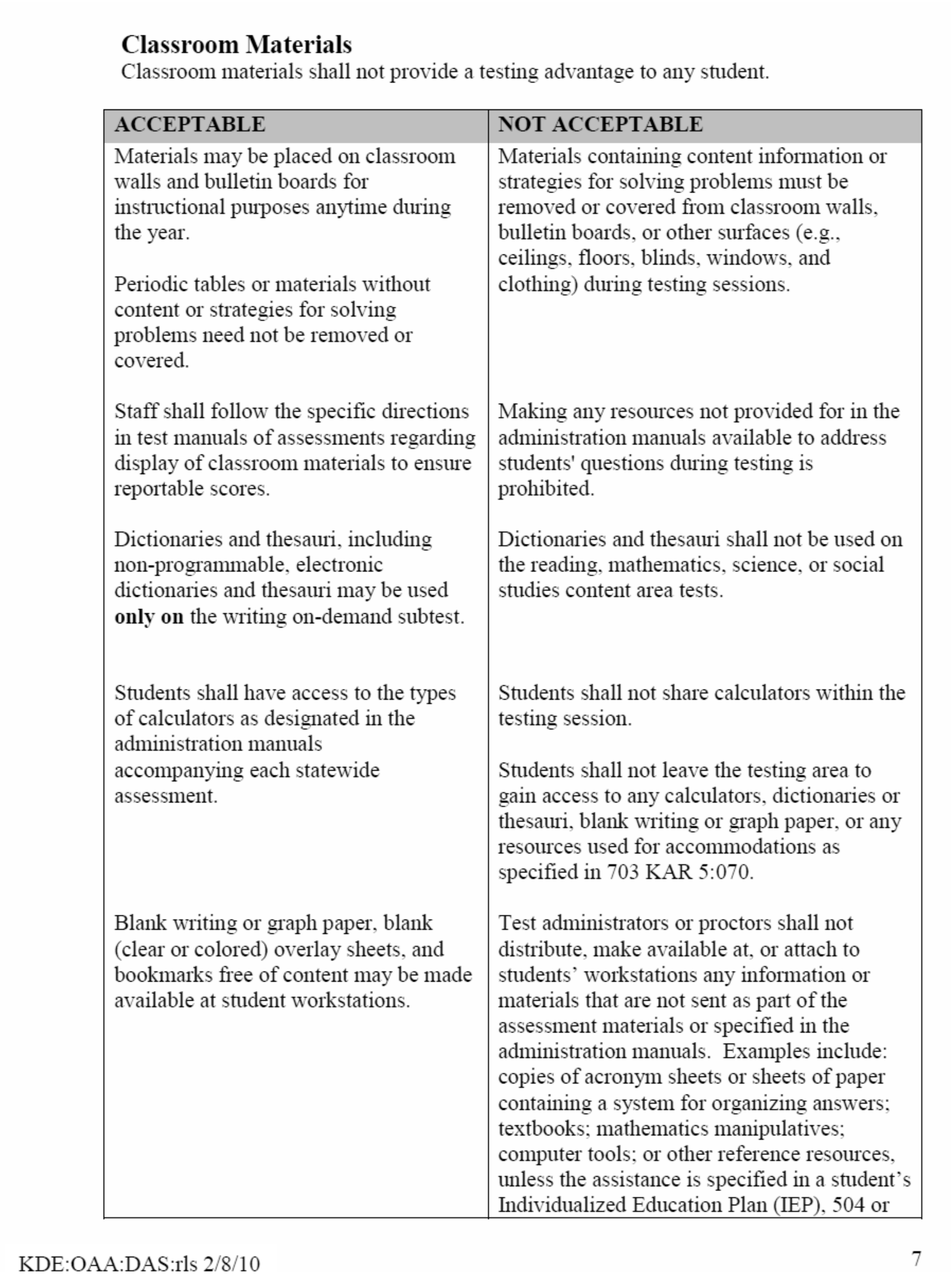


ACCEPTABLE	NOT ACCEPTABLE
Test Administration Manuals shall be distributed to administrators/proctors prior to the testing window.	Test booklets shall not be made available to administrators/proctors until the first scheduled day of testing and shall be secured between testing sessions.
Tests shall be distributed in the order in which they are received in the shrink-wrapped packages.	No one may have test booklets without authorization from the DAC or BAC.
Test Administrators and BACs shall ensure that any testing materials reused from previous years are free of any marks made by students who have used them in the past.	Local district staff may not show items in the test booklets to anyone not administering the test.
	Test booklets cannot be stored in classrooms unless double locked (such as a lockable storage unit inside a locked room). Access to these locks shall be limited to authorized personnel.
	Test booklets outside of locked storage shall not be left unattended.

Procedures for Reporting Errors in Assessment Materials

If an error is found in secure test materials, the following procedure shall be followed:

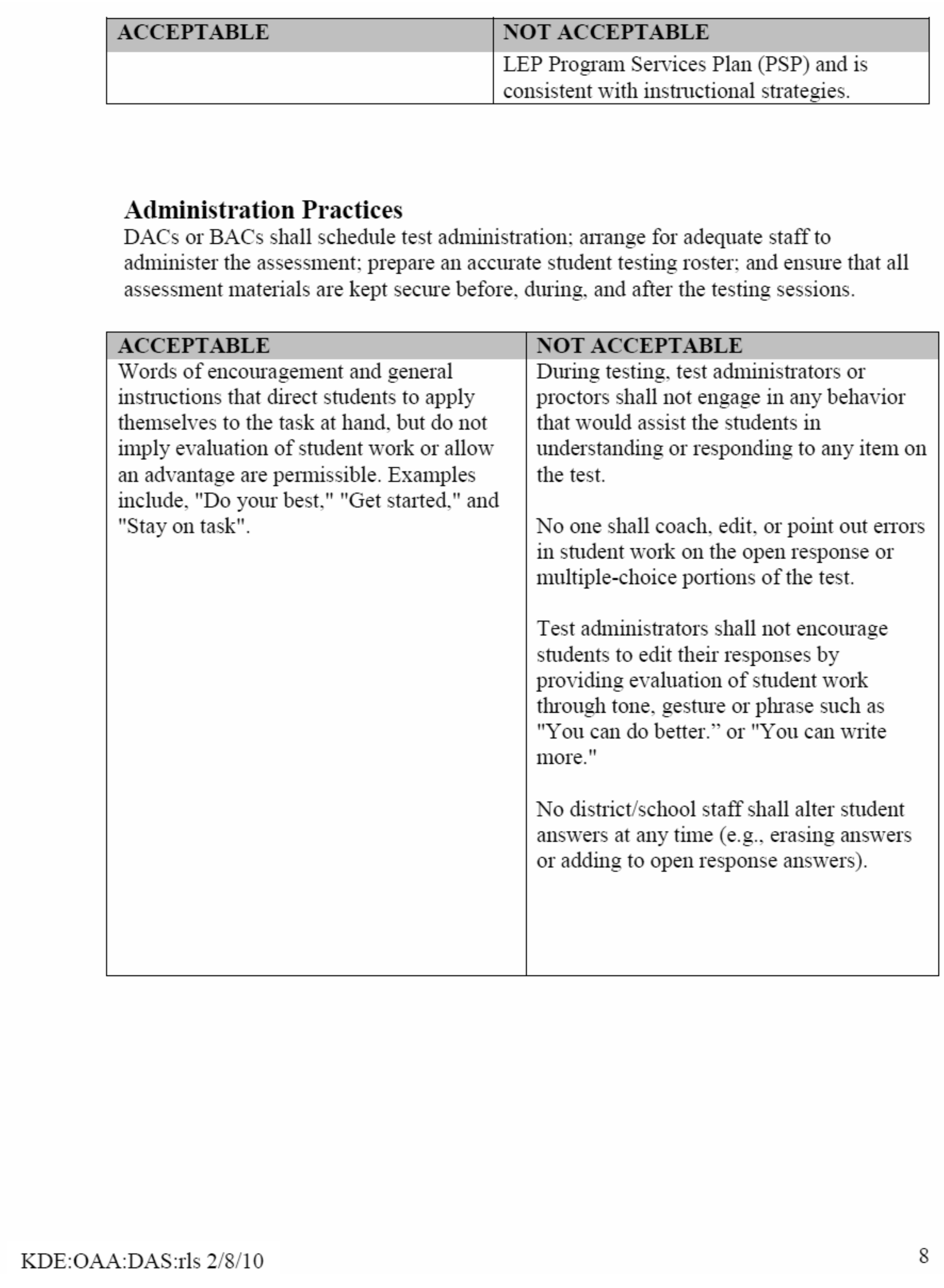
- Do not reproduce the test item in any way (photocopying, photographing, handwriting, typing, or e-mailing the question in whole, in part or paraphrasing in any way);
- Identify the location of the error (grade level, subject area, form number or letter, item number, and page number);
- Summarize and/or document the error in general and the documentation shall not unduly compromise the security of the assessment. Acceptable reporting is as follows: Grade 4, Reading, Form 1A, Multiple Choice Item number 2, page 30, no correct answer choice provided.
- Notify the local DAC who shall then notify the KDE, Office of Assessment and Accountability and forward any requested documentation.



Classroom Materials

Classroom materials shall not provide a testing advantage to any student.

ACCEPTABLE	NOT ACCEPTABLE
Materials may be placed on classroom walls and bulletin boards for instructional purposes anytime during the year. Periodic tables or materials without content or strategies for solving problems need not be removed or covered.	Materials containing content information or strategies for solving problems must be removed or covered from classroom walls, bulletin boards, or other surfaces (e.g., ceilings, floors, blinds, windows, and clothing) during testing sessions.
Staff shall follow the specific directions in test manuals of assessments regarding display of classroom materials to ensure reportable scores.	Making any resources not provided for in the administration manuals available to address students' questions during testing is prohibited.
Dictionaries and thesauri, including non-programmable, electronic dictionaries and thesauri may be used **only on** the writing on-demand subtest.	Dictionaries and thesauri shall not be used on the reading, mathematics, science, or social studies content area tests.
Students shall have access to the types of calculators as designated in the administration manuals accompanying each statewide assessment.	Students shall not share calculators within the testing session. Students shall not leave the testing area to gain access to any calculators, dictionaries or thesauri, blank writing or graph paper, or any resources used for accommodations as specified in 703 KAR 5:070.
Blank writing or graph paper, blank (clear or colored) overlay sheets, and bookmarks free of content may be made available at student workstations.	Test administrators or proctors shall not distribute, make available at, or attach to students' workstations any information or materials that are not sent as part of the assessment materials or specified in the administration manuals. Examples include: copies of acronym sheets or sheets of paper containing a system for organizing answers; textbooks; mathematics manipulatives; computer tools; or other reference resources, unless the assistance is specified in a student's Individualized Education Plan (IEP), 504 or

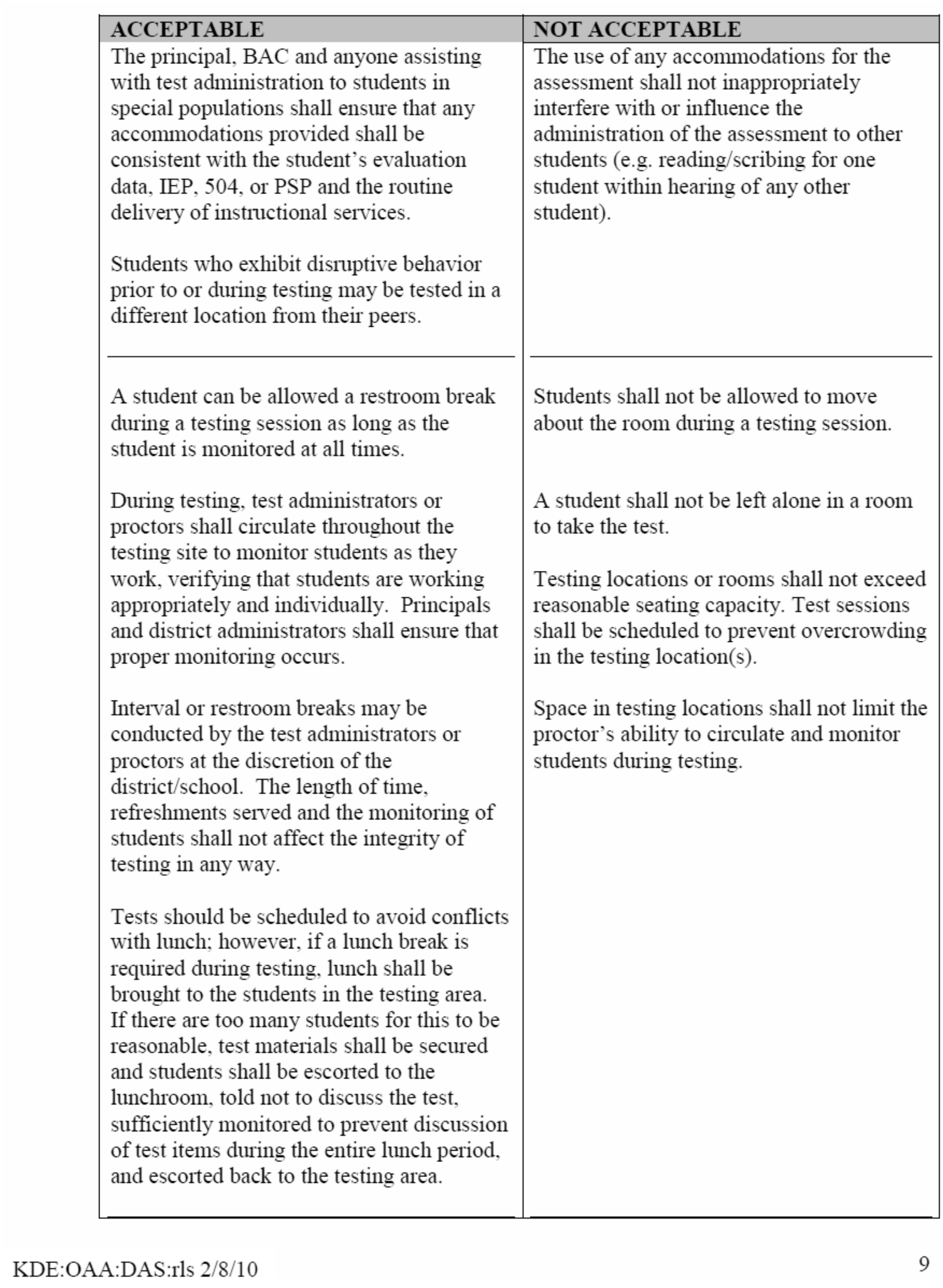


ACCEPTABLE	NOT ACCEPTABLE
	LEP Program Services Plan (PSP) and is consistent with instructional strategies.

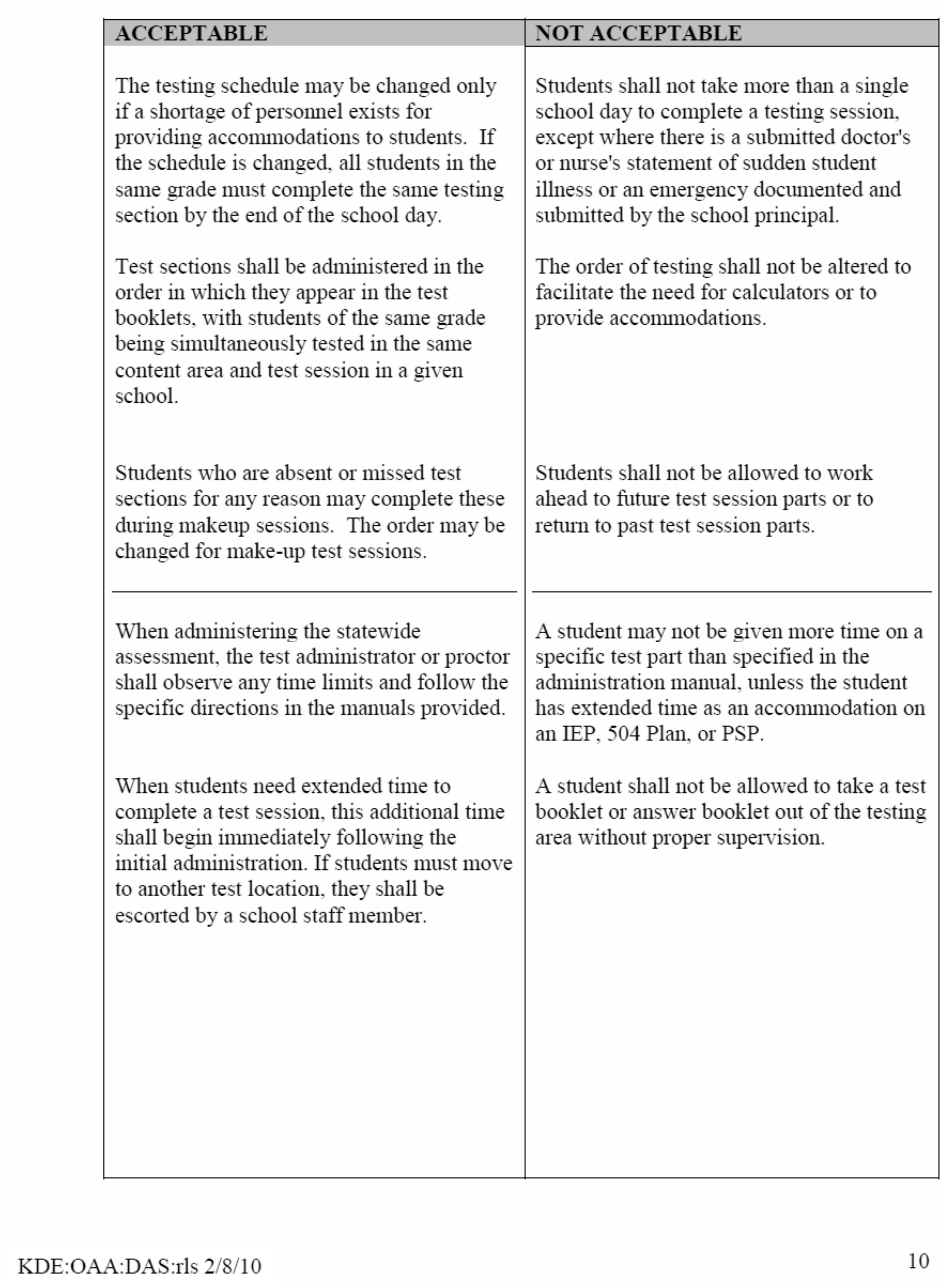
Administration Practices

DACs or BACs shall schedule test administration; arrange for adequate staff to administer the assessment; prepare an accurate student testing roster; and ensure that all assessment materials are kept secure before, during, and after the testing sessions.

ACCEPTABLE	NOT ACCEPTABLE
Words of encouragement and general instructions that direct students to apply themselves to the task at hand, but do not imply evaluation of student work or allow an advantage are permissible. Examples include, "Do your best," "Get started," and "Stay on task".	During testing, test administrators or proctors shall not engage in any behavior that would assist the students in understanding or responding to any item on the test. No one shall coach, edit, or point out errors in student work on the open response or multiple-choice portions of the test. Test administrators shall not encourage students to edit their responses by providing evaluation of student work through tone, gesture or phrase such as "You can do better." or "You can write more." No district/school staff shall alter student answers at any time (e.g., erasing answers or adding to open response answers).



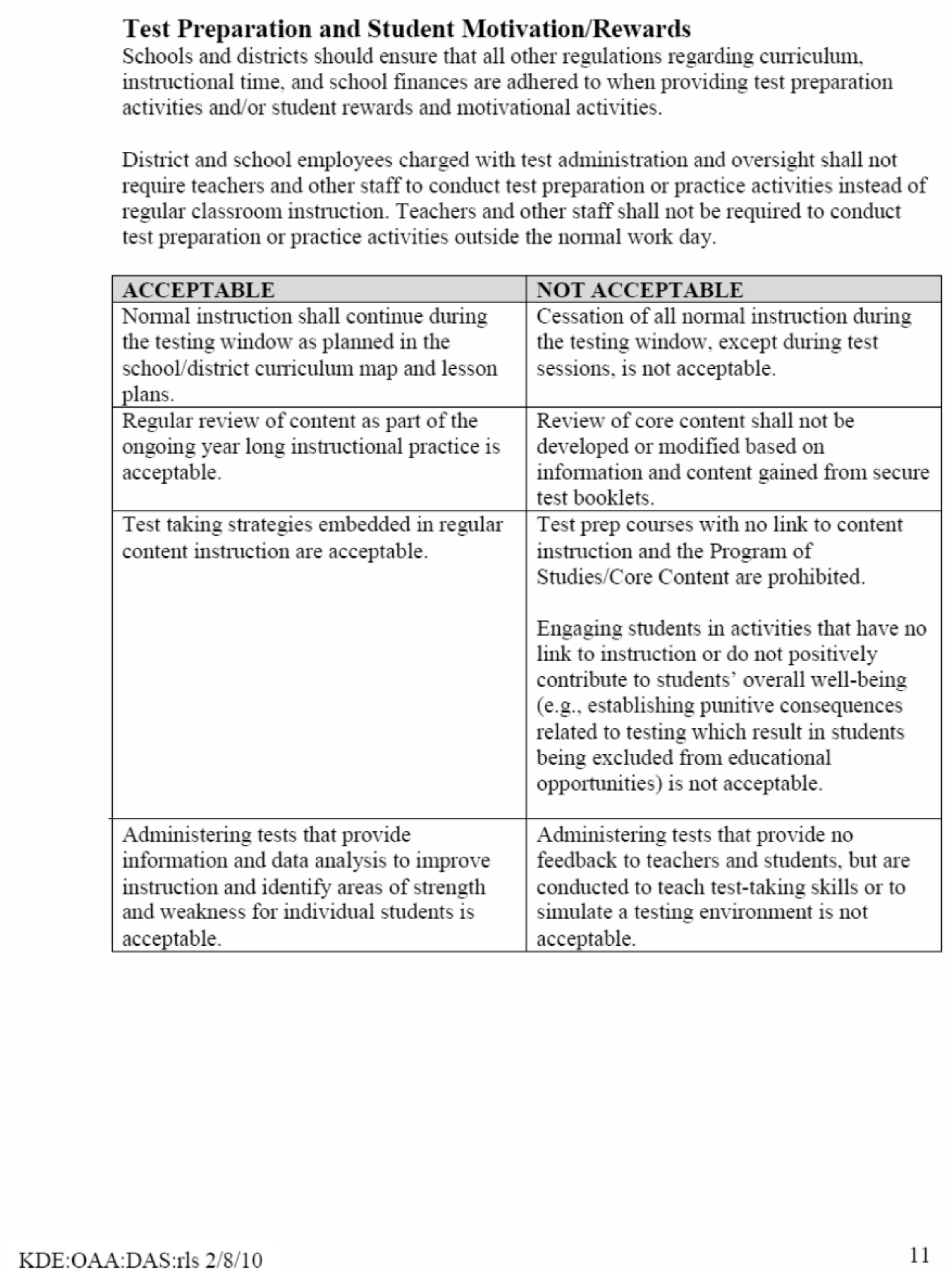
ACCEPTABLE	NOT ACCEPTABLE
The principal, BAC and anyone assisting with test administration to students in special populations shall ensure that any accommodations provided shall be consistent with the student's evaluation data, IEP, 504, or PSP and the routine delivery of instructional services. Students who exhibit disruptive behavior prior to or during testing may be tested in a different location from their peers.	The use of any accommodations for the assessment shall not inappropriately interfere with or influence the administration of the assessment to other students (e.g. reading/scribing for one student within hearing of any other student).
A student can be allowed a restroom break during a testing session as long as the student is monitored at all times. During testing, test administrators or proctors shall circulate throughout the testing site to monitor students as they work, verifying that students are working appropriately and individually. Principals and district administrators shall ensure that proper monitoring occurs. Interval or restroom breaks may be conducted by the test administrators or proctors at the discretion of the district/school. The length of time, refreshments served and the monitoring of students shall not affect the integrity of testing in any way. Tests should be scheduled to avoid conflicts with lunch; however, if a lunch break is required during testing, lunch shall be brought to the students in the testing area. If there are too many students for this to be reasonable, test materials shall be secured and students shall be escorted to the lunchroom, told not to discuss the test, sufficiently monitored to prevent discussion of test items during the entire lunch period, and escorted back to the testing area.	Students shall not be allowed to move about the room during a testing session. A student shall not be left alone in a room to take the test. Testing locations or rooms shall not exceed reasonable seating capacity. Test sessions shall be scheduled to prevent overcrowding in the testing location(s). Space in testing locations shall not limit the proctor's ability to circulate and monitor students during testing.



ACCEPTABLE	NOT ACCEPTABLE
The testing schedule may be changed only if a shortage of personnel exists for providing accommodations to students. If the schedule is changed, all students in the same grade must complete the same testing section by the end of the school day.	Students shall not take more than a single school day to complete a testing session, except where there is a submitted doctor's or nurse's statement of sudden student illness or an emergency documented and submitted by the school principal.
Test sections shall be administered in the order in which they appear in the test booklets, with students of the same grade being simultaneously tested in the same content area and test session in a given school.	The order of testing shall not be altered to facilitate the need for calculators or to provide accommodations.
Students who are absent or missed test sections for any reason may complete these during makeup sessions. The order may be changed for make-up test sessions.	Students shall not be allowed to work ahead to future test session parts or to return to past test session parts.
When administering the statewide assessment, the test administrator or proctor shall observe any time limits and follow the specific directions in the manuals provided.	A student may not be given more time on a specific test part than specified in the administration manual, unless the student has extended time as an accommodation on an IEP, 504 Plan, or PSP.
When students need extended time to complete a test session, this additional time shall begin immediately following the initial administration. If students must move to another test location, they shall be escorted by a school staff member.	A student shall not be allowed to take a test booklet or answer booklet out of the testing area without proper supervision.

KDE:OAA:DAS:rls 2/8/10

10

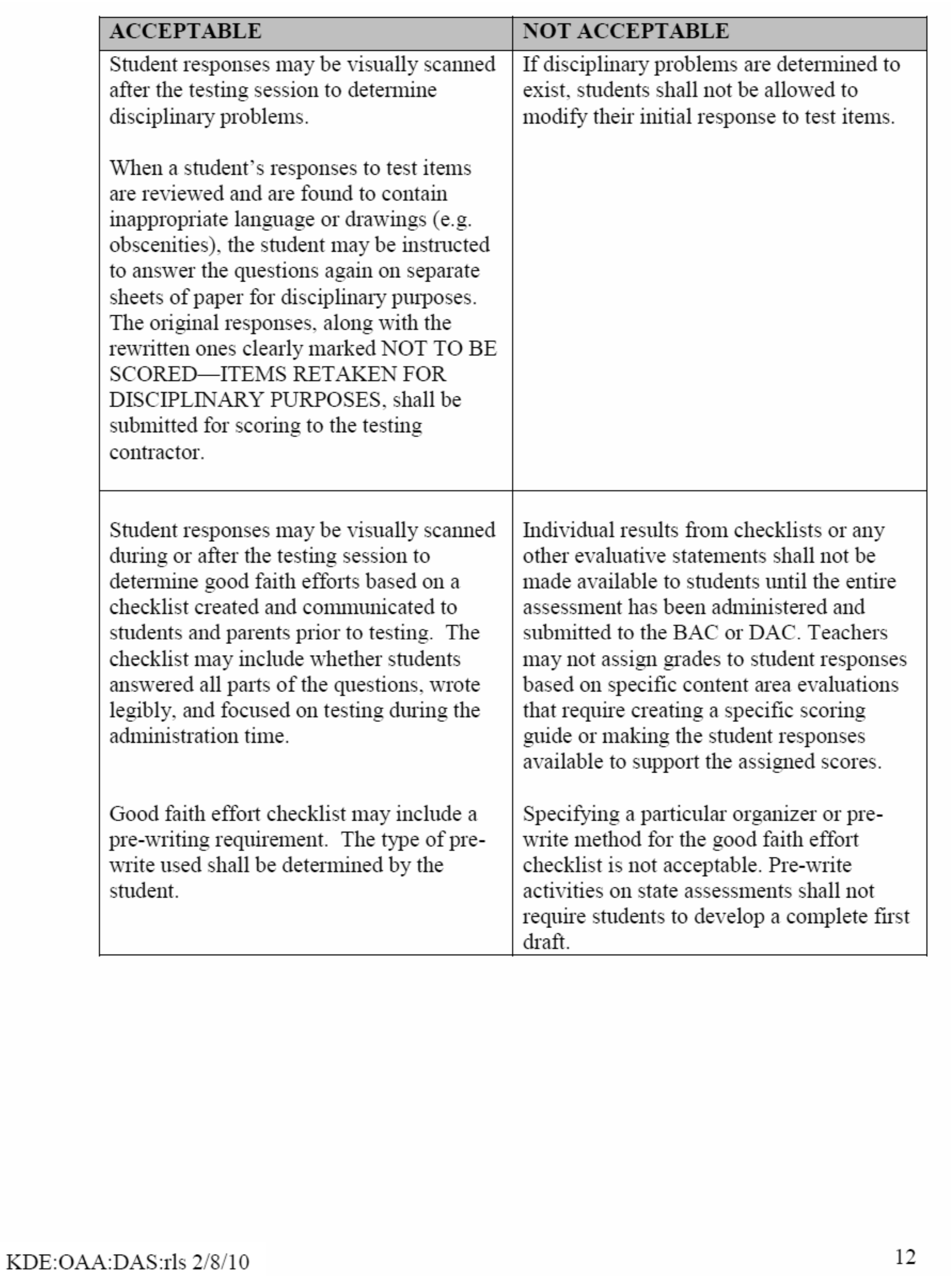


Test Preparation and Student Motivation/Rewards

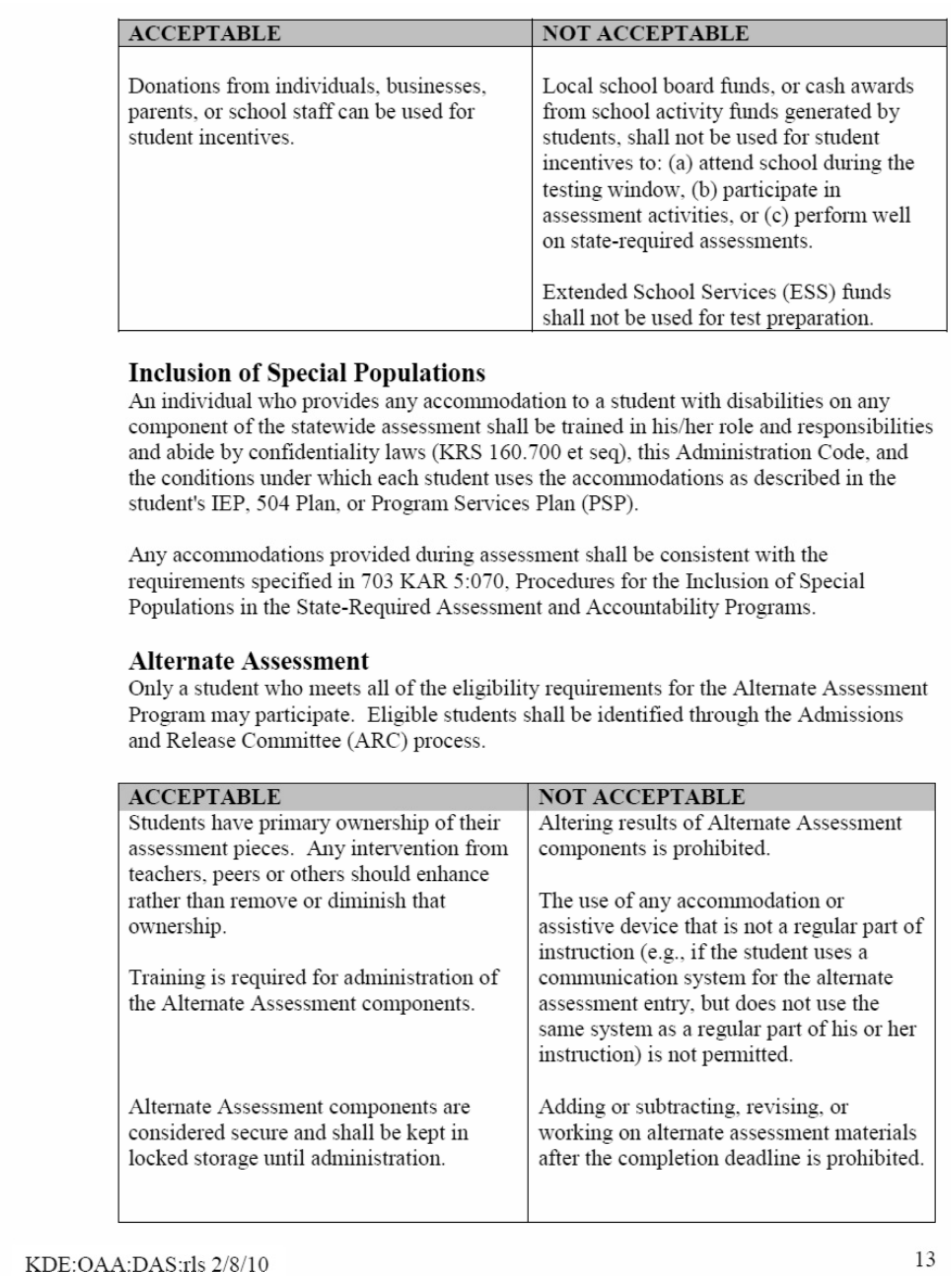
Schools and districts should ensure that all other regulations regarding curriculum, instructional time, and school finances are adhered to when providing test preparation activities and/or student rewards and motivational activities.

District and school employees charged with test administration and oversight shall not require teachers and other staff to conduct test preparation or practice activities instead of regular classroom instruction. Teachers and other staff shall not be required to conduct test preparation or practice activities outside the normal work day.

ACCEPTABLE	NOT ACCEPTABLE
Normal instruction shall continue during the testing window as planned in the school/district curriculum map and lesson plans.	Cessation of all normal instruction during the testing window, except during test sessions, is not acceptable.
Regular review of content as part of the ongoing year long instructional practice is acceptable.	Review of core content shall not be developed or modified based on information and content gained from secure test booklets.
Test taking strategies embedded in regular content instruction are acceptable.	Test prep courses with no link to content instruction and the Program of Studies/Core Content are prohibited. Engaging students in activities that have no link to instruction or do not positively contribute to students' overall well-being (e.g., establishing punitive consequences related to testing which result in students being excluded from educational opportunities) is not acceptable.
Administering tests that provide information and data analysis to improve instruction and identify areas of strength and weakness for individual students is acceptable.	Administering tests that provide no feedback to teachers and students, but are conducted to teach test-taking skills or to simulate a testing environment is not acceptable.



ACCEPTABLE	NOT ACCEPTABLE
Student responses may be visually scanned after the testing session to determine disciplinary problems. When a student's responses to test items are reviewed and are found to contain inappropriate language or drawings (e.g. obscenities), the student may be instructed to answer the questions again on separate sheets of paper for disciplinary purposes. The original responses, along with the rewritten ones clearly marked NOT TO BE SCORED—ITEMS RETAKEN FOR DISCIPLINARY PURPOSES, shall be submitted for scoring to the testing contractor.	If disciplinary problems are determined to exist, students shall not be allowed to modify their initial response to test items.
Student responses may be visually scanned during or after the testing session to determine good faith efforts based on a checklist created and communicated to students and parents prior to testing. The checklist may include whether students answered all parts of the questions, wrote legibly, and focused on testing during the administration time. Good faith effort checklist may include a pre-writing requirement. The type of pre-write used shall be determined by the student.	Individual results from checklists or any other evaluative statements shall not be made available to students until the entire assessment has been administered and submitted to the BAC or DAC. Teachers may not assign grades to student responses based on specific content area evaluations that require creating a specific scoring guide or making the student responses available to support the assigned scores. Specifying a particular organizer or pre-write method for the good faith effort checklist is not acceptable. Pre-write activities on state assessments shall not require students to develop a complete first draft.



ACCEPTABLE	NOT ACCEPTABLE
Donations from individuals, businesses, parents, or school staff can be used for student incentives.	Local school board funds, or cash awards from school activity funds generated by students, shall not be used for student incentives to: (a) attend school during the testing window, (b) participate in assessment activities, or (c) perform well on state-required assessments. Extended School Services (ESS) funds shall not be used for test preparation.

Inclusion of Special Populations

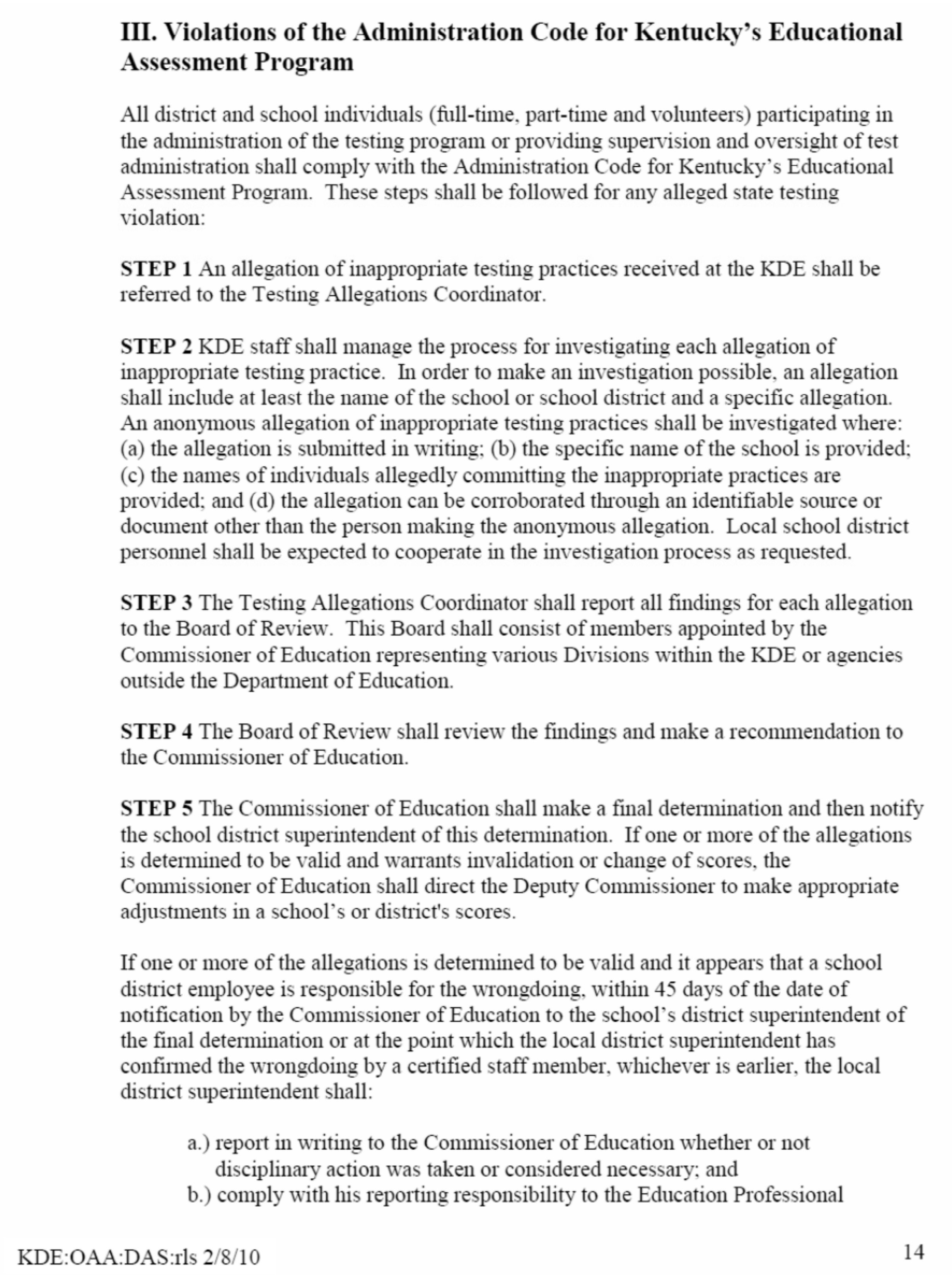
An individual who provides any accommodation to a student with disabilities on any component of the statewide assessment shall be trained in his/her role and responsibilities and abide by confidentiality laws (KRS 160.700 et seq), this Administration Code, and the conditions under which each student uses the accommodations as described in the student's IEP, 504 Plan, or Program Services Plan (PSP).

Any accommodations provided during assessment shall be consistent with the requirements specified in 703 KAR 5:070, Procedures for the Inclusion of Special Populations in the State-Required Assessment and Accountability Programs.

Alternate Assessment

Only a student who meets all of the eligibility requirements for the Alternate Assessment Program may participate. Eligible students shall be identified through the Admissions and Release Committee (ARC) process.

ACCEPTABLE	NOT ACCEPTABLE
Students have primary ownership of their assessment pieces. Any intervention from teachers, peers or others should enhance rather than remove or diminish that ownership. Training is required for administration of the Alternate Assessment components.	Altering results of Alternate Assessment components is prohibited. The use of any accommodation or assistive device that is not a regular part of instruction (e.g., if the student uses a communication system for the alternate assessment entry, but does not use the same system as a regular part of his or her instruction) is not permitted.
Alternate Assessment components are considered secure and shall be kept in locked storage until administration.	Adding or subtracting, revising, or working on alternate assessment materials after the completion deadline is prohibited.



III. Violations of the Administration Code for Kentucky's Educational Assessment Program

All district and school individuals (full-time, part-time and volunteers) participating in the administration of the testing program or providing supervision and oversight of test administration shall comply with the Administration Code for Kentucky's Educational Assessment Program. These steps shall be followed for any alleged state testing violation:

STEP 1 An allegation of inappropriate testing practices received at the KDE shall be referred to the Testing Allegations Coordinator.

STEP 2 KDE staff shall manage the process for investigating each allegation of inappropriate testing practice. In order to make an investigation possible, an allegation shall include at least the name of the school or school district and a specific allegation. An anonymous allegation of inappropriate testing practices shall be investigated where: (a) the allegation is submitted in writing; (b) the specific name of the school is provided; (c) the names of individuals allegedly committing the inappropriate practices are provided; and (d) the allegation can be corroborated through an identifiable source or document other than the person making the anonymous allegation. Local school district personnel shall be expected to cooperate in the investigation process as requested.

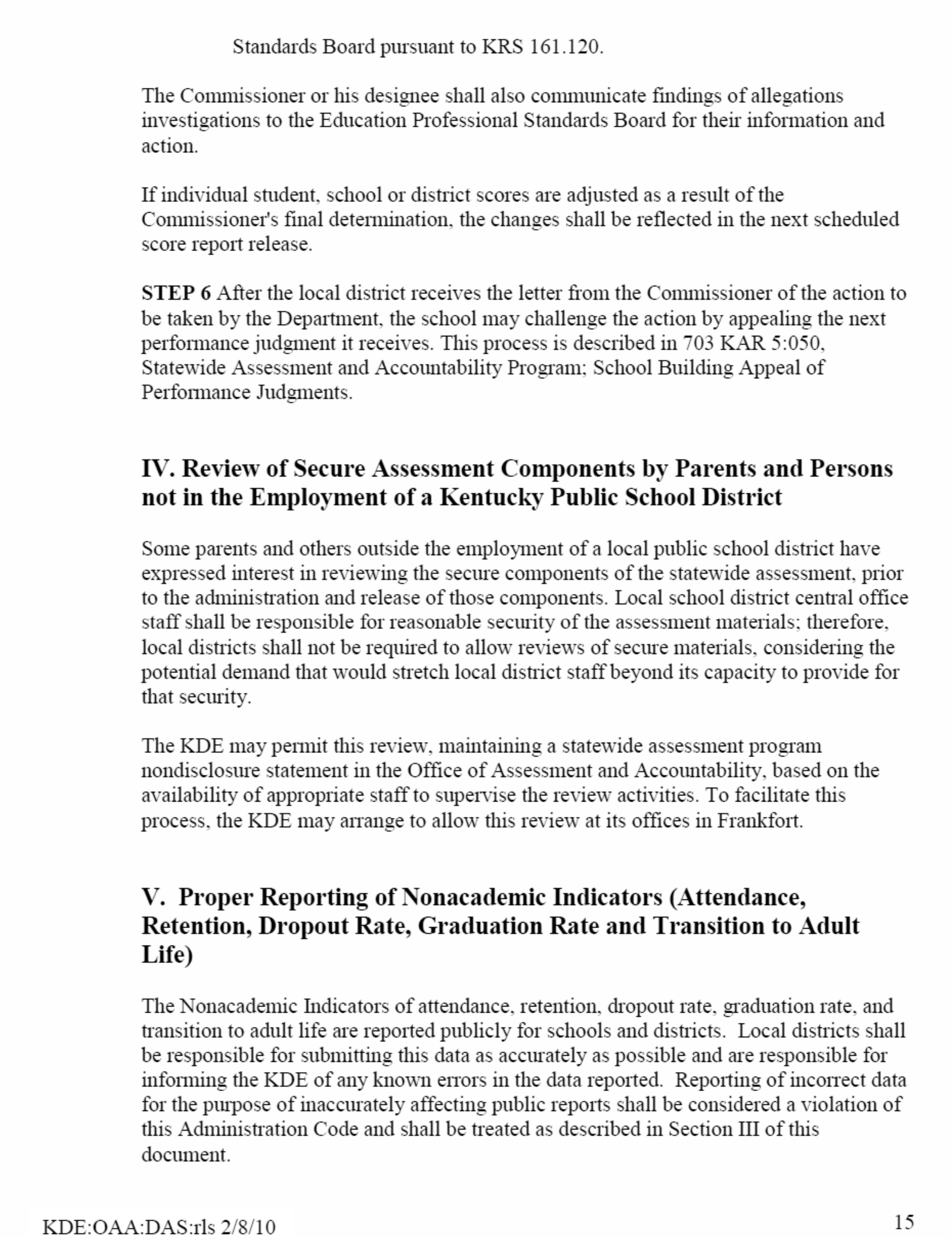
STEP 3 The Testing Allegations Coordinator shall report all findings for each allegation to the Board of Review. This Board shall consist of members appointed by the Commissioner of Education representing various Divisions within the KDE or agencies outside the Department of Education.

STEP 4 The Board of Review shall review the findings and make a recommendation to the Commissioner of Education.

STEP 5 The Commissioner of Education shall make a final determination and then notify the school district superintendent of this determination. If one or more of the allegations is determined to be valid and warrants invalidation or change of scores, the Commissioner of Education shall direct the Deputy Commissioner to make appropriate adjustments in a school's or district's scores.

If one or more of the allegations is determined to be valid and it appears that a school district employee is responsible for the wrongdoing, within 45 days of the date of notification by the Commissioner of Education to the school's district superintendent of the final determination or at the point which the local district superintendent has confirmed the wrongdoing by a certified staff member, whichever is earlier, the local district superintendent shall:

a.) report in writing to the Commissioner of Education whether or not disciplinary action was taken or considered necessary; and
b.) comply with his reporting responsibility to the Education Professional



Standards Board pursuant to KRS 161.120.

The Commissioner or his designee shall also communicate findings of allegations investigations to the Education Professional Standards Board for their information and action.

If individual student, school or district scores are adjusted as a result of the Commissioner's final determination, the changes shall be reflected in the next scheduled score report release.

STEP 6 After the local district receives the letter from the Commissioner of the action to be taken by the Department, the school may challenge the action by appealing the next performance judgment it receives. This process is described in 703 KAR 5:050, Statewide Assessment and Accountability Program; School Building Appeal of Performance Judgments.

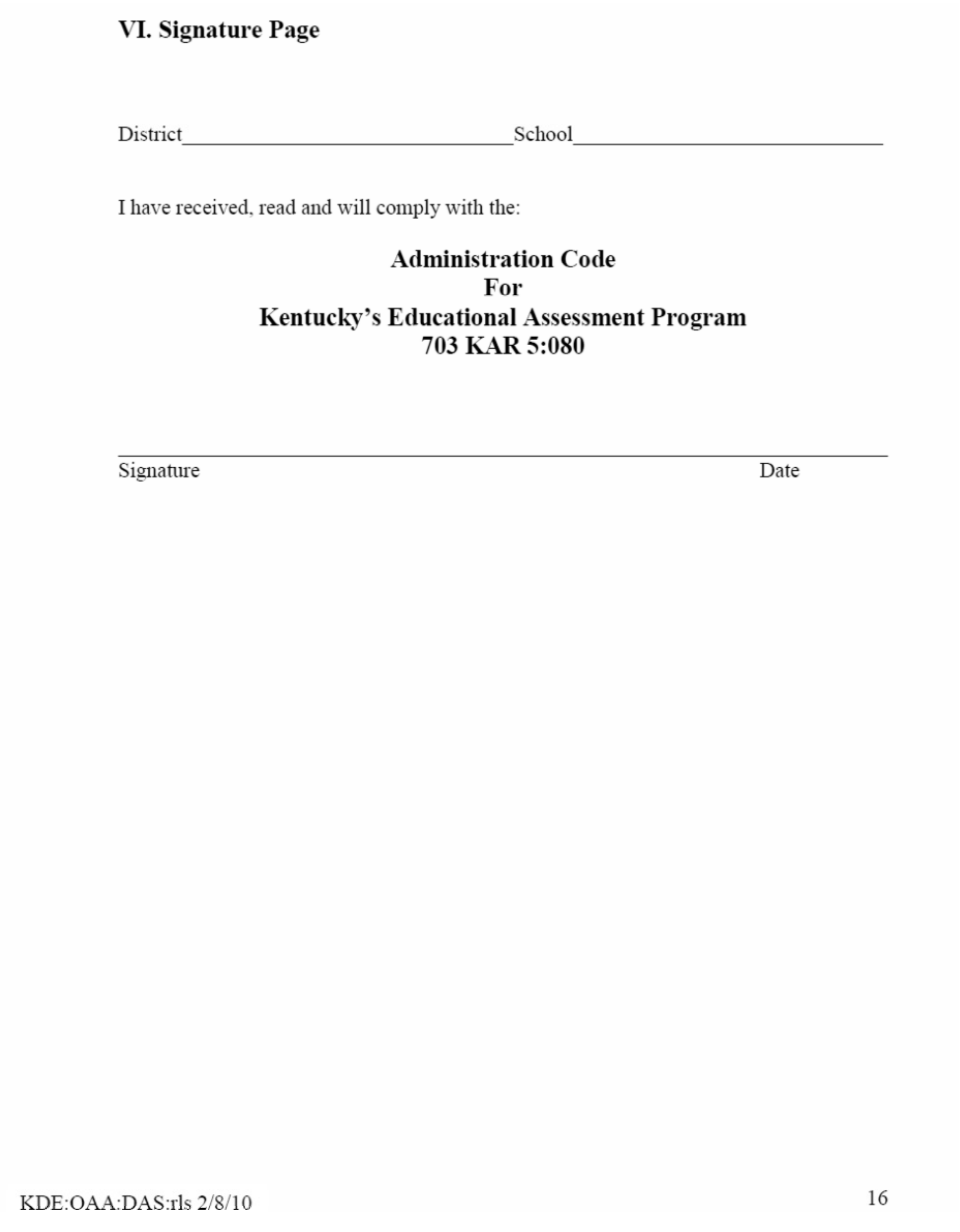
IV. Review of Secure Assessment Components by Parents and Persons not in the Employment of a Kentucky Public School District

Some parents and others outside the employment of a local public school district have expressed interest in reviewing the secure components of the statewide assessment, prior to the administration and release of those components. Local school district central office staff shall be responsible for reasonable security of the assessment materials; therefore, local districts shall not be required to allow reviews of secure materials, considering the potential demand that would stretch local district staff beyond its capacity to provide for that security.

The KDE may permit this review, maintaining a statewide assessment program nondisclosure statement in the Office of Assessment and Accountability, based on the availability of appropriate staff to supervise the review activities. To facilitate this process, the KDE may arrange to allow this review at its offices in Frankfort.

V. Proper Reporting of Nonacademic Indicators (Attendance, Retention, Dropout Rate, Graduation Rate and Transition to Adult Life)

The Nonacademic Indicators of attendance, retention, dropout rate, graduation rate, and transition to adult life are reported publicly for schools and districts. Local districts shall be responsible for submitting this data as accurately as possible and are responsible for informing the KDE of any known errors in the data reported. Reporting of incorrect data for the purpose of inaccurately affecting public reports shall be considered a violation of this Administration Code and shall be treated as described in Section III of this document.



VI. Signature Page

District______________________________School____________________________

I have received, read and will comply with the:

Administration Code
For
Kentucky's Educational Assessment Program
703 KAR 5:080

____
Signature Date

Appendix D

Outlier Identification Methodology

The methodology for identifying outliers was prescribed in "Robust Outlier Identification Using SAS*"* (Shoemaker). Complementary outlier identification methods were described in *Data Analysis with Microsoft Excel* (Berk 154-5). Both texts list methods originated by John Tukey in an effort to employ quartiles rather than variance to minimize the impact of extreme values.

The methods referenced above were applied to between-year difference scores. Difference scores were calculated by subtracting the previous year's score from the current year's score. School size was determined by establishing each school's percentile rank based on the number of students tested. A percentile is the proportion of total data points that are at or below a given data point (Heiman). Tiny schools were defined as those that fell at or below the 25th percentile, small schools fell between the 26th and 50th percentiles, medium schools were those between the 51st and 75th percentiles, and large schools were at or above the 76th percentile based on number of students tested. While the Educational Planning and Assessment System (EPAS) routinely generates composite scores from which it was easy to derive difference scores, the Kentucky Core Content Test (KCCT) system does not generate composite scores. Difference scores for the KCCT were generated by subtracting the raw 2009 KCCT multiple-choice average from the 2010 multiple-choice average. In order to analyze data for outliers, education assessment data for each respective instrument was linked by school number in order to establish prior and current year score columns. Schools in which no students tested or for which no scores were available were excluded from the research pool.

Once difference scores, composite scores, and size category fields were populated for each school, the respective score distributions for each size category were computed, and outlier parameters were established. Difference scores that were greater than the third quartile plus 1.5 times the interquartile range (or IQR, which is the difference between the third and first quartiles) or less than the first quartile minus 1.5 times the IQR were identified as outliers for each respective size category (Shoemaker).

Office of Education Accountability

Legislative Research Commission
Frankfort, Kentucky
lrc.ky.gov