Wyoming Department of Education Controls Over State Assessment Scoring

FINAL AUDIT REPORT

ED-OIG/A09I0012
July 2009

Our mission is to promote the efficiency, effectiveness, and integrity of the Department's programs and operations.

U.S. Department of Education
Office of Inspector General
Sacramento, CA

NOTICE

Statements that managerial practices need improvements, as well as other conclusions and recommendations in this report represent the opinions of the Office of Inspector General. Determinations of corrective action to be taken will be made by the appropriate Department of Education officials.

In accordance with the Freedom of Information Act (5 U.S.C. § 552), reports issued by the Office of Inspector General are available to members of the press and general public to the extent information contained therein is not subject to exemptions in the Act.

UNITED STATES DEPARTMENT OF EDUCATION
OFFICE OF INSPECTOR GENERAL

Audit Services
Sacramento Region

July 10, 2009

Dr. James M. McBride
State Superintendent of Public Instruction
Wyoming Department of Education
2300 Capitol Avenue
Hathaway Building, 2nd Floor
Cheyenne, WY 82002-0050

Dear Dr. McBride,

Enclosed is our final audit report, Control Number ED-OIG/A09I0012, entitled *Wyoming Department of Education Controls Over State Assessment Scoring.* This report incorporates the comments you provided in response to the draft report. If you have any additional comments or information that you believe may have a bearing on the resolution of this audit, you should send them directly to the following Education Department official, who will consider them before taking final Departmental action on this audit:

Joseph Conaty
Executive Administrator Delegated the Authority to Perform
the Functions and Duties of the Assistant Secretary for
Office of Elementary and Secondary Education
U.S. Department of Education
400 Maryland Avenue SW, Room 3E314
Washington, D.C. 20202

It is the policy of the U. S. Department of Education to expedite the resolution of audits by initiating timely action on the findings and recommendations contained therein. Therefore, receipt of your comments within 30 days would be appreciated.

In accordance with the Freedom of Information Act (5 U.S.C. § 552), reports issued by the Office of Inspector General are available to members of the press and general public to the extent information contained therein is not subject to exemptions in the Act.

Sincerely,

/s/

Raymond Hendren
Regional Inspector General for Audit

Enclosures

Acronyms/Abbreviations Used in this Report

ARRA	American Recovery and Reinvestment Act of 2009
AYP	Adequate Yearly Progress
CCSSO	Council of Chief State School Officers
ELA	English Language Arts
ELC	Education Leaders Council
ESEA	Elementary and Secondary Education Act of 1965, as amended by the No Child Left Behind Act of 2001
FERPA	Family Educational Rights and Privacy Act
LEA	Local Educational Agency (includes school districts)
OESE	Office of Elementary and Secondary Education (U.S. Department of Education)
PAWS	Proficiency Assessments for Wyoming Students
SEA	State Educational Agency
TAC	Technical Advisory Committee
WDE	Wyoming Department of Education
WISE	Wyoming Integrated Statewide Education (information system)

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The objective of our audit was to determine whether controls over scoring of assessments at the Wyoming Department of Education (WDE) were adequate to provide reasonable assurance that assessment results are reliable. Our review covered assessments administered in school year 2007-2008 used for evaluating individual student achievement and making adequate yearly progress (AYP) determinations under Section 1111(b)(3) of the Elementary and Secondary Education Act (ESEA), as amended by the No Child Left Behind Act of 2001. Congress recently reemphasized the importance of using academic assessments to measure student achievement in Section 14005(d)(4)(A) of the American Recovery and Reinvestment Act of 2009 (ARRA) by specifically targeting recovery funds toward "enhancing the quality of assessments" that States administer under the ESEA. ARRA funding will provide an opportunity for States to enhance the assessments that the President indicates in his education agenda will "be used to track student learning in a timely and individualized manner." The Proficiency Assessments for Wyoming Students (PAWS) was developed to meet ESEA requirements but is also used, consistent with the President's agenda, to support student learning and teaching in Wyoming public classrooms.

We evaluated the adequacy of WDE's controls over scoring the PAWS based on current uses, which includes assessing individual student achievement without linking results to high-stakes outcomes (graduation, grade promotion, college entrance, etc.) and for making statistically reliable adequate yearly progress (AYP) decisions about Wyoming schools, local educational agencies (LEAs), and WDE. We concluded that WDE implemented adequate controls over scoring. This provided reasonable assurance that reliable assessment results were used to determine AYP and to evaluate individual student achievement in school year 2007-2008.

WDE worked with its assessment contractor to implement a system of controls over scoring that included automated student assessment record tracking, multiple levels of data validation, and other quality control procedures that were important to obtaining reasonable assurance in the results. WDE also established a collaborative control environment with the contractor that encouraged frequent communication, fostered continuous improvement of existing processes, and ensured correction of problems once identified. However, we did identify controls over PAWS scoring that could be enhanced consistent with recommendations from the U.S. Department of Education, Office of Elementary and Secondary Education (OESE); Council of Chief State School Officers (CCSSO); and Education Leaders Council (ELC) to provide greater assurance in results. Specifically, we recommend that WDE document its existing procedures for monitoring contractor performance and for reviewing contractor-provided assessment results. We also recommend that WDE allow additional time for accountability staff to review contractor-provided assessment results and involve school personnel in the review process before results are published.

As noted in the Other Matters section of this report, errors have been publicly disclosed in each of the last 2 years (2006-2007 and 2007-2008) that the PAWS has been administered. For the one scoring error that occurred in 2007-2008, WDE determined that the error affected 37 of the more than 44,000 students assessed and did not impact AYP determinations. However, the error was publicly reported, and in conjunction with publicly reported errors from 2006-2007, may affect user confidence in the reliability of PAWS results.

We provided a draft report to WDE on May 18, 2009, for comment. WDE concurred with our findings in its comments to the draft report and described actions that have been taken or will be taken to implement each of our recommendations. WDE's comments on the draft report are summarized at the end of each finding and included in their entirety as Enclosure 2 to this report.

BACKGROUND

ESEA §1111(b)(3) requires each State to use a set of annual student academic assessments to determine whether the performance of the State educational agency (SEA), local educational agencies (LEAs), and schools meet the State's academic achievement standards. States must also use these assessments to determine whether individual students are meeting minimum State proficiency standards in mathematics, reading or language arts, and science. ESEA §1111(b)(3)(C)(iii) requires the assessments to be used for valid and reliable purposes consistent with relevant, nationally recognized professional and technical standards. Nationally recognized professional and technical standards are contained in the *Standards for Educational and Psychological Testing* (Standards) jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. ESEA §1111(b)(3)(C)(iv), requires SEAs to demonstrate to the U.S. Department of Education (Department) that their assessments adhere to technical requirements in the Standards for validity and reliability to use them for ESEA accountability purposes. In November 2007, the Assistant Secretary of Elementary and Secondary Education reported, based on external peer reviews and Department staff evaluations of WDE and contractor evidence, that the primary PAWS math, writing, and reading assessments met all ESEA requirements.[1]

The Standards differentiate between high- and low-stakes testing based on the importance of the results for individuals, organizations, and groups. According to the Standards--

> At the individual level, when significant educational paths or choices of an individual are directly affected by test performance, such as whether a student is promoted or retained at a grade level, graduated, or admitted or placed into a desired program, the test use is said to have high stakes.
> . . . Testing programs for institutions can have high stakes when aggregate performance of a sample or of the entire population of test takers is used to infer the quality of service provided, and decisions are made about institutional status, rewards, or sanctions based on the test results.
> . . . The higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality.

Thus, we concluded that calculating and reporting AYP for SEAs, LEAs, and schools would be a high-stakes use of ESEA assessments. However, using ESEA assessments to evaluate individual student achievement could be either high- or low-stakes depending on whether decisions, such as high-school graduation, were based on student performance on the assessments.

[1] WDE must submit final 2007-2008 PAWS participation data for the science exam to OESE to determine whether the science exam meets all ESEA requirements.

WDE State Assessments

WDE developed the PAWS to comply with assessment provisions added to the ESEA in 2002 with passage of the No Child Left Behind Act. In September 2004, WDE contracted with Harcourt Assessment (Harcourt) to create and score individual student assessments for school years 2005-2006 through 2007-2008. In January 2008, Pearson Education, Inc. (Pearson) purchased Harcourt and assumed responsibility for the contract. The term "contractor" as used in this report refers to Harcourt/Pearson. Of the $23.1 million paid to the contractor from February 2005 through October 2008 to develop and score the PAWS, about $8.5 million, or 37 percent, was paid using Federal funds.[2]

The PAWS administered in school year 2007-2008 included exams for reading, writing, and mathematics for all students in grades 3-8 and 11. Students in grades 4, 8, and 11 also took a science exam for the first time in 2007-2008. The reading, mathematics, and science exams were composed of multiple choice and constructed response items while the writing exam included only constructed response items. Constructed response items require a student to write in an answer that is later graded by a person rather than a machine. Multiple-choice questions on the PAWS exams (for every grade and subject area) were taken online at school sites and electronically scored.

Students in grades 3-8 entered their responses to constructed response items in student test booklets that were later imaged into an electronic system at the contractor's facility for subsequent scoring. Grade 11 exams were administered completely online for reading and writing, but responses to mathematics and science constructed response items were recorded in student test booklets similar to grade 3-8 exams.

Constructed response items were scored by contractor personnel at facilities in Texas and Ohio. Wyoming educators also graded constructed response items at a scoring institute held in Casper, Wyoming, that replicated the same training and grading processes used by the contractor at its facilities. The scoring institute provided educators a first-hand understanding of the PAWS grading process, which they could take back to their districts and schools to share with colleagues.

PAWS results are used by LEAs and schools to evaluate whether individual students have attained proficiency in meeting the State's reading, writing, mathematics, and science content standards. PAWS results are also used to improve teaching and learning in Wyoming classrooms. PAWS results are *not* used to make high-stakes decisions about individual students, such as advancement to the next grade or eligibility for a high school diploma.[3]

[2] ESEA Title VI, Part A, Improving Academic Achievement Grant; ESEA, Title III, Language Instruction for Limited English Proficient and Immigrant Students Grant; and Individuals with Disabilities Education Act (IDEA), Title I, Part B, Assistance for Education of All Children with Disabilities Grant.

[3] The *Standards for Educational and Psychological Testing* require evidence of higher technical quality for high-stakes assessments. "When the stakes for an individual are high, and important decisions depend substantially on test performance, the test needs to exhibit higher standards of technical quality for its avowed purposes than might be expected of tests used for lower-stakes purposes."

In contrast with the 23 States that require students to pass a single State assessment for graduation, Wyoming LEAs are required to use a "Body of Evidence" to determine whether a student has met graduation requirements. The Wyoming Body of Evidence system provides multiple measures to assess student mastery of the content standards and to determine whether individual students should graduate.[4] WDE conducts reviews of each LEA's Body of Evidence system which can include district assessments, common assessments across classrooms, course-embedded assessments, ratings of performance on projects, and successful completion of courses with passing grades. An LEA may include PAWS results as a component of its Body of Evidence system.

WDE also uses PAWS results to determine AYP for LEAs and schools. When an SEA identifies LEAs or schools that do not meet AYP performance requirements, the SEA is required to take specific actions to improve student academic achievement in those LEAs and schools.[5] Because consequences are tied to WDE AYP determinations, the PAWS is considered a "high-stakes" assessment for Wyoming LEAs and schools. The ESEA requires SEAs to use assessment results to calculate and report statistically reliable AYP determinations. As allowed by Federal regulation, WDE determines AYP reliability by (1) using confidence intervals to calculate assessment results for LEAs and schools within a margin of error, (2) averaging assessment results across multiple years, (3) employing minimum student counts in its evaluation of demographic subgroup performance, and (4) using safe-harbor calculations to monitor annual gains in schools needing improvement.

Elements of WDE's Scoring Process

Before any item was included on the operational PAWS exam, it was field-tested in a prior year to determine its suitability for inclusion on the actual exam. Contractor and WDE content experts and State educators reviewed the statistical indicators and content relevance for each field tested item. When a question was found to be valid for measuring student performance against State content standards, the question was included in the PAWS operational test bank for use in future years. WDE reviewed and approved all operational test items and test booklets before they were used.

WDE collected demographic data on every Wyoming student in October 2007 and March 2008 and tracked each student using a unique identifier. Unique identifiers were assigned using the Wyoming Integrated Statewide Education (WISE) system, which was also used to track

[4] The *Standards for Educational and Psychological Testing* state that using multiple measures increases assessment validity. "The validity of individual interpretations can be enhanced by taking into account other relevant information about individual students before making important decisions. It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions."

[5] Schools receiving Title I funding are identified for improvement after failing to make AYP for two consecutive years in the same content area. A Title I school in its first year of improvement must provide parents with the public school choice option, if available. If the Title I school identified for improvement again fails to make AYP (second year of improvement), it must also offer supplemental educational services to low-income students. If the Title I school identified for improvement fails to make AYP in the content area in subsequent years, the school must implement corrective action (third year of improvement), create a plan for restructuring (fourth year of improvement), and undergo restructuring (fifth year of improvement).

individual student achievement over time. Student records in WISE were validated against LEA records and then submitted to the contractor to create a Pre-ID data file that was used to create scannable test labels, organize test materials by school, and assign exam forms to the right students. Use of pre-assigned test materials, unique student identifiers, and validation processes enhance accountability over assessment materials, expedite the reporting process, and improve accuracy. The Pre-ID file was also used by the contractor to create individual user profiles for every student taking the PAWS.

The online portion of the PAWS was administered on a secure computer system that limited students to answering one item at a time, which reduced the impact of internet connectivity problems and concerns about exam security. All schools in Wyoming were equipped with high-speed internet access to allow for the effective delivery of online PAWS exams. All computer workstations used for the online assessment were set to allow only the assessment program to operate, which prevented students from accessing the internet or other resources on the computer for assistance. A key check report was created to validate the keys used to score the online multiple choice items after a sufficient number of students had taken the exam. The key check process was performed early enough to allow sufficient time for any necessary changes to be made. After the key check process was completed by the contractor, preliminary score reports were made available online to allow teachers access to student results for multiple choice items. The contractor operated a hotline for schools to report potential anomalies.

Paper assessment booklets containing constructed response items were sent to LEAs with a scannable Pre-ID label attached that included student name, unique student identifier, grade, school, district, gender, and date of birth. When the assessment booklets were received at the contractor site for scoring, they were electronically imaged into the grading system. The paper assessments were then matched and merged to each student's corresponding online testing record using a five-data-field validation process (unique identifier, first name, last name, date of birth, and gender) to ensure that results were correctly matched to each student. The merge process also included quality checks to ensure that raw scores were accurately applied for all constructed response items and that all multiple-choice items had a corresponding constructed response record when required.

Matching paper booklets to an equivalent online record enabled the contractor to identify missing or erroneous records. Booklets that could not be matched to an online record were flagged for review. The contractor kept a real-time inventory of every test booklet issued using the scanning label affixed to each, which allowed for effective tracking of mismatched records and positive accountability for all materials. Discrepancies were individually reconciled by the contractor and WDE.

Test administrators were required by WDE to complete PAWS administration and security training. WDE also approved all materials used to train contractor personnel in grading the PAWS, replicated the constructed response scoring process using teachers, reviewed the education and training qualifications of graders, and had an onsite representative at the contractor's facility during scoring. Constructed response graders were required to have at least a bachelor's degree and to demonstrate proficiency in a training environment before participating in the scoring process.

During scoring, constructed response items were evaluated against anchor tests that described expected responses for each potential score point.[6] Once scoring began, 20 percent of each grader's constructed response items were rescored each day by another independent grader who did not know the original score. In accordance with the assessment contract, graders maintained an inter-rater reliability[7] (agreement) rate exceeding 65 percent for rescored items. The contractor tracked inter-rater reliability between its graders and provided WDE with daily inter-rater reliability and score distribution reports.

Team leaders reviewed grader scoring in real-time by reading behind randomly selected constructed response items to monitor scoring and improve reader reliability. The contractor also required graders to complete five calibration sets during live scoring each day to ensure that readers maintained intra-reader reliability. Calibration sets consisted of five student papers of mixed quality in random order, which were pre-scored by expert team leaders who were familiar with WDE's scoring parameters. Grader's scores were compared with the known scores by the contractor and a calibration report was prepared. Any reader who failed to score 80 percent of constructed response items correctly was required to be retrained.

Once scoring was completed, WDE subjected the contractor-provided electronic file containing assessment results to a validation process to verify the accuracy of the fields in the file before performing detailed reviews. After the results file was validated and accepted by WDE, accountability staff performed various queries to analyze and test for blank scores that affect participation rates, expected total student counts, comparisons to prior year results, and other quality checks. Concurrently, WDE's Accountability Supervisor performed various reasonableness tests on the contractor-provided assessment results using a sampling-based review of item-level entries, compared multiple-choice scoring to constructed response item scoring, and looked for inconsistencies in the data. The Accountability Supervisor's specific knowledge of individual LEAs' historical performance levels added value to the reasonableness checks performed at WDE. The quality checking activities performed by WDE were essential to ensuring that contractor provided results were accurate and complete when reported to LEAs.

Quality control procedures are used throughout the entire process for developing, scoring, and reporting the PAWS. A standard quality procedure at Pearson is to create a test deck for all testing programs. The test deck process enters intentionally misstated student data into the scoring system to assess the system's ability to detect errors and to ensure that all aspects of scanning, editing, scoring, and reporting are functioning properly. An issues log with sign-off approvals is used to address any issues that arise during the review of the test deck data. Contractor quality control checklists were completed, maintained, and available for review at the scoring site.

WDE staff met weekly with the contractor to discuss general testing issues and deadlines for different procedures. During live scoring, daily reports were provided to WDE documenting data trends, potential problems, and corrective actions. Errors were logged and tracked by WDE and the contractor to ensure that compensating controls were working effectively.

[6] WDE educators reviewed anchor tests before the PAWS was administered.

[7] According to the *Standards for Educational and Psychological Testing*, inter-rater reliability is the consistency with which two or more judges rate the work or performance of test takers.

Wyoming has recruited industry experts to form a Technical Advisory Committee (TAC) to assist in the design, development, and implementation of the PAWS. The TAC meets in-person three times each year and corresponds monthly with WDE assessment staff about the PAWS. The TAC reviewed the contractor-developed and WDE-approved PAWS Technical Report for the Spring 2008 administration, which described the technical characteristics of the PAWS.

AUDIT RESULTS

We concluded that WDE and the contractor implemented adequate controls over scoring. This provided reasonable assurance that reliable assessment results were used to make AYP determinations and to evaluate individual student achievement in school year 2007-2008. Controls for scoring the PAWS were developed by the contractor and WDE, but implementation of controls was primarily performed by the contractor. WDE worked with its assessment contractor to implement a system of controls over scoring that included automated student assessment record tracking, multiple levels of data validation, and quality control procedures that were important to obtaining reasonable assurance in the results. WDE also established a collaborative control environment with the contractor that encouraged frequent communication, fostered continuous improvement of existing processes, and ensured correction of problems once identified.

We evaluated the reliability of WDE's controls over scoring based on current uses of the PAWS, which includes assessing individual student achievement without linking results to high-stakes outcomes (graduation, grade promotion, college entrance, etc.) and for making statistically reliable AYP decisions about Wyoming schools, LEAs, and WDE. We identified two areas where WDE can enhance the level of assurance provided by its scoring controls. Specifically, WDE should document its existing procedures for monitoring contractor performance and reviewing contractor-provided assessment results, which would enhance management's control over operations and allow others to perform and review standard quality control tasks. WDE should also allow additional time for its accountability staff to review contractor-provided assessment results and involve school personnel in the review process. Implementing the above measures may decrease the risk that WDE will publish incorrect results and then have to restate student assessment results and AYP determinations, or both.

As noted in the Other Matters section of this report, in each of the 3 years that the PAWS has been administered, WDE identified errors in the contractor-provided assessment results. In all cases, WDE required the contractor to determine and correct the cause of the errors and produce corrected results. We noted that errors from school years 2006-2007 and 2007-2008 required both public disclosure and revision of PAWS reports sent to parents. According to WDE, the errors in 2007-2008 affected the individual performance classifications of 37 of the more than 44,000 students that took the PAWS. Still, the 2007-2008 error was reported in local newspapers, and in conjunction with publicly reported errors from 2006-2007, may affect user confidence in the reliability of PAWS results.

FINDING – WDE Should Document Its Control Processes and Adjust Timelines to Provide Greater Assurance That Assessment Results Are Reliable

WDE has not documented its procedures for monitoring certain contractor performance areas or reviewing contractor provided assessment results. Written policies and procedures enhance management's control over operations, allow others to perform and review standard quality control tasks, and provide evidence of the technical quality required by the Standards. Furthermore, WDE accountability staff were afforded 19 days to complete reviews of contractor-provided assessment results. Providing more time to follow-up on anomalies with the contractor is especially important to WDE given the problems they have experienced with contractor-provided assessment results in each of the last 3 years. WDE would gain additional time for review by receiving contractor-provided assessment results earlier. Following timelines recommended by the Education Leaders Council report entitled *Model Contractor Standards and State Responsibilities for State Testing Programs* would have provided WDE 48 days to review results in 2007-2008.[8] Developing written policies and procedures and allowing sufficient time for reviews would enhance the level of assurance provided by WDE controls.

WDE Should Develop Written Procedures for Contractor Monitoring and Assessment Results Reviews

The Standards establish technical requirements for assessments to be valid and reliable based on each assessment's intended use including making high- or low-stakes decisions, but do not prescribe specific technical methods. However, as the ELC recommends in its *Model Contractor Standards and State Responsibilities for State Testing Programs* report, a State should "develop its own quality assurance policies to monitor the work of the vendor." Although WDE reviews were embedded sufficiently into contractor processes to obtain reasonable assurance in the results for WDE's uses, WDE could enhance its level of assurance by documenting procedures and reviews over certain contractor processes and results. As an activity using Federal ESEA Title VI grant funds, WDE also has a responsibility under 34 C.F.R. § 80.40(a) for managing the day-to-day operations of grant activities and monitoring grant activities to assure that performance goals are being achieved. Specifically, we noted WDE did not develop written policies and procedures documenting its monitoring of contractor performance or WDE staff reviews of contractor provided results.

Monitoring Contractor Performance. WDE staff described how they verified grader qualifications and training and how reviews of test booklets, score reports, and assessment results were documented in program plans and processing procedures. However, WDE did not provide sufficient documentation for us to confirm that the reviews were performed on a consistent basis or that all procedures were always performed. For example, no written

[8] The ELC report entitled *Model Contractor Standards and State Responsibilities for State Testing Programs* was developed with input from industry testing and education leaders from 11 States. The report states, "for vendors, commitment to following the 'vendor standards' described [in the report] can be cited as evidence of self-regulation and adherence to best practices." "For states, the outlined 'state responsibilities' are intended to provide a model for what is necessary to create a high quality testing program and to serve as guidelines for policymakers enacting reforms in state testing programs."

procedure was available to communicate the process WDE used to verify grader educational and training qualifications for constructed response item scoring. WDE's Assessment Director advised us that she reviewed grader qualifications of randomly selected graders at the contractor's facility and personally observed the live scoring process. The contractor provided us a sample of the grader qualification documents that the Assessment Director reviewed to demonstrate that reviews did occur. However, WDE could enhance its controls by developing written procedures describing how grader education and training qualifications should be verified and by documenting that reviews were performed.

As part of our review, we examined contractor records on the educational qualifications and training for individuals (graders) hired to score Wyoming students' constructed responses. From the population of 485 graders eligible to score the PAWS assessment, we requested educational and training records for a random sample of 25 graders. The contractor and WDE provided evidence demonstrating that 23 of the 25 graders possessed the required 4-year degree to score the PAWS. The contractor's temporary staffing agency could not locate the educational documentation for 2 of the graders. Our sample of 25 graders included 13 employees from the contractor's temporary hiring agency, 10 contractor employees, and 2 Wyoming teachers from the scoring institute. All 21 of the graders that we reviewed (10 contractor employees plus 11 of 13 temporary staffing agency employees) had degrees from accredited U.S. institutions of higher-education. We also reviewed contractor hiring notices and anonymously contacted contractor human resources (HR) personnel regarding employment to assess whether a grader could be hired without a 4-year degree. We concluded from our review that it was unlikely a grader without a 4-year degree could be hired.

The contractor was only able to provide the initial training documentation for 4 of the 25 graders.[9] Although the PAWS contract did not require contractors to retain training documentation, the contractor's Manager of Scoring Resources stated that the training documentation should have nevertheless been retained. The Manager of Scoring Resources advised us that the contractor's paper-based training system will be replaced in school year 2008-2009 by an online system that electronically maintains grader training records. WDE procedures should require the contractor to retain documentation of both grader educational qualifications and training.

WDE's Assessment Director also described the role of its Technical Advisory Committee (TAC) in reviewing contractor reported calibrating, equating, scaling, linking, statistical, and psychometric[10] decisions (refer to Enclosure 1: Glossary in this report for definitions of these

[9] The contractor used a number of controls to ensure the reliability of grader scoring of constructed responses. In addition to the initial training, the contractor had procedures requiring (1) graders to pass contractor-generated calibration exams randomly each day, (2) real-time, second scoring of student responses by a supervisor to ensure consistency with scoring standards, and (3) daily blind rescores of constructed response items to ensure reliability from reader to reader. WDE also approved all training materials and replicated the scoring process and controls using Wyoming educators as the graders at a scoring institute conducted by the contractor in Casper, Wyoming, in May 2008.

[10] Psychometrics is the branch of psychology that deals with the design, administration, and interpretation of quantitative tests for the measurement of psychological variables such as intelligence, aptitude, and personality traits.

processes) as well as the annual PAWS Technical Report.[11] The Council of Chief State School Officers (CCSSO) recommends in its *Quality Control Checklist for Processing, Scoring, and Reporting* that State educational agencies perform reviews in each of the areas that WDE described as being within the scope of the TAC's responsibilities. However, WDE could not provide sufficient documentation for us to confirm that the reviews over these areas were performed.

We determined that the contractor implemented reviews over the areas described in the CCSSO report. For example, the contractor had two psychometricians independently replicate the calibrating, equating, scaling, linking, statistical, and psychometric calculations required for the PAWS. The psychometrician's work was then reviewed by a senior psychometrician before being finalized. However, documenting the TAC procedures used, and the review process itself, would improve the level of assurance provided by WDE's controls over scoring.

Reviewing Assessment Results. WDE's Accountability Supervisor performed error-checking reviews of the contractor-provided assessment results using a risk-based approach that included factors such as known prior errors, knowledge of results most critical to accurate AYP determinations, and knowledge of LEAs' historical performance on State assessments. Another WDE staff, who works closely with the Accountability Supervisor, performed general reasonableness and accuracy checks of the results using database queries to identify trends or problems that might warrant additional review. We requested a list of the reasonableness checks and the accompanying query programming code, but WDE could not provide them. Although WDE's error-checking and reasonableness techniques have detected issues in assessment results prior to publication in each of the last 3 years, the reviews are not systematic or documented to allow for review or consistent replication if one of the employees were to leave WDE. Therefore, WDE should document its processes for reviewing contractor-provided assessment results. WDE's Accountability Supervisor acknowledged that reviews should be more systematic and documented.

Developing written policies and procedures for monitoring assessment contractors and reviewing assessment results would enhance the level of assurance provided by WDE's controls. Documenting its policies and procedures would also be useful to WDE managers in controlling their operations, to new staff involved with performing control functions, and to others involved in analyzing or evaluating quality control operations.

WDE Should Provide Additional Time for Reviewing the Assessment Results Received From Its Contractor

WDE reduced the amount of time staff was provided to detect and correct misstatements each year since the first PAWS administration in 2005-2006. In school year 2007-2008, WDE required the contractor to provide assessment results by June 26, 2008, while it established a target date of July 15 to communicate preliminary AYP determinations to LEAs and schools. Thus, WDE afforded itself about 19 days to validate and review the assessment results provided by the contractor, communicate any problems to the contractor, review corrected results, and

[11] We reviewed the qualifications of the members of the TAC and found that all had an extensive background in educational measurement.

compute preliminary AYP determinations. Reducing the time afforded to WDE staff to detect and correct misstatements in the results makes it more important that WDE receive accurate results from the contractor on the first submission.

We noted that WDE submitted an amended accountability plan to OESE on February 15, 2008, extending the delivery date for initial contractor-provided assessment results from June 2 to June 30. This change was not identified on OESE's list of approved amendments dated July 15, 2008. One OESE official that we contacted stated that substantive changes to State accountability plans must be reviewed, but this change would be considered a non-substantive change that would probably not require review. However, the official noted it would have been better if WDE had discussed the change ahead of time to identify any intended or unintended consequences.

The Department's Non-Regulatory Guidance, entitled *Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting* (April 2006), indicates that when developing reporting timelines, SEAs should allow "substantial time" for followup when data anomalies, missing items, and other data quality issues are identified by their reviews. The Department's guidance also cautions that hurried or ad hoc reporting greatly increases the potential for quality problems. The ELC recommends in its *Model Contractor Standards and State Responsibilities* report that results for examinations containing constructed response items, which the PAWS includes in each tested subject, should be produced by the end of the tested semester or no later than 6 weeks after the contractor receives the exam for scoring. To follow the ELC 6-week recommendation for receiving results, WDE needed to obtain the 2007-2008 results from the contractor no later than May 28, 2008. WDE obtained the results on June 26, 2008, nearly a month after the 6-week date recommended by the ELC report.

Additional time for review is important because the PAWS assessment includes constructed response items in each subject area. Constructed response items increase the time required to obtain assessment results and take longer to score than exclusively machine-scored, multiple-choice items.[12] The use of both multiple-choice and constructed response items in the subject area exams creates two separate scoring procedures for the PAWS with computers scoring multiple-choice items and graders manually scoring constructed response items. The results generated from the two scoring processes must then be combined and matched to each individual student through a complex data translation process. Thus, sufficient time should be allowed for State and local review of the assessment results provided by the contractor.

State-level Reviews. Although the deadline for reporting preliminary AYP results (July 15) has not changed since the PAWS was first administered in April 2006, the agreed-upon date for the contractor to provide assessment results to WDE has been later in each successive year. For the first PAWS assessment cycle in school year 2005-2006, the contractor was required to submit assessment results to WDE for review and analysis by May 15, 2006. In school year 2006-2007, WDE allowed the contractor until June 2, 2007, to deliver results to WDE, 3 weeks later than the prior year. In school year 2007-2008, WDE allowed the contractor until June 26, 2008, to provide results to WDE, nearly 6 weeks later than the first year. Accordingly, the time allotted

[12] Constructed responses are not required by the ESEA.

for WDE staff to review results has declined from 61 days in 2005-2006 to 43 days in 2006-2007 and to 19 days in 2007-2008.

The importance of providing more time for State review of the assessment results is demonstrated by WDE's own experiences. In each of the last 3 years, WDE staff identified problems with the first submission of contractor-provided assessment results that required correction before WDE could issue preliminary AYP determinations.[13] As we mentioned earlier, the timeline for school year 2007-2008 gave WDE less than 3 weeks to review the contractor-provided assessment results. The contractor provided results to WDE on the contractually required deadline. However, WDE staff detected a problem with the initial contractor submission. After the contractor corrected the error, WDE had about 1 week to review the corrected assessment results and make preliminary AYP determinations. Accountability staff described working long days during this period to meet WDE's July 15 deadline for delivering preliminary AYP results to LEAs. Obtaining results within 6 weeks of the end of testing, as recommended by the ELC's report, would have provided WDE 48 days to review contractor-provided assessment results in 2007-2008. Allowing additional time to complete reviews would enhance the level of assurance provided by WDE's review process and would provide more opportunities for timely reviews of assessment results at the local level.

Local-level Reviews. Congress recognized the importance of providing parents and teachers with access to scores as one of the best strategies for ensuring that mistakes that do occur are both identified and corrected in an expedient manner.[14] In cases where assessment results lead a federally funded school to be identified as needing improvement, ESEA §1116(b)(2) requires that a school be provided with the opportunity to review school-level assessment results before AYP determinations are finalized.

For school year 2007-2008, LEA personnel were provided an opportunity to review PAWS results from July 15 to July 30, 2008. However, the LEA review period occurred during the summer break when school personnel that would likely be in the best position to evaluate the results of individual students, such as teachers, were not available to perform the reviews. WDE's Accountability Supervisor explained that LEA personnel are usually the only personnel available to review the information during the appeals window established by WDE. Historically, most LEA appeals have focused on the accuracy of student demographic information and not on individual student assessment results. The three AYP appeals that WDE received from LEAs for school year 2007-2008 were related to suspected errors in student demographic information, not the reliability of assessment results.

Obtaining timely school-level reviews can help prevent and detect errors that might otherwise go undetected and that could require results to be restated. For example, a school-level review of assessment results occurring after final results had been publicly released detected an error that identified 37 students whose writing assessment scores that had not been recorded. In this case, a school official notified the contractor's hotline that a teacher had noticed that two students had erroneous blank scores for expressive writing in their individual student reports. The contractor

[13] The Other Matters section of the report provides additional details on the problems noted in each year.

[14] *Congressional Record*, 12/12/2001, p. H9952, para. 96.

made WDE aware of the suspected error. Of the more than 44,000 students tested, the contractor identified 625 students that had a blank score recorded for either of the two writing questions included in the exam. The contractor determined that 37 of the 625 students had actually completed the writing exam without receiving credit. In response, the contractor ensured that appropriate writing scores were included for the affected students and produced corrected individual student reports for each.

The contractor identified a single team leader as the grader responsible for all 37 of the erroneous blank scores. Normally, 20 percent of a grader's work is subjected to a second review by an independent grader. However, team leaders scoring constructed response items were not subject to this review requirement. Thus, the lack of controls over operational scoring by team leaders enabled the entry of erroneous blank scores to get through the scoring system without being detected by the contractor's rescore procedures. In response to the error, additional scoring controls were implemented including a requirement that all team leaders be held to the existing standards of reader reliability of subordinate graders and that all blank scores be reviewed for accuracy.

According to WDE, the error affected only a small number of students. However, newspaper articles did not always communicate the isolated nature of the error. As a result, stakeholder perceptions about the reliability of the PAWS may have been adversely affected. Implementing a PAWS scoring timeline that allows for local school reviews of contractor-provided assessment results before publication of AYP determinations could enhance WDE's ability to detect errors and increase user confidence in PAWS reporting.

Recommendations

We recommend that the Assistant Secretary for Elementary and Secondary Education advise the Wyoming Department of Education to—

1.1 Develop written, standardized policies and procedures for performing and documenting the monitoring of its assessment contractor(s) and reviewing assessment results provided by its contractor(s), including those pertaining to grader qualifications and training, Technical Advisory Committee reviews, and contractor-provided data validation.

1.2 Evaluate the timeline for scoring the Proficiency Assessments for Wyoming Students and consider providing more time for State and local reviews of the assessment results provided by the contractor before reporting preliminary adequate yearly progress results to local educational agencies and schools.

WDE Comments

In its comments, WDE concurred with the finding. WDE stated that procedures have been developed for documenting control processes used to review contractor performance and contractor-provided results. WDE also stated that timelines for delivery of PAWS results will be revised starting in 2010 to allow WDE and LEAs to perform extended reviews of contractor-provided assessment results as well as additional time to followup on any issues identified.

WDE also provided comments detailing specific corrective actions taken in response to each recommendation. We have not modified our recommendations based on WDE's comments.

- Recommendation 1.1. In response to our recommendation, WDE's Superintendent of Public Instruction indicated that WDE has developed an action plan to monitor and document assessment vendor performance every quarter against specifications identified in the action plan. The Superintendent also described how WDE has created written standardized procedures for performing reviews of scoring sites and scorer educational and training requirements in order to ensure consistent monitoring activities are performed in the future. Additional corrective actions identified by the Superintendent include documenting the assessment results review process performed by the Technical Advisory Committee and the development of procedures to document WDE's reasonableness and accuracy checks of assessment results.

- Recommendation 1.2. In response to our recommendation, WDE's Superintendent of Public Instruction indicated that WDE will receive assessment results during the first week of June beginning with the 2010 PAWS administration. Earlier receipt of results will provide 6 weeks for WDE to review assessment results and followup on any issues before issuing preliminary AYP results to LEAs and schools. The Superintendent also explained that WDE is working to provide LEAs with preliminary results by mid-June to allow LEAs a month and a half to review results before AYP results are made public.

OIG Response

WDE did not provide us with the process and procedure documents described in the Superintendent's comments, but the actions described by the Superintendent would appear to address our finding. We also noted that WDE intends to adjust its timeline for scoring the PAWS by requiring its assessment vendor to provide results in the first week of June beginning in 2010 rather than the last week of June under the prior policy. We commend WDE for developing policies and procedures and adjusting its timelines to enhance controls over assessment scoring.

Although WDE intends to provide LEAs with preliminary assessment results by the middle of June in the future, the personnel in the best position to identify individual student errors, namely teachers, may not have an opportunity to review the results under this plan. For our audit period, publicly disclosed errors were detected by teachers after returning from the summer break. We recognize WDE has developed its AYP reporting timelines to meet legal

requirements, but we encourage WDE to continue refining timelines in an effort to provide teachers with preliminary PAWS results before the school year ends.

OTHER MATTERS

WDE informed us of several errors that have occurred since the implementation of the PAWS. Some errors did not involve scoring or were detected before results were distributed to parents and instructional staff, while other errors were not corrected prior to public release of results. When errors affect only a small group of students and do not impact AYP results, negative publicity may still undermine parent, teacher, and other stakeholder confidence in the reliability of PAWS as an effective measure of student achievement. As a result, users may not use PAWS assessments to identify the specific academic needs of individual students, as intended by Federal law.

Reporting Errors for School Year 2007-2008 That Did Not Impact Scoring of Student Assessments

When WDE received the assessment results for school year 2007-2008, the results contained incorrect "lexile" numbers for high school students. Lexile numbers are diagnostic scores used to match a student's reading level to appropriate reading material. Lexile numbers are not used to identify student performance levels or to make AYP determinations. However, lexile numbers are included in student reports provided to parents. Even though the lexile numbers were slightly misstated, the contractor needed to regenerate the results file and WDE had to revalidate the new file. The errors in the lexile numbers occurred because the contractor did not update a prior year file for an equating error that occurred in the preparation of the assessment results for school year 2006-2007. (See section below on errors detected in prior periods.)

The other error that occurred in school year 2007-2008 also did not impact student performance levels. While investigating the cause and impact of the 37 students with improper blank writing scores (discussed in the Finding), the contractor discovered that the "Writing Item Score Analysis" section of the 2007-2008 PAWS Student Reports displayed incorrect numbers for expressive and expository writing results because of a programming error. As a result, 2007-2008 PAWS reports that were sent to parents had to be reissued. The programming error did not affect the student's overall writing score, performance level, or AYP determinations.

Errors Detected in Prior Periods That Impacted the Scoring of Student Assessments

School Year 2006-2007. The assessment results provided to WDE by the contractor in 2006-2007 did not include the writing scores for more than 2,000 students. WDE detected the problem when conducting reasonableness checks of information in the assessment results file and discovered that participation rates on the writing exam were well below the expected participation rates. The contractor revised the results for WDE 3 times in 3 weeks to correct the error. One LEA's preliminary results were delayed for a week beyond the rest of the LEAs, but

WDE staff received the revised results in sufficient time to meet the public AYP reporting deadline. Correction of the error prevented one school from being identified for restructuring and changed the AYP determinations for two other schools. The contractor determined that the problem was caused by a programming error in the merge process, in which the online student results are combined with the written student results to form the completed student record. Controls over the merge process were improved for school year 2007-2008 when the contractor replaced an outdated system. The new system includes error alerts and a process that matches five data fields (unique identifier, first name, last name, date of birth, and gender) prior to the final merging of the online and paper records.

After publication of final 2006-2007 assessment results, two additional scoring errors were identified while a contractor employee was reviewing another employee's work prior to administration of the 2007-2008 PAWS exam. In one case, the contractor had determined that the cutoff score used to separate about 200 5th grade students between "proficient" and "basic" on the English Language Arts (ELA) component of the assessment was off by one point. One of the contractor's administrative staff caused the error by inputting the wrong cut-score into the contractor's database. In the other case, the number of score points possible on some questions changed from 1-4 in 2005-2006 to 0-4 in 2006-2007 but was not appropriately accounted for in the equating process between the 2 years. Therefore, student results had to be restated for school year 2006-2007. A Wyoming district superintendent we spoke with explained how parents received the second edition of the 2006-2007 individual student reports in September 2008 about a week after receiving the revised student reports for school year 2007-2008 and about 3 weeks after receiving the original 2007-2008 student reports. As a result, some parents received 3 separate PAWS reports, covering student assessment results for 2 different school years, in a span of 3 weeks.

School Year 2005-2006. WDE's Accountability Supervisor explained how several hundred student score sheets were not scored. The supervisor detected the missing scores while reviewing participation rates contained in the contractor provided results. The contractor ultimately found the missing student score sheets in a box that was supposed to contain only non-scoring materials. The contractor graded the student score sheets and provided corrected results to WDE before the release of preliminary AYP determinations. The supervisor attributed the problem to not tracking extra exam materials sent to schools with a serial number and allowing every school to send their testing materials directly back to the contractor. Exam materials intended for a specific student in 2005-2006 had serial numbers, but extra exams did not. To prevent the error from recurring, WDE implemented a policy in 2006-2007 that required serial numbers on all exam materials for accountability and tracking purposes. WDE also developed a new process for returning exam materials to the contractor that included a reconciliation process to ensure the number of materials shipped equaled the number received at the contractor after testing was complete.

OBJECTIVES, SCOPE, AND METHODOLOGY

The objective of our audit was to determine whether controls over scoring of assessments at WDE were adequate to provide reasonable assurance that assessment results are reliable. Our review covered assessments administered in school year 2007-2008 that were used for evaluating individual student achievement and making AYP determinations under Section 1111(b)(3) of the ESEA.

To gain an understanding of the criteria and relevant issues applicable to State assessments, we reviewed Federal laws, regulations, and guidance related to assessments, the Standards for Educational and Psychological Testing (1999), Congressional Record transcripts from passage of the No Child Left Behind Act, multiple professional journals on educational performance measurement, prior Office of Inspector General audit reports and memoranda, peer review guidance and decision letters from the Department's Office of Elementary and Secondary Education, and Government Accountability Office reports on State assessments and internal controls.

To gain an understanding of Wyoming's system of controls over scoring, as well as a general overview of the PAWS we--

- Reviewed media coverage of the PAWS, the Wyoming Single Audit Report for the period ended June 30, 2007, and Department monitoring reports.
- Conducted interviews with personnel from the Department, WDE, and the contractor and conducted a walk-through of the contractor's San Antonio scoring facility.
- Reviewed requests for proposals, vendor proposals, and contracts for the PAWS as well as documentation supporting WDE's procurement process.
- Obtained and reviewed WDE and contractor documentation describing PAWS development, administration, security, processing, scoring, and reporting procedures.
- Reviewed PAWS score merging procedures (matching of results to correct student), test deck and key check procedures, quality control checklists, and psychometric control documents.
- Reviewed the list of errors compiled by the National Board on Educational Testing and Public Policy based at Boston College to identify risk areas in WDE's system for scoring State assessments.
- Reviewed a list of appeals from Wyoming LEAs regarding PAWS data.
- Identified and analyzed funding expended on the PAWS during the contract period.

To evaluate WDE's controls over scoring of assessments and their adequacy for producing reliable results we--

- Reviewed a random sample of 25 PAWS graders from a population of 485 to confirm that PAWS graders possessed the contractually required educational and training qualifications to score the examination.
- Reviewed PAWS inter-reader reliability statistics for compliance with quality assurance requirements established in the contract between WDE and the contractor.

- Compared the WDE and contractor's system of controls to guidelines contained in the January 2003 *Quality Control Checklist for Processing, Scoring, and Reporting* issued by the CCSSO, the February 2002 *Model Contractor Standards and State Responsibilities for State Testing Programs* issued by the ELC, and the Department's April 2006 Non-Regulatory Guidance on *Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting.* As required by ESEA § 1111(b)(3)(C)(iii) and the Standards, we considered the low-stakes individual use and high-stakes AYP use of WDE's assessment in our evaluation of WDE's internal control framework.
- Analyzed current and prior year errors to identify their effect on the reliability of results and determine whether WDE and the contractor took appropriate action to prevent recurrence of the same or similar errors.
- Obtained school year 2006-2007 and school year 2007-2008 PAWS results files from WDE to test controls we classified as high risk.

Consistent with OIG's national review of State assessments, we reviewed the sufficiency of WDE's contractual safeguards for protecting student information obtained during the assessment process in accordance with the Family Educational Rights and Privacy Act (FERPA). We found that relevant FERPA student information safeguards were contained in the WDE contract.

Our review was limited to the main PAWS exam without accommodations. Wyoming also uses an alternate version of the PAWS exam to test less than 1 percent of the State's students with the most serious cognitive disabilities and an exam to assess the proficiency of English Language Learners on reading skills. Although we gained an understanding of these alternate exams, we did not include them in our review.

We performed our onsite fieldwork at WDE's Assessment Office in Laramie, Wyoming, and the contractor's San Antonio, Texas, scoring facility. We held an exit briefing with WDE officials on March 10, 2009. We conducted this performance audit in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objective.

Enclosure 1: Glossary [15]

Anchor Test: A common set of items administered with each of two or more different forms of a test for the purpose of equating the scores obtained on these forms.

Calibrating: In linking test score scales, the process of setting the test score scale, including mean, standard deviation, and possibly shape of score distribution, so that scores on a scale have the same relative meaning as scores on a related scale.

Confidence Interval: An interval between two values on a score scale within which, with specified probability, a score or parameter of interest lies.

Constructed Response Item: An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Short answer items require a few words or a number as an answer, whereas extended-response items require at least a few sentences.

Cut Score: A specified point on a score scale, such that scores at or above that point are interpreted differently from scores below that point.

Equating: Putting two or more essentially parallel tests on a common scale.

Field Test: A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting.

Inter-rater Reliability: The consistency with which two or more judges rate the work or performance of test takers.

Item: A statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task.

Lexile Score: According to WDE's 2008 PAWS Interpretative Guide, a lexile score provides a common scale for matching reader ability and text difficulty.

Linking (Linkage): The result of placing two or more tests on the same scale, so that scores can be used interchangeably.

Psychometrics: According to the American Heritage Dictionary, 4th Edition, the branch of psychology that deals with the design, administration, and interpretation of quantitative tests for the measurement of psychological variables such as intelligence, aptitude, and personality traits.

Reliability: The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable.

Scaling: The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale or by producing scale scores designed to support criterion-referenced or norm-referenced score interpretations.

Technical Manual: A publication prepared by test authors and publishers to provide technical and psychometric information on a test.

Validity: The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

[15] All definitions are from the *Standards for Psychological and Educational Testing (1999)* unless otherwise noted.

Enclosure 2: WDE Comments

Wyoming Department of Education
Dr. Jim McBride, Superintendent of Public Instruction
Hathaway Building, 2nd Floor, 2300 Capitol Avenue
Cheyenne, WY 82002-0050
Phone 307-777-7673 Fax 307-777-6234 Website www.k12.wy.us

May 28, 2009

Raymond Hendren
Regional Inspector General for Audit
U.S. Department of Education
Office of Inspector General
501 I Street Suite 9-200
Sacramento, CA 95814

Dear Mr. Hendren:

The Wyoming Department of Education (WDE) has received and reviewed the draft audit report, Control Number ED-OIG/A09I0012, which contains the findings and recommendations of the audit entitled *Wyoming Department of Education Controls Over State Assessment Scoring.* Please find the WDE's comments regarding these findings and recommendations below.

Finding: WDE should document its control processes and adjust timelines to provide greater assurance that assessment results are reliable.

The WDE concurs with this finding. The following corrective actions have already taken place:

- The WDE has developed processes and procedures for documenting the control processes for reviewing contractor performance areas and the contractor-provided assessment results.
- The WDE has adjusted timelines for the delivery of the Proficiency Assessments for Wyoming Students (PAWS) assessment results beginning in 2010 to allow for complete review of the contractor-provided assessment results by both the WDE and LEAs, and time for the WDE to follow up when issues are identified.

Recommendation 1.1: Develop written, standardized policies and procedures for performing and documenting the monitoring of its assessment contractor(s) and reviewing assessment results provided by its contractor(s), including those pertaining to grader qualifications and

training, Technical Advisory Committee reviews, and contractor provided validation.

The WDE concurs with this recommendation. The following corrective actions have already taken place:

- The WDE developed an Action Plan to monitor all aspects of the 2008-2012 contract with our assessment vendor. The contractor performance is reviewed against the specifications in the Action Plan and is documented on a quarterly basis.
- During the 2008-2009 school year, the WDE worked with our Technical Advisory Committee (TAC) to develop written, standardized processes and procedures for the monitoring of the assessment contractor's scoring sites, and scorer qualifications and training. Members of the WDE Standards and Assessment Unit visited scoring sites following the 2009 PAWS administration window. The Wyoming Department of Education's newly developed monitoring tool, *Check List for Monitoring Scoring Sites,* consistent with all contractual requirements, was utilized during onsite scoring site visit. The checklist includes the review of scorer qualifications and training, scoring procedures, as well as the monitoring of the safety and security of the scoring facilities. This document will assure that all future monitoring will be performed on a consistent basis.
- The TAC reviews the PAWS assessment results provided in the initial draft version of each year's Technical Manual. The WDE will document the TAC review through written comments provided to the department via email and through minutes following the TAC meeting in which the results are provided.
- The WDE has developed procedures described in *Quality Controls for State Assessment Data,* to document the general reasonableness and accuracy checks of the assessment results. The risk-based approach to evaluate assessment results is included in this document. The *Quality Controls for State Assessment Data* will be available to new staff involved with performing control functions, and to others involved in analyzing or evaluating quality control operations.

Recommendation 1.2: Evaluate the timeline for scoring the Proficiency Assessments for Wyoming Students and consider providing more time for state and local reviews of the assessment results provided by the contractor before reporting preliminary adequate yearly progress results to local educational agencies and schools.

Raymond Hendren
May 28, 2009
Page 3

The WDE concurs with this recommendation. The following corrective actions are in development:

- The WDE will receive assessment results during the first week of June beginning with the 2010 PAWS administration. This will allow the WDE six weeks to review the results before providing LEAs and schools with preliminary adequate yearly progress (AYP) results.

- The WDE is developing processes to provide the LEAs their results in mid-June, allowing a full month and a half for their review before the WDE publicly releases AYP and State Assessment results.

Thank you for allowing the WDE to provide written comments on the findings and recommendations regarding the audit.

Sincerely,

/s/

Jim McBride, Ed.D.

JM:LW