Spring 2017 Georgia Milestones Assessment Desktop Audit Results

January 2018

THE GOVERNOR'S OFFICE OF
STUDENT ACHIEVEMENT

Table of Contents

Annual Assessment Analysis Summary

The Governor's Office of Student Achievement (GOSA) serves as the reporting and accountability agency for Georgia education. As such, it is charged by law with inspecting academic records of schools to ensure that education institutions are faithful to performance accountability requirements. Through an academic audit, GOSA reviews student assessment data and other school records reported to the State to confirm accuracy and explore the effectiveness of local school initiatives in improving achievement.

The Georgia Milestones Assessment Audit is divided into two separate analyses:

The Answer Change Analysis, formerly known as the Erasure Analysis, identifies classrooms and schools where the number of wrong answers that have been changed to right answers on individual student answer sheets is well above the state average. It is conducted on the following assessments:

- **Assessments Included**
  - Grades 3 to 8 End of Grade Assessments (EOG) in English-Language Arts and Mathematics, and in Grades 5 and 8 in Science and Social Studies.
  - Grades 7-12 End of Course Assessments (EOC) in the following ten high school courses: Ninth Grade Literature and Composition, American Literature and Composition, Coordinate Algebra, Algebra I, Analytic Geometry, Geometry, Physical Science, Biology, U.S. History, and Economics.
- **Flagging Methodology**
  - EOG (Grades 3-8)
    - Five percent or more of classrooms in a school are flagged at four standard deviations or greater, OR
    - One classroom is flagged at seven standard deviations or greater.
  - EOC (Grades 9-12)
    - Five percent or more of classrooms in a school are flagged at five standard deviations or greater, OR
    - One classroom is flagged at seven standard deviations or greater.

The Unusual Response Pattern Analysis, identifies schools that have unexpected test score gains across years (95th percentile or higher) using a cohort of students as well as unexpected patterns in student answers (95th percentile or higher). For unexpected patterns, it examines the following four areas: The patterns examined include (1) unlikely blocks of consecutive, identical answers, (2) highly correlated answers across tests, (3) correlation of responses across test items, and (4) cases where students miss easy items but answer difficult answers correctly.

- **Assessments Included**
  - Grades 4-8 English Language Arts and Mathematics
- **Flagging Methodology**
  - Schools where two or more testing groups had test score gains and unusual response patterns that were in the 95th percentile, OR
  - Schools where one testing group had test score gains and unusual response patterns that were in the 99th percentile.

Appendices B, C, and D provide more in-depth information on the calculation formulas and business rules. It is important to note that the results of both analyses are used as an initial flag to spur further investigation of many indicators to determine if any cheating occurred. It is important to note that the flags do not indicate that cheating has occurred.

Analysis Results Summary

After the desktop audit, 83 schools in 44 LEAs have been identified for further inquiry, split out by analysis below:[1]

EOG Answer Change Analysis
177 classrooms in 69 schools in 37 LEAs were identified for an initial desktop audit. After the audit, 37 schools in 26 LEAs require further inquiry.

EOC Answer Change Analysis
30 classrooms in 12 schools in 5 LEAs were identified for an initial desktop audit. After the audit, 9 schools in 4 LEAs require further inquiry.

EOG Unusual Response Pattern Analysis
81 testing groups in 43 schools in 28 LEAs require further inquiry. No desktop audit was conducted for this analysis.

The following report contains the results of GOSA's desktop audit and recommendations to the State Board of Education for actions to be taken in schools requiring further inquiry by the State. Once inquiries are complete in spring 2018, GOSA will publish a report with detailed information on each school's inquiry results.

[1] The number of schools flagged in each analysis does not add up to the total of 86 because three schools were flagged in two of the analyses.

EOG Answer Change Desktop Audit Results

Analysis Overview

GOSA identified schools for a desktop audit when five percent or more of classrooms in a school were flagged at four standard deviations or greater, or one classroom was flagged at seven standard deviations or greater. In total, 177 classrooms in 69 schools in 37 LEAs were identified for an initial desktop audit to determine a possible explanation for the flag that would remove the need for further inquiry. In this audit analysis, many school-level factors, outlined in detail in Appendix A, were reviewed holistically and discussed as a team before any determinations were made. GOSA placed schools in one of two categories: "further inquiry needed," or "no further inquiry needed."

After the desktop audit, 37 schools in 26 LEAs, listed on the next page, require further inquiry. These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons exist to explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred. Appendix B lists the number of classrooms at each school requiring further inquiry for both EOG and EOC.

Schools Requiring Further Inquiry*

**Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.*

- Baldwin County, Blandy Hills Elementary School
- Baldwin County, Creekside Elementary School
- Buford City, Buford Middle School
- Camden County, Matilda Harris Elementary School
- Carroll County, Central Elementary School
- Chatham County, Hubert Middle School
- Chatham County, The STEM Academy at Bartlett
- Cherokee County, Teasley Middle School
- Clarke County, Cleveland Road Elementary School
- Clarke County, Oglethorpe Avenue Elementary
- Cobb County, Campbell Middle School
- Cobb County, Kennesaw Charter School
- Cobb County, Tapp Middle School
- Colquitt County, Stringfellow Elementary School
- DeKalb County, Kittredge Magnet School
- DeKalb County, Montgomery Elementary School
- DeKalb County, Narvie Harris Elementary School
- Douglas County, Factory Shoals Middle School
- Fayette County, Whitewater Middle School
- Floyd County, Armuchee Elementary School
- Forsyth County, I-Achieve Academy
- Fulton County, Amana Academy School
- Fulton County, Northwestern Middle School
- Fulton County, Spalding Drive Elementary
- Gwinnett County, Kanoheda Elementary School
- Gwinnett County, Sycamore Elementary School
- Gwinnett County, Trip Elementary School
- Houston County, Matthew Arthur Elementary School
- Jackson County, East Jackson Middle School
- Jefferson City, Jefferson Academy
- Laurens County, Northwest Laurens Elementary School
- Lee County, Lee County Elementary School
- Muscogee County, North Columbus Elementary School
- Newton County, Fairview Elementary School
- Pulaski County, Pulaski County Elementary School
- Rockdale County, Sims Elementary School
- State Charter School, Georgia Cyber Academy

EOC Answer Change Desktop Audit Results

Schools were identified for a desktop audit when five percent or more of classrooms in a school were flagged at five standard deviations or greater, or one classroom was flagged at seven standard deviations or greater. In total, 76 classrooms in 30 schools in 12 LEAs were identified for an initial desktop audit to determine a possible explanation for the flag that would remove the need for further inquiry. In this audit analysis, many school-level factors, outlined in detail in Appendix A, were reviewed holistically and discussed as a team before any determinations were made. GOSA placed schools in one of two categories: "further inquiry needed," or "no further inquiry needed."

After the desktop audit, nine schools in four LEAs, listed below, require further inquiry. These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons exist to explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred. Appendix B lists the number of classrooms at each school requiring further inquiry for both EOG and EOC.

Schools Requiring Further Inquiry*
**Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.*

- DeKalb County, Chamblee Charter High School
- Forsyth County, Forsyth Central High School
- Forsyth County, Lambert High School
- Forsyth County, Piney Grove Middle School
- Forsyth County, Riverwatch Middle School
- Fulton County, Crabapple Middle School
- Fulton County, Northwestern Middle School
- Fulton County, Webb Bridge Middle School
- Muscogee County, Columbus High School

EOG Unusual Response Pattern Desktop Audit Results

Analysis Overview

GOSA identified schools for further inquiry that met one of the following criteria:

- Schools where two or more testing groups had test score gains and unusual response patterns that were in the 95th percentile, OR
- Schools were one testing group had test score gains and unusual response patterns that were in the 99th percentile.

Each testing group is the total number of students by grade level and subject area (ELA or mathematics) who took a certain test form (A or B) regardless of classroom assignment. For example, all students in a school who took the 4th grade mathematics Georgia Milestones Form A assessment are a testing group.

All schools identified under the criteria are kept for further inquiry.

Using these criteria, 43 schools in 28 LEAs, listed on the next page, require further inquiry. These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons exist to explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred. Appendix B lists the subject areas, test forms, and grade levels flagged for each school.

Schools Requiring Further Inquiry*

**Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.*

- Appling County, Appling County Middle School
- Atlanta Public Schools, Atlanta Classical Academy
- Atlanta Public Schools, KIPP Strive Primary
- Atlanta Public Schools, Morningside Elementary School
- Bryan County, Bryan County Elementary School
- Buford City, Buford Middle School
- Carroll County, Bowdon Elementary School
- Catoosa County, Ringgold Middle School
- Cherokee County, Hickory Flat Elementary School
- Cherokee County, Liberty Elementary School
- Cherokee County, Sixes Elementary School
- Cobb County, Dodgen Middle School
- Cobb County, Kemp Elementary School
- Cobb County, Mountain View Elementary School
- Coweta County, East Coweta Middle School
- Coweta County, Smokey Road Middle School
- DeKalb County, DeKalb Academy of Technology
- DeKalb County, Edward L. Bouie, Sr. Elementary School
- DeKalb County, Kittredge Magnet School
- DeKalb County, Stephenson Middle School
- DeKalb County, Stone Mountain Elementary School
- Dougherty County, Albany Middle School
- Douglas County, New Manchester Elementary School
- Fayette County, Spring Hill Elementary School
- Fayette County, Whitewater Middle School
- Forsyth County, George A. Whitlow Elementary School
- Forsyth County, Riverwatch Middle School
- Forsyth County, South Forsyth Middle School
- Fulton County, Webb Bridge Middle School
- Gilmer County, Gilmer Middle School
- Gwinnett County, Berkeley Lake Elementary School
- Gwinnett County, Trip Elementary School
- Heard County, Heard County Middle School
- Henry County, Locust Grove Middle School
- Lowndes County, Hahira Elementary School
- Madison County, Madison County Middle School
- Mitchell County, Mitchell County Middle School
- Newton County, Newton County Theme School at Ficquett
- Peach County, Byron Middle School
- Pelham City, Pelham City Middle School
- Richmond County, Wheeless Road Elementary School

- State Charter School, Cirrus Charter Academy
- State Charter School, Georgia Connections Academy

GOSA Recommendations in Schools Requiring Further Inquiry

Overall, 83 schools in 43 LEAs require further inquiry. GOSA recommends the following actions to the SBOE in these schools:

GOSA will:

1. Share EOG/EOC data files with superintendents of LEAs that have schools requiring further inquiry to facilitate:
    - LEA internal investigation of reason(s) for flags, and
    - Submission of online inquiry form to GOSA with results of investigation and an explanation of testing protocols in place.
2. Require schools with outstanding inquiry concerns to rotate teachers for the 2018 Georgia Milestones test administration (EOG).
3. Assign state monitors to observe and inspect identified schools requiring further inquiry for the 2018 Georgia Milestones test administration as necessary (EOG and EOG).
4. Conduct on-site audits as necessary.

Appendix A: EOG/EOC Schools Requiring Further Inquiry Summary Table

The following list includes the number of classrooms flagged in the 83 schools requiring further inquiry after the desktop audit. Many reasons may explain why the school requires further inquiry. It is important to note that the flag does not mean that cheating has occurred.

System Name	School Name	Answer Change Classrooms Requiring Further Inquiry	Unusual Response Subject/Grade Form Requiring Further Inquiry	Answer Change Classrooms Requiring Further Inquiry
These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.				
APPLING COUNTY	APPLING COUNTY MIDDLE SCHOOL		Math 6A, English 7B	
ATLANTA PUBLIC SCHOOLS	ATLANTA CLASSICAL ACADEMY		Math 7A	
ATLANTA PUBLIC SCHOOLS	KIPP STRIVE PRIMARY		English 4A/4B	
ATLANTA PUBLIC SCHOOLS	MORNINGSIDE ELEMENTARY SCHOOL		English 5A/5B	
BALDWIN COUNTY	BLANDY HILLS ELEMENTARY SCHOOL	2		
BALDWIN COUNTY	CREEKSIDE ELEMENTARY SCHOOL	3		
BRYAN COUNTY	BRYAN COUNTY ELEMENTARY SCHOOL		English 4B	
BUFORD CITY	BUFORD MIDDLE SCHOOL		Math 6B, English 8B	
CAMDEN COUNTY	MATILDA HARRIS ELEMENTARY SCHO	3		
CARROLL COUNTY	BOWDON ELEMENTARY SCHOOL		Math 4A/4B	
CARROLL COUNTY	CENTRAL ELEMENTARY SCHOOL	1		
CATOOSA COUNTY	RINGGOLD MIDDLE SCHOOL		Math 8A/8B	
CHATHAM COUNTY	HUBERT MIDDLE SCHOOL	1		
CHATHAM COUNTY	THE STEM ACADEMY AT BARTLETT	7		
CHEROKEE COUNTY	HICKORY FLAT ELEMENTARY SCHOOL		English 5A/5B	
CHEROKEE COUNTY	LIBERTY ELEMENTARY SCHOOL		Math 5B, English 5A	
CHEROKEE COUNTY	SIXES ELEMENTARY SCHOOL		Math 4A/4B	
CHEROKEE COUNTY	TEASLEY MIDDLE SCHOOL	9		
CLARKE COUNTY	CLEVELAND ROAD ELEMENTARY SCHO	4		
CLARKE COUNTY	OGLETHORPE AVENUE ELEMENTARY S	5		
COBB COUNTY	CAMPBELL MIDDLE SCHOOL	1		
COBB COUNTY	DODGEN MIDDLE SCHOOL		Math 6A, English 6A	
COBB COUNTY	KEMP ELEMENTARY SCHOOL		Math 5A	
COBB COUNTY	KENNESAW CHARTER SCHOOL	1		
COBB COUNTY	MOUNTAIN VIEW ELEMENTARY SCHOOL		Math 4A	
COBB COUNTY	TAPP MIDDLE SCHOOL	11		
COLQUITT COUNTY	STRINGFELLOW ELEMENTARY SCHOOL	1		
COWETA COUNTY	EAST COWETA MIDDLE SCHOOL		Math 6A/6B	

System Name	School Name	Answer Change Classrooms Requiring Further Inquiry	Unusual Response Subject/Grade Form Requiring Further Inquiry	Answer Change Classrooms Requiring Further Inquiry
These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.				
COWETA COUNTY	SMOKEY ROAD MIDDLE SCHOOL		English 8A/8B	
DEKALB COUNTY	CHAMBLEE CHARTER HIGH SCHOOL			3
DEKALB COUNTY	DEKALB ACADEMY OF TECHNOLOGY		Math 6A, English 7A/4A	
DEKALB COUNTY	EDWARD L. BOUIE, SR. ELEMENTARY SCHOOL		Math 5A/5B	
DEKALB COUNTY	KITTREDGE MAGNET SCHOOL	2	English 5A/5B	
DEKALB COUNTY	MONTGOMERY ELEMENTARY SCHOOL	2		
DEKALB COUNTY	NARVIE HARRIS ELEMENTARY SCHOO	2		
DEKALB COUNTY	STEPHENSON MIDDLE SCHOOL		Math 8A/8B	
DEKALB COUNTY	STONE MOUNTAIN ELEMENTARY SCHOOL		Math 5A/5B	
DOUGHERTY COUNTY	ALBANY MIDDLE SCHOOL		Math 6A/6B	
DOUGLAS COUNTY	FACTORY SHOALS MIDDLE SCHOOL	6		
DOUGLAS COUNTY	NEW MANCHESTER ELEMENTARY SCHOOL		Math 5A/5B	
FAYETTE COUNTY	SPRING HILL ELEMENTARY SCHOOL		English 4A/4B	
FAYETTE COUNTY	WHITEWATER MIDDLE SCHOOL	2	English 7A/8B	
FLOYD COUNTY	ARMUCHEE ELEMENTARY SCHOOL	4		
FORSYTH COUNTY	FORSYTH CENTRAL HIGH SCHOOL			1
FORSYTH COUNTY	GEORGE A. WHITLOW ELEMENTARY SCHOOL		English 4A/4B	
FORSYTH COUNTY	I-ACHIEVE ACADEMY	1		
FORSYTH COUNTY	LAMBERT HIGH SCHOOL			9
FORSYTH COUNTY	PINEY GROVE MIDDLE SCHOOL			2
FORSYTH COUNTY	RIVERWATCH MIDDLE SCHOOL		English 6B	2
FORSYTH COUNTY	SOUTH FORSYTH MIDDLE SCHOOL		English 6A/6B	
FULTON COUNTY	AMANA ACADEMY SCHOOL	1		
FULTON COUNTY	CRABAPPLE MIDDLE SCHOOL			2
FULTON COUNTY	NORTHWESTERN MIDDLE SCHOOL	1		2
FULTON COUNTY	SPALDING DRIVE ELEMENTARY	1		
FULTON COUNTY	WEBB BRIDGE MIDDLE SCHOOL		Math 7A/7B	2
GEORGIA CYBER ACADEMY	GEORGIA CYBER ACADEMY	2		
GILMER COUNTY	GILMER MIDDLE SCHOOL		Math 6A, English 5A/5B	
GWINNETT COUNTY	BERKELEY LAKE ELEMENTARY SCHOOL		Math 4A/4B	
GWINNETT COUNTY	KANOHEDA ELEMENTARY SCHOOL	7		
GWINNETT COUNTY	SYCAMORE ELEMENTARY SCHOOL	1		
GWINNETT COUNTY	TRIP ELEMENTARY SCHOOL	6	Math 5A	
HEARD COUNTY	HEARD COUNTY MIDDLE SCHOOL		Math 6A/8A	

System Name	School Name	Answer Change Classrooms Requiring Further Inquiry	Unusual Response Subject/Grade Form Requiring Further Inquiry	Answer Change Classrooms Requiring Further Inquiry
These schools will submit an inquiry form to GOSA describing the reasoning for the flag and steps taken to reduce the likelihood for future flags. Many reasons may explain why a school requires further inquiry. It is important to note that the flags do not indicate that cheating has occurred.				
HENRY COUNTY	LOCUST GROVE MIDDLE SCHOOL		Math 8A/8B	
HOUSTON COUNTY	MATTHEW ARTHUR ELEMENTARY SCHO	2		
JACKSON COUNTY	EAST JACKSON MIDDLE SCHOOL	1		
JEFFERSON CITY	JEFFERSON ACADEMY	3		
LAURENS COUNTY	NORTHWEST LAURENS ELEMENTARY	4		
LEE COUNTY	LEE COUNTY ELEMENTARY SCHOOL	5		
LOWNDES COUNTY	HAHIRA ELEMENTARY SCHOOL		Math 5B	
MADISON COUNTY	MADISON COUNTY MIDDLE SCHOOL		Math 7A/7B	
MITCHELL COUNTY	MITCHELL COUNTY MIDDLE SCHOOL		Math 5A/6A	
MUSCOGEE COUNTY	COLUMBUS HIGH SCHOOL			2
MUSCOGEE COUNTY	NORTH COLUMBUS ELEMENTARY	3		
NEWTON COUNTY	FAIRVIEW ELEMENTARY	2		
NEWTON COUNTY	NEWTON COUNTY THEME SCHOOL AT FICQUETT		Math 4A, English 6A	
PEACH COUNTY	BYRON MIDDLE SCHOOL		Math 8A/8B	
PELHAM CITY	PELHAM CITY MIDDLE SCHOOL		Math 6A	
PIKE COUNTY	PIKE COUNTY ELEMENTARY SCHOOL	1		
PULASKI COUNTY	PULASKI COUNTY ELEMENTARY SCHOOL	1		
RICHMOND COUNTY	WHEELESS ROAD ELEMENTARY SCHOOL		English 4A/4B	
ROCKDALE COUNTY	SIMS ELEMENTARY SCHOOL	2		
STATE CHARTER SCHOOL	CIRRUS CHARTER ACADEMY		Math 4A/4B/5B	
STATE CHARTER SCHOOL	GEORGIA CONNECTIONS ACADEMY		Math 8A/8B	
Totals: 44 LEAs	**83 schools**	**111 classrooms**	**81 Form/Grade/Subject Area Testing Groups**	**25 classrooms**

Appendix B: Desktop Audit Indicators

Desktop Audit Indicators Reviewed
Number of classrooms flagged in each school and whether the flagged classrooms had different test administrators.
Total answer changes and number of wrong-to-right (w-t-r) at the classroom level, including student-level data to determine whether answer changes are concentrated in a small number of students. Classrooms where more than 50% of students in a classroom have zero answer changes and/or w-t-r answer changes reduce the likelihood of systematic or widespread changes in answers from wrong to right.
The severity of the individual flagged classroom (i.e. the standard deviation value or how far from what is considered normal behavior is the class positioned). EOC flags between 5.0 and 6.0 SDs are of less concern than those over 6.0 SDs.
Percentage of total classroom answer changes changed from w-t-r. Generally, classrooms with greater than 60% of answer changes being w-t-r are of concern, unless a classroom with multiple students had one student with many w-t-r answer changes, suggesting that systematic cheating was unlikely.
The number of students in each classroom. (Example: Extremes in classroom populations on both ends of the distribution can skew post-calculation metrics and in turn cause flagged classrooms.).
Classroom percentile ranks of wrong-to-right answer changes by student to observe the distribution of answer changes in a classroom and compare that distribution to the state distribution. For example, comparing a classroom's 50^{th} and 90^{th} percentile with the state 50^{th} and 90^{th} percentiles can identify whether abnormal distributions and/or outliers.
The type of school (i.e. high transient population, alternative education program, residential treatment facilities, etc.).
School demographics and groups (ELL population, gifted, magnet, students with disabilities, etc.).
Variance in performance level data from previous years (not applicable in 2014-2015 due to Georgia Milestones transition).
History as a school of concern.
Prior test monitoring and/or an on-site audit by state personnel.
District personnel and/or policies currently implemented to support test security.
Review of state monitor notes and/or forms.

Appendix C: 2017 DRC EOG Answer Change Executive Report

Analysis of Answer Changes

Submitted by DRC

October 2017

With the high-stakes nature of large-scale assessments such as the Milestones End of Grade (EOG), there are times when students' responses, and hence their scores, may not be a true representation of their own abilities. Various activities may take place, such as a student copying from another student's paper, students receiving inappropriate assistance before or during testing, or students' responses altered after testing. To maintain the integrity of the Milestones EOG and the validity of the results, it is important that any such instances be discovered.

The present study investigated student responses on the English Language/Arts, Mathematics, Science and Social Studies tests of the 2017 Spring Milestones EOG that a) an answer choice was replaced by a different answer choice and b) changed from a wrong answer to a right answer (wrong-to-right).

It should be emphasized that results from the erasure analyses performed in 2017 should only be used to identify potential problems within individual classrooms. That is, these types of analyses must be supported by additional, collateral information before conclusions regarding any improprieties are reached.

Answer Changes for Paper Administrations

The GA Milestones EOG paper-pencil answer documents were processed using high speed 5000i optical scanners which reliably captured document images and optical mark read data. The sophisticated proprietary scoring software system, specifically Optical Mark Recognition (OMR) software, reviews the integrity of each batch of documents scanned according to pre-defined guidelines and services.

The OMR software provides a mechanism for identifying multiple-marks and identification of erasures for scanned data to support answer change analysis. The basis of the answer change analysis is to count erasures for multiple-choice items where two or more responses have been made with a specified intensity. Erasure analyses provide a mechanism to differentiate between three kinds of answer changes: a) wrong-to-wrong, b) right-to-wrong and c) wrong-to-right. Capturing the frequency of answer changes from wrong-to-right can be useful for identifying potential instances of cheating at the student level. Erasure analyses results can be grouped to tentatively identify problems at the classroom and school levels.

Answer Changes for Online Administrations

The test administration software that delivers the Georgia Milestones assessment system, INSIGHT, captures answer changes during online testing sessions. Similar to paper based administrations where answer changes are determined by examining erasure marks, the INSIGHT system records changes to answers within an online test administration that are made either before leaving an item or upon returning to the item and making a change. Answer change analyses for

students testing online also focuses on the three kinds of changes: a) wrong-to-wrong, b) right-to-wrong and c) wrong-to-right. As with paper based erasure analyses, capturing the frequency of answer changes from wrong-to-right can be useful for identifying potential instances of cheating at the student level in online testing. Analyses results can be grouped to tentatively identify problems at the classroom and school levels.

Method

The basis for the answer change analysis is to count erasures in items where an answer choice was erased and replaced with another answer choice; online, an item was selected and then later changed to a different answer choice. Herein, both actions are referred to as an erasure. Often the data captured is useful for identifying cases of cheating. During erasure analysis, two sets of erasures were analyzed: all erasures and wrong-to-right erasures where an incorrect answer choice was erased and replaced with the correct answer choice. Operational and field test multiple-choice (MC) items were used for the answer change analyses implemented for the 2017 Georgia Milestones.

The basic idea underlying the procedure is a statistical test of the null hypothesis (H0) that the mean number of erasures for a class constitutes a random sample from the state distribution of erasures. The hypothesis is tested against the (right-sided) alternative (H1) that the mean number is too high to be explained by random sampling. Classes for which H0 has to be rejected are flagged for further scrutiny. A well-known central limit theorem in statistics tells us that the sampling distribution of the mean number of erasures for class i (mi) is asymptotically normal with mean and standard deviation (SD)

$$mean(m_i) = \mu \qquad (1)$$

$$SD(m_i) = \frac{\sigma}{\sqrt{n_i}} \qquad (2)$$

where ni and mi denote the size and mean number of erasures for class i, respectively. In addition, μ and σ denote the mean and the SD of the distribution of the number of erasures of the population of individual students in the state of Georgia.

The classes were flagged if their mi was larger than $\mu + 4\frac{\sigma}{\sqrt{n_i}}$. Statistically, the flagging criterion set at or above 4σ is conservative. The standard normal table shows that under random sampling the (asymptotic) probability of a sample mean being more than four SDs above the population mean is around 0.00003. However, rejection of H0 only tells us that the observed mean number of erasures is unlikely to be the result of random sampling.

It is evident in the formula that the class flagging criterion for each class is adjusted for the number of test takers in a classroom. For example, if the state mean and SD of erasure count are 1.73 and 2.11, respectively, the flagging criterion for a class size of 20 is adjusted to 3.62 ($1.73 + 4\frac{2.11}{\sqrt{20}} = 3.62$).

This adjustment ensures that the flagging criterion is equally stringent for classes with considerably different numbers of test takers. In addition, minimizing the probability of false

positive (Type I) errors in this statistical test is crucial in this analysis.

Results

Tables 1 and 2 reports the state summary of erasure counts for paper-pencil and online respectively. The tables include the number of students, the total number of all types of erasures, the mean and the SD of all types of erasures, the correlation between all erasures and wrong-to-right erasures, the number of erasures at the 50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th} percentiles, and the maximum number of all types of erasures. The mean number of paper-pencil erasures across all courses ranged from 0.80 to 1.40, and mean number of online answer changes ranged from 6.48 to 11.23 for the 2017 Spring Milestones EOG. In other words, approximately 1 to 2 answer changes were made per student paper-pencil answer sheet on average, and 6 to 11 answer changes were made per online student assessment. The erasure count at specific percentile points (50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th}) is also reported. The erasure count at the 95^{th} percentile point was between 3 and 5 on paper-pencil answer sheets, and between 14 and 24 online.

Tables 3 and 4 report the state summary of wrong-to-right erasure counts for paper-pencil and online respectively. The tables include the number of students, the number of wrong-to-right erasures, the mean and the SD of wrong-to-right erasures, the correlation between all erasures and wrong-to-right erasures, the number of wrong-to-right erasure at the 50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th} percentiles, and the maximum number of wrong-to-right erasures. As can be expected, the mean wrong-to-right erasure count and the count at the specific percentile points were lower than those obtained from all erasure counts. The mean number of paper-pencil wrong-to-right erasures ranged from 0.39 to 0.74, and mean number of online answer changes ranged from 2.81 to 5.26 for the 2017 Spring Milestones EOG. In other words, approximately 0 to 1 wrong-to-right answer changes were made per student paper-pencil answer sheet on average, and approximately 3 to 5 wrong-to-right answer changes were made per online student. The wrong-to-right erasure count at specific percentile points (50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th}) is also reported. The wrong-to-right erasure count at the 95^{th} percentile point was between 2 and 3 on paper-pencil answer sheets, and between 7 and 12 online.

Tables 5 and 6 present a summary of the number of schools flagged across four content areas - English Language/Arts, Mathematics, Science and Social Studies – within each analysis of the Milestones EOG for paper-pencil and online respectively. For each analysis, the number of schools was computed in two ways: flagged for at least one content area or flagged for all four content areas. The number/percentage of schools that had zero flags for all erasures and wrong-to-right erasures in English Language/Arts, Mathematics, Science and Social Studies is provided in Tables 7 and 8. The number/percentage of schools that had less than 1% of the classes flagged for all erasures and wrong-to-right erasures in English Language/Arts, Mathematics, Science and Social Studies and across grades is provided in Tables 9 and 10.

Discussion

With respect to the erasure analyses, the following caveats are always applicable:

1. The normal distribution holds only for large classes; for smaller classes, the result is approximate.

2. Rejection of H_0 does not necessarily imply cheating. Alternative explanations are possible.

3. The flagging criterion should thus be taken as a stimulus to look for additional evidence and find out what happened in the school.

4. The groups of students taking the tests online and paper are not equivalent. Comparing the magnitude of answer changes between testing modes cannot be supported given that the groups likely differ in ability and other key background characteristic.

This answer change analysis is considered a check for unusual numbers of answer changes to student responses. Without additional layers added to the analysis, this kind of check only addresses the possibility, not the certainty, of teachers or administrators altering the responses of students. The 2017 erasure analyses represent an important step in helping to maintain the integrity of future administrations of the Milestones EOG.

Table 1. State Summary Statistics for All Types of Erasure (ERA) Counts by Content/Grade Paper-Pencil

Content	Grade	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
							50	75	90	95	99	99.9	
ELA	3	43705	51027	1.17	1.64	0.79	1	2	3	4	7	13.296	48
	4	39553	32163	0.81	1.42	0.79	0	1	2	3	6	10.448	53
	5	24396	19585	0.80	1.39	0.78	0	1	2	3	6	12	35
	6	41410	41749	1.01	1.51	0.80	1	1	3	4	7	12	26
	7	39607	35637	0.90	1.42	0.78	0	1	3	4	6	11	36
	8	24722	21256	0.86	1.43	0.81	0	1	3	4	6	11	52
MA	3	43705	52528	1.20	1.75	0.82	1	2	3	5	8	12	23
	4	39553	46230	1.17	1.79	0.82	1	2	3	5	8	13	66
	5	24396	25046	1.03	1.61	0.82	0	1	3	4	7	12	33
	6	41410	50247	1.21	1.74	0.82	1	2	3	5	8	13	34
	7	39607	33383	0.84	1.48	0.82	0	1	3	4	6	12	48
	8	24722	21382	0.86	1.63	0.83	0	1	3	4	7	13	20
SC	5	24396	29971	1.23	1.96	0.83	1	2	3	5	9	16	57
	8	24722	20482	0.83	1.82	0.83	0	1	3	4	8	16	39
SS	5	24396	31772	1.30	1.99	0.83	1	2	4	5	9	16	30
	8	24722	34567	1.40	2.05	0.84	1	2	4	5	9	17	39

Table 2. State Summary Statistics for All Types of Erasure (ERA) Counts by Content/Grade Online

Content	Grade	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
							50	75	90	95	99	99.9	
ELA	3	92882	741668	7.99	4.81	0.69	7	11	14	17	23	33	53
	4	96534	684718	7.09	4.52	0.71	6	9	13	15	22	30	53
	5	108829	822191	7.55	4.58	0.71	7	10	14	16	22	31	51
	6	88563	663467	7.49	4.59	0.75	7	10	13	16	22	31	49
	7	90239	692028	7.67	4.69	0.75	7	10	14	16	22	31	50
	8	103465	774255	7.48	4.57	0.77	7	10	13	16	22	30	51
MA	3	93187	682948	7.33	4.55	0.69	7	10	13	16	21	33	65
	4	96897	710484	7.33	4.50	0.67	7	10	13	15	21	32	66
	5	109289	754384	6.90	4.39	0.67	6	9	13	15	21	31	66
	6	88721	575314	6.48	4.14	0.69	6	9	12	14	19	29	56
	7	90203	661635	7.33	4.55	0.70	7	10	13	16	21	31	62
	8	98389	722529	7.34	5.13	0.73	7	10	14	16	23	32	58
SC	5	109062	1114742	10.22	5.96	0.75	9	13	18	21	29	42	72
	8	96555	872413	9.04	6.74	0.79	8	13	18	21	29	42	73
SS	5	108427	1111332	10.25	6.11	0.76	9	13	18	21	30	44	73
	8	102837	1155026	11.23	6.81	0.82	10	15	20	24	33	48	72

Table 3. State Summary Statistics for Wrong-to-Right (WTR) Erasure Counts by Content/Grade Paper-Pencil

Content	Grade	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
							50	75	90	95	99	99.9	
ELA	3	43705	28096	0.64	0.97	0.79	0	1	2	2	4	7	18
	4	39553	16318	0.41	0.88	0.79	0	1	1	2	3	6	42
	5	24396	9487	0.39	0.80	0.78	0	1	1	2	3	6	20
	6	41410	21373	0.52	0.91	0.80	0	1	2	2	4	7	18
	7	39607	17498	0.44	0.85	0.78	0	1	1	2	4	7	22
	8	24722	11254	0.46	0.89	0.81	0	1	1	2	4	7	39
MA	3	43705	28714	0.66	1.11	0.82	0	1	2	3	5	8	17
	4	39553	24552	0.62	1.11	0.82	0	1	2	3	5	8	37
	5	24396	13381	0.55	0.99	0.82	0	1	2	2	4	7	14
	6	41410	26496	0.64	1.09	0.82	0	1	2	3	5	8	29
	7	39607	17127	0.43	0.89	0.82	0	1	1	2	4	7	13
	8	24722	10415	0.42	0.93	0.83	0	1	1	2	4	7	12
SC	5	24396	15699	0.64	1.17	0.83	0	1	2	3	5	9	16
	8	24722	9951	0.40	1.01	0.83	0	0	1	2	4	9	27
SS	5	24396	17026	0.70	1.25	0.83	0	1	2	3	5	10	22
	8	24722	18426	0.75	1.29	0.84	0	1	2	3	6	11	25

Table 4. State Summary Statistics for Wrong-to-Right (WTR) Erasure Counts by Content/Grade Online

Content	Grade	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
							50	75	90	95	99	99.9	
ELA	3	92882	305708	3.29	2.43	0.69	3	5	6	8	11	16	42
	4	96534	305395	3.16	2.40	0.71	3	4	6	8	11	16	32
	5	108829	352477	3.24	2.46	0.71	3	4	6	8	11	17	38
	6	88563	311901	3.52	2.68	0.75	3	5	7	9	12	18	33
	7	90239	302902	3.36	2.64	0.75	3	5	7	8	12	18	30
	8	103465	377767	3.65	2.72	0.77	3	5	7	9	12	18	32
MA	3	93187	306277	3.29	2.35	0.69	3	5	6	8	10	15	48
	4	96897	311516	3.21	2.32	0.67	3	4	6	7	10	15	50
	5	109289	317556	2.91	2.15	0.67	3	4	6	7	9	13	50
	6	88721	249557	2.81	2.10	0.69	2	4	6	7	9	13.28	34
	7	90203	291823	3.24	2.32	0.70	3	5	6	7	10	14	42
	8	98389	309367	3.14	2.59	0.73	3	5	7	8	11	15	38
SC	5	109062	502234	4.61	3.24	0.75	4	6	9	11	15	23	56
	8	96555	389744	4.04	3.54	0.79	3	6	9	11	15	22	59
SS	5	108427	496452	4.58	3.35	0.76	4	6	9	11	15	26	63
	8	102837	540582	5.26	3.92	0.82	4	7	10	12	18	30	56

Table 5. Number of Schools Flagged for Erasure and WTR Erasure and WTR Analysis Paper-Pencil

Grade	Total Number of Schools	All Erasure Analyses		Wrong-to-Right Erasure Analyses	
		Number of Schools Flagged for at Least One Content Area	Number of Schools Flagged for All Content Areas	Number of Schools Flagged for at Least One Content Area	Number of Schools Flagged for All Content Areas
3	541	58	14	44	3
4	483	48	9	30	4
5	388	41	1	37	1
6	266	29	7	13	2
7	258	29	6	19	3
8	212	51	2	36	0

Table 6. Number of Schools Flagged for Erasure and WTR Erasure and WTR Analysis Online

Grade	Total Number of Schools	All Erasure Analyses		Wrong-to-Right Erasure Analyses	
		Number of Schools Flagged for at Least One Content Area	Number of Schools Flagged for All Content Areas	Number of Schools Flagged for at Least One Content Area	Number of Schools Flagged for All Content Areas
3	1125	114	25	53	8
4	1134	112	28	37	7
5	1190	162	6	76	2
6	552	69	12	27	4
7	545	71	19	34	7
8	545	88	6	76	2

Table 7. The number/percentage of schools that had zero flags for all erasures and wrong-to-right erasures Paper-Pencil

Grade	English/Language Arts			Mathematics			Science			Social Studies		
	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags
3	541	491	91%	541	497	92%						
4	483	442	92%	483	453	94%						
5	388	364	94%	388	368	95%	388	369	95%	388	368	95%
6	266	247	93%	266	244	92%						
7	258	240	93%	258	234	91%						
8	212	199	94%	212	179	84%	212	175	83%	212	201	95%

Table 8. The number/percentage of schools that had zero flags for all erasures and wrong-to-right erasures Online

Grade	English/Language Arts			Mathematics			Science			Social Studies		
	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags	No. of Schools	No. Of Schools with Zero Flags	% of Schools with Zero Flags
3	1124	1042	93%	1125	1032	92%						
4	1134	1046	92%	1134	1054	93%						
5	1190	1136	95%	1189	1077	91%	1188	1097	92%	1188	1121	94%
6	552	514	93%	552	491	89%						
7	545	494	91%	545	479	88%						

Table 9. The number/percentage of schools that had less than 1% of the classes flagged for all erasures and wrong-to-right erasures Paper-Pencil

English/Language Arts			Mathematics			Science			Social Studies		
No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades
992	975	98%	992	968	98%	575	525	91%	575	548	95%

Table 10. The number/percentage of schools that had less than 1% of the classes flagged for all erasures and wrong-to-right erasures Online

English/Language Arts			Mathematics			Science			Social Studies		
No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades	No. of Schools	No. of Schools with <1% Flag Across Grades	% of Schools with <1% Flag Across Grades
1742	1726	99%	1742	1709	98%	1666	1519	91%	1671	1573	94%

Appendix D: 2016 DRC EOC Answer Change Executive Report

Analysis of Answer Changes

Submitted by DRC

October 2017

With the high-stakes nature of large-scale assessments such as the Milestones End of Course (EOC), there are times when students' responses, and hence their scores, may not be a true representation of their own abilities. Various activities may take place, such as a student copying from another student's paper, students receiving inappropriate assistance before or during testing, or students' responses altered after testing. To maintain the integrity of the Milestones EOC and the validity of the results, it is important that any such instances be discovered.

The present study investigated student responses on the Ninth Grade Literature & Composition, American Literature & Composition, Coordinate Algebra, Analytic Geometry, Algebra I, Geometry, Biology, Physical Science, United States History, and Economics tests of the 2017 Spring Milestones EOC that a) an answer choice was replaced by a different answer choice and b) changed from a wrong answer to a right answer (wrong-to-right).

It should be emphasized that results from the erasure analyses performed in 2017 should only be used to identify potential problems within individual classrooms. That is, these types of analyses must be supported by additional, collateral information before conclusions regarding any improprieties are reached.

Answer Changes for Paper Administrations

The GA Milestones EOC paper-pencil answer documents were processed using high speed 5000i optical scanners which reliably captured document images and optical mark read data. The sophisticated proprietary scoring software system, specifically Optical Mark Recognition (OMR) software, reviews the integrity of each batch of documents scanned according to pre-defined guidelines and services.

The OMR software provides a mechanism for identifying multiple-marks and identification of erasures for scanned data to support answer change analysis. The basis of the answer change analysis is to count erasures for multiple-choice items where two or more responses have been made with a specified intensity. Erasure analyses provide a mechanism to differentiate between three kinds of answer changes: a) wrong-to-wrong, b) right-to-wrong and c) wrong-to-right. Capturing the frequency of answer changes from wrong-to-right can be useful for identifying potential instances of cheating at the student level. Erasure analyses results can be grouped to tentatively identify problems at the classroom and school levels.

Answer Changes for Online Administrations

The test administration software that delivers the Georgia Milestones assessment system, INSIGHT, captures answer changes during online testing sessions. Similar to paper based administrations where answer changes are determined by examining erasure marks, the INSIGHT system records changes to answers within an online test administration that are made either before

leaving an item or upon returning to the item and making a change. Answer change analyses for students testing online also focuses on the three kinds of changes: a) wrong-to-wrong, b) right-to-wrong, and c) wrong-to-right. As with paper based erasure analyses, capturing the frequency of answer changes from wrong-to-right can be useful for identifying potential instances of cheating at the student level in online testing. Analyses results can be grouped to tentatively identify problems at the classroom and school levels.

Method

The basis for the answer change analysis is to count erasures in items where an answer choice was erased and replaced with another answer choice; online, an item was selected and then later changed to a different answer choice. Herein, both actions are referred to as an erasure. Often the data captured is useful for identifying cases of cheating. During erasure analysis, two sets of erasures were analyzed: all erasures and wrong-to-right erasures where an incorrect answer choice was erased and replaced with the correct answer choice. Operational and field test items were used for the answer change analyses implemented for the 2017 Georgia Milestones.

The basic idea underlying the procedure is a statistical test of the null hypothesis (H0) that the mean number of erasures for a class constitutes a random sample from the state distribution of erasures. The hypothesis is tested against the (right-sided) alternative (H1) that the mean number is too high to be explained by random sampling. Classes for which H0 has to be rejected are flagged for further scrutiny. A well-known central limit theorem in statistics tells us that the sampling distribution of the mean number of erasures for class i (m_i) is asymptotically normal with mean and standard deviation (SD)

$$mean(m_i) = \mu \qquad (1)$$

$$SD(m_i) = \frac{\sigma}{\sqrt{n_i}} \qquad (2)$$

where n_i and m_i denote the size and mean number of erasures for class i, respectively. In addition, μ and σ denote the mean and the SD of the distribution of the number of erasures of the population of individual students in the state of Georgia.

The classes were flagged if their m_i was larger than $\mu + 5\frac{\sigma}{\sqrt{n_i}}$. Statistically, the flagging criterion set at or above 5σ is conservative. The standard normal table shows that under random sampling the (asymptotic) probability of a sample mean being more than five SDs above the population mean is around 0.00003. However, rejection of H0 only tells us that the observed mean number of erasures is unlikely to be the result of random sampling.

It is evident in the formula that the class flagging criterion for each class is adjusted for the number of test takers in a classroom. For example, if the state mean and SD of erasure count are 1.73 and 2.11, respectively, the flagging criterion for a class size of 20 is adjusted to 4.11 ($1.73 + 5\frac{2.11}{\sqrt{20}} = 4.11$).

This adjustment ensures that the flagging criterion is equally stringent for classes with considerably different numbers of test takers. In addition, minimizing the probability of false positive (Type I) errors in this statistical test is crucial in this analysis.

Results

Tables 1 and 2 reports the state summary of erasure counts for paper-pencil and online respectively. The tables include the number of students, the total number of all types of erasures, the mean and the SD of all types of erasures, the correlation between all erasures and wrong-to-right erasures, the number of erasures at the 50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th} percentiles, and the maximum number of all types of erasures. The mean number of paper-pencil erasures across all courses ranged from 0.61 to 1.86, and mean number of online answer changes ranged from 5.15 to 10.67 for the 2017 Spring Milestones EOC. In other words, approximately 1 to 2 answer changes were made per student paper-pencil answer sheet on average, and 5 to 11 answer changes were made per online student assessment. The erasure count at specific percentile points (50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th}) is also reported. The erasure count at the 95^{th} percentile point was between 3 and 7 on paper-pencil answer sheets, and between 12 and 23 online.

Tables 3 and 4 report the state summary of wrong-to-right erasure counts for paper-pencil and online respectively. The tables include the number of students, the number of wrong-to-right erasures, the mean and the SD of wrong-to-right erasures, the correlation between all erasures and wrong-to-right erasures, the number of wrong-to-right erasure at the 50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th} percentiles, and the maximum number of wrong-to-right erasures. As can be expected, the mean wrong-to-right erasure count and the count at the specific percentile points were lower than those obtained from all erasure counts. The mean number of paper-pencil wrong-to-right erasures ranged from 0.34 to 0.97, and mean number of online answer changes ranged from 2.85 to 5.05 for the 2017 Spring Milestones EOC. In other words, approximately 0 to 1 wrong-to-right answer changes were made per student paper-pencil answer sheet on average, and approximately 3 to 5 wrong-to-right answer changes were made per online student. The wrong-to-right erasure count at specific percentile points (50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th}, and 99.9^{th}) is also reported. The wrong-to-right erasure count at the 95^{th} percentile point was between 2 and 4 on paper-pencil answer sheets, and between 7 and 12 online.

Tables 5 and 6 present a summary of the number of schools flagged for total erasures and wrong-to-right erasures based on Milestones EOC paper-pencil and online respectively. Table 7 presents a summary of all schools with at least one class taking the Milestones EOC for at least one subject.

Discussion

With respect to the erasure analyses, the following caveats are always applicable:

1. The normal distribution holds only for large classes; for smaller classes the result is approximate.

2. Rejection of $H0$ does not necessarily imply cheating. Alternative explanations are possible.

3. The flagging criterion should thus be taken as a stimulus to look for additional evidence and find out what happened in the school.

4. The groups of students taking the tests online and paper are not equivalent.

Comparing the magnitude of answer changes between testing modes cannot be supported given that the groups likely differ in ability and other key background characteristic.

This answer change analysis is considered a check for unusual numbers of answer changes to student responses. Without additional layers added to the analysis, this kind of check only addresses the possibility, not the certainty, of teachers or administrators altering the responses of students. The 2017 erasure analyses represent an important step in helping to maintain the integrity of future administrations of the Milestones EOC.

Table 1. State Summary Statistics for All Types of Erasure (ERA) Counts by Course Paper-Pencil

Course	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
						50	75	90	95	99	99.9	
9LCO	3780	4070	1.08	1.56	0.80	1	2	3	4	7	11	18
AGEO	1318	1853	1.41	2.14	0.80	1	2	4	5	10	15	16
ALG1	2217	2808	1.27	2.07	0.80	1	2	3	5	10	19	23
AMLC	3549	2155	0.61	1.14	0.84	0	1	2	3	5	9	12
BIOL	1811	2090	1.15	2.17	0.86	0	2	3	5	9	21	33
CALG	2009	2879	1.43	2.03	0.73	1	2	4	5	9	17	18
ECON	325	605	1.86	3.40	0.89	1	2	5	7	17	28	31
GEOM	2341	2746	1.17	1.88	0.84	1	2	3	4	9	15	25
HIST	1423	2052	1.44	2.20	0.85	1	2	4	6	9	14	38
PHSC	1289	2120	1.64	2.45	0.81	1	2	4	6	11	17	33

Table 2. State Summary Statistics for All Types of Erasure (ERA) Counts by Course Online

Course	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
						50	75	90	95	99	99.9	
9LCO	111787	623656	5.58	3.70	0.81	5	7	10	12	17	24	5
AGEO	19866	167594	8.44	4.87	0.72	8	11	15	17	23	33	8
ALG1	98982	859857	8.69	5.11	0.77	8	11	15	18	24	35	8
AMLC	96484	496647	5.15	3.58	0.82	4	7	10	12	16	23	4
BIOL	104894	1118903	10.67	6.55	0.82	9	14	19	23	32	45	9
CALG	22119	202880	9.17	5.17	0.75	8	12	16	19	25	36	8
ECON	57341	602358	10.50	6.57	0.81	9	14	19	23	31	46	9
GEOM	78288	564605	7.21	4.44	0.74	6	10	13	15	21	31	6
HIST	94799	993324	10.48	6.67	0.80	9	14	19	23	32	45	9
PHSC	85134	829826	9.75	5.57	0.78	9	13	17	20	27	38	9

Table 3. State Summary Statistics for Wrong-to-Right (WTR) Erasure Counts by Course Paper-Pencil

Course	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
						50	75	90	95	99	99.9	
9LCO	3780	2112	0.56	0.95	0.80	0	1	2	2	4	6	8
AGEO	1318	836	0.63	1.12	0.80	0	1	2	3	5	8	8
ALG1	2217	1232	0.56	1.04	0.80	0	1	2	2	4	9	11
AMLC	3549	1196	0.34	0.73	0.84	0	0	1	2	3	6	7
BIOL	1811	990	0.55	1.17	0.86	0	1	2	3	5	11	16
CALG	2009	1221	0.61	1.09	0.73	0	1	2	3	5	10	13
ECON	325	316	0.97	1.88	0.89	0	1	3	4	8	14	14
GEOM	2341	1376	0.59	1.08	0.84	0	1	2	3	5	8	16
HIST	1423	1024	0.72	1.22	0.85	0	1	2	3	5	9	14
PHSC	1289	999	0.78	1.36	0.81	0	1	2	3	5	9	22

Table 4. State Summary Statistics for Wrong-to-Right (WTR) Erasure Counts by Course Online

Course	N	No. of Erasures	Mean	SD	Correlation between ERA and WTR	Number of Erasure by Percentiles						Max
						50	75	90	95	99	99.9	
9LCO	111787	334886	3.00	2.45	0.81	3	4	6	8	11	16	32
AGEO	19866	70607	3.55	2.58	0.72	3	5	7	8	11	16	25
ALG1	98982	379194	3.83	2.81	0.77	3	5	7	9	13	18	39
AMLC	96484	274999	2.85	2.38	0.82	2	4	6	7	11	15	27
BIOL	104894	529788	5.05	3.80	0.82	4	7	10	12	17	27	51
CALG	22119	90185	4.08	2.87	0.75	4	6	8	9	13	18	28
ECON	57341	285454	4.98	3.70	0.81	4	7	10	12	17	26	43
GEOM	78288	248278	3.17	2.41	0.74	3	4	6	8	11	16	33
HIST	94799	461724	4.87	3.72	0.80	4	7	10	12	17	27	54
PHSC	85134	387151	4.55	3.20	0.78	4	6	9	10	14	21	38

Table 5. Number of Schools Flagged for Erasure and WTR Erasure and WTR Analysis Paper-Pencil

Course	Total Number of Schools	All Erasure Analyses		Wrong-to-Right Erasure Analyses	
		Number of Schools Flagged	Percent of Schools Flagged	Number of Schools Flagged	Percent of Schools Flagged
9LCO	105	1	0.95	1	0.95
AGEO	51	0	0.00	0	0.00
ALG1	60	1	1.67	1	1.67
AMLC	88	2	2.27	1	1.14
BIOL	93	1	1.08	0	0.00
CALG	58	0	0.00	0	0.00
ECON	37	0	0.00	0	0.00
GEOM	47	1	2.13	0	0.00
HIST	79	0	0.00	0	0.00
PHSC	86	0	0.00	0	0.00

Table 6. Number of Schools Flagged for Erasure and WTR Erasure and WTR Analysis Online

Course	Total Number of Schools	All Erasure Analyses		Wrong-to-Right Erasure Analyses	
		Number of Schools Flagged	Percent of Schools Flagged	Number of Schools Flagged	Percent of Schools Flagged
9LCO	569	15	2.64	11	1.93
AGEO	210	3	1.43	4	1.90
ALG1	653	7	1.07	22	3.37
AMLC	449	6	1.34	0	0.00
BIOL	477	9	1.89	16	3.35
CALG	248	4	1.61	4	1.61
ECON	433	10	2.31	2	0.46
GEOM	396	10	2.53	7	1.77
HIST	451	8	1.77	3	0.67
PHSC	694	4	0.58	16	2.31

Table 7. Number of Schools Flagged (WTR) in any Course for Milestones EOC

Course	Total Number of Schools	Number of Schools Flagged (WTR)	% of Schools Flagged (WTR)	Number of Schools Not Flagged (WTR)	% of Schools Not Flagged (WTR)
Paper Tests	202	3	1.49	199	98.51
Online Tests	859	58	6.75	801	93.25

Appendix E: 2017 DRC EOG Unusual Response Executive Report

Modified Jacob and Levitt Analyses

Submitted by DRC

December 2017

With the high-stakes nature of large-scale assessments such as the Milestones End of Grade (EOG), there are times when students' responses, and hence their scores, may not be a true representation of their own abilities. Various activities may take place, such as a student copying from another student's paper, students receiving inappropriate assistance before or during testing, or students' responses altered after testing. To maintain the integrity of the Milestones EOG and the validity of the results, it is important that any such instances be discovered.

The present study investigated student responses on the English Language/Arts and Mathematics assessments of the 2017 Spring Milestones EOG using a modified application of the analysis described by Jacob and Levitt (2003). This method is designed to identify schools with both large score fluctuations across years and unexpected patterns in student answers.

Method

This method included a combination of two indices: (1) unexpected test score fluctuations across years using a cohort of students and (2) unexpected patterns in student answers. The first indicator ranked each school's average test score gains relative to other schools' gains for a particular grade and subject. The second index ranked schools regarding unexpected patterns in student answers. The student answer pattern analyses were examined in four ways. Schools were ranked on four measures that were combined to provide an overall index of unexpected patterns in student answers. The analyses identified the:

- most unlikely block of identical answers,
- degree of correlation in student answers across the test,
- degree of variance in the correlation of responses across items, and
- extent to which student responses were congruent with respect to item difficulty and student ability.

It was possible for a school to experience a large increase in tests scores due to, for example, the introduction of a new curriculum or after-school program. It was also possible for unexpected answer patterns to appear without inappropriate behavior having occurred. For these reasons, a school had to be in the 95th percentile on both indices to be flagged. Having to be within the 95th percentile on both indices, in this context, was a way to limit the number of schools being identified due to Type I error. In this case, a Type I error would be incorrectly identifying a school for suspicious behavior. By ranking the schools on both indices Type I

error is made smaller than if using only one index and the schools are protected from being falsely identified.

INDEX 2

The second index ranked schools regarding unexpected patterns in student answers. The student answer pattern analyses were examined in four ways. Schools' rankings on the four measures were combined to provide an overall index of unexpected patterns in student answers. The analyses identified the following measures:

1. most unlikely block of identical answers,
2. highly correlated answers across the test,
3. degree of variance in the correlation of responses across items, and
4. cases in which students miss easy items while answering difficult items correctly.

Measure 1 identifies the most unlikely block of identical answers given by students on consecutive items using a multinomial logit model. The likelihood of each student choosing each possible answer on every item is calculated based on the student's current year's test responses and previous year's test scores. All combinations of students and consecutive items are compared to find the block of identical answers that were least likely to have arisen by chance.

First, a multinomial logit model is used to calculate every student's likelihood on each item:

1. $P(Y_{isc} = k) = \frac{e^{\beta_k x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}}$, $\quad k = 1, \ldots, J$

where *s* is the student, *c* is the cohort, *k* is the selected answer option, *J* is the total number of options, and *x* is the vector of previous year's test scores.

Second, the likelihood of a student's answer for item *i* is found by selecting the appropriate value from Equation 1:

1. $P_{isck} = \frac{e^{\beta_k x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}}$, where *k* is the response actually chosen by student *s* on item *i*.

Third, identify strings of items, *m* to *n*, for which the cohort gave identical responses; then the likelihood of this string for student *s* is the product of the item likelihoods from Equation 2:

2. $P_{sc}^{mn} = \prod_{i=m}^{n} P_{isck}$.

Fourth, the product across all students in the cohort who had identical responses in the string is

3. $\tilde{P}_{sc}^{mn} = \prod_{s\in\omega} P_{sc}^{mn}$, where □□is the group of students who have identical responses to items *m* to *n*. The calculations are repeated for all strings of five consecutive operational items.

Finally, the minimum value of this measure for each cohort is recorded as Measure 1.

4. *Measure 1* = $min_s\left(\tilde{P}_{sc}^{mn}\right)$

The smallest values are associated with more improbable answer strings within a cohort.

Measure 2 examines the degree of correlation in student responses across the test, particularly for unexpected answers. It was based on the assumption that teachers who cheated will have students with highly correlated answers. Measure 2 is the average of the item residual values. Higher values indicate cohorts with highly correlated answers.

5. $e_{jisc} = \begin{matrix} 0 - P_{isck} & if\ j \neq k \\ 1 - P_{isck} & if\ j = k \end{matrix}$.

where e_{ijsc} is a residual for item *i* for *j*th item answer option for cohort student sc, and P_{isck} is the probability of the cohort student, *sc*, select the answer option *k* for item *i*. Then residual for each option are summed across students within the cohort:

6. $e_{jic} = \sum_s e_{jisc}$.

This sum of residual, e_{ijsc} should be approximately zero if there is no within-school/administrator correlation in the way students responded to item *i*, response *j*. The residual for all possible responses were summed for each item within school/administrator. Then, the sum of squared residuals is divided by the squared number of students to normalize for cohort size. This is analogous to average test residual for cohort.

1. $v_{ic} = \frac{\sum_j e_{jic}^2}{n^2}$.

Measure 2 is average value of v_{ic} across all items for cohort, and a measure of correlation across test.

2. *Measure 2* = $\overline{v_c} = \frac{\sum_i v_{ic}}{ni}$.

Measure 3 calculates the variance of test residual for cohort, v_{ic}. If a teacher cheated by changing or providing answers for multiple students on selected questions, the within-cohort residual deviation on those particular questions will be extremely high, while the within-cohort residual deviation on other questions is likely to be normal. Thus, a large degree of variance in the residual of responses across items would occur. The variance is calculated as follows:

3. *Measure 3* = $\sigma_{v_c} = \frac{\sum_i (v_{ic} - \bar{v}_c)^2}{ni}$, where *ni* is the number of items on the exam.

Measure 4 compares the answers of students within a cohort to the answers from other students with same total scores in the sample. It detects students who missed easy items while answering difficult items correctly. For each student in cohort, deviation of item response from all students with aggregate score *A* is computed as follows:

4. $z_{sc} = \sum_i (q_{isc} - \bar{q}_i^A)^2$,

where superscript *A* indicate all students with aggregate total score *A*; q_{isc} equals one if the student answers item *i* correctly and zero otherwise; and $\bar{q}_i^A$is the proportion of A students answering item *i* correctly. Squared deviation is summed for items, making test level deviation for each student. Measure 4 is average value of test level deviation from students with total score A for each cohort.

5. *Measure 4* = $\frac{\sum_S (z_{sc} - \bar{z}^A)}{n_{sc}}$

High values of this index indicate the answer from a large number of students in the cohort deviated from students with same total scores in other cohorts.

With all four measures computed, the schools/groups are ranked on each of the four measures. For this, quantile regression is used to remove the effect of sample size inherent in the indexes used by Jacob and Levitt (2003). While *least squares* regression minimizes the sum of squared deviations from the regression line and passes through the mean, quantile regression on the median, for example, minimizes the sum of absolute values of the deviations from the line which is the median (Koenker, 2005). The result is that exactly 50% of the data points will be above the line and 50% below. It is also true that 50% of the points are expected to be above the line for any value of the independent variable. In other words, the quantile regression line is the median of the dependent variable conditional on the independent variable.

In the quantile analysis, the Modified Jacob and Levitt measures were the dependent variables and school enrollment was the independent variable. Quantile regression can be generalized to any percentile, typically denoted as τ between 0 and 99. The preceding discussion used the median ($\tau = 0.5$). By iterating on τ using

the *R* package *quantreg* (R Development Core Team, 2003; Koenker, 2011), it is possible to determine the percentile rank for any value of the measure conditional on school enrollment. The percentile ranks for each cohort, on each measure, are then combined to form the Index 2 as follows:

1. $Index\ 2 = Measure1\_rank^2 + Measure2\_rank^2 + Measure3\_rank^2 + Measure4\_rank^2$.

COMBINING INDEX 1 and INDEX 2

It was possible for a school to experience a large increase in tests scores due to, for example, the introduction of a new curriculum or after-school program. It was also possible for unexpected answer patterns to appear without inappropriate behavior having occurred. For these reasons, a school had to be in the 95th percentile on both indexes to be flagged. Having to be within the 95th percentile on both indexes, in this context, was a way to limit the schools being identified due to Type I error.

Results

Unlike the other forensic methods, for a school to be identified by this analysis it had to have a minimum sample size of eight and be in the 95th percentile for both unexpected score changes and unexpected patterns in student response indices. It is expected that fewer schools would be identified with this method since it is designed to detect only extreme cases of potential misconduct.

As can be seen in Table 1, the number of schools flagged in ELA ranged from 1 schools to 11 and from 4 schools to 19 schools in Mathematics. Grade 3 was excluded from this method since there was no data from an immediately preceding year to use in the analysis. Similar to last year, more schools were flagged within the Mathematics analyses. However, the overall rate of schools that are flagged within the analysis is similar to the prior year.

Table 1. Count of Schools Flagged within 2017 Jacob and Levitt Analyses

Grade	English Language Arts				Mathematics			
	Form A		Form B		Form A		Form B	
	Count	%	Count	%	Count	%	Count	%
3	-	-	-	-	-	-	-	-
4	9	0.73%	8	0.90%	19	1.54%	18	2.01%
5	11	0.90%	8	0.76%	19	1.55%	15	1.42%
6	8	1.42%	6	1.42%	14	2.50%	8	1.93%
7	1	0.18%	2	0.49%	9	1.65%	4	0.99%
8	2	0.36%	5	1.07%	9	1.66%	10	2.18%
TOTAL	31	0.75%	29	0.89%	70	1.70%	55	1.70%

Discussion

The goal of psychometric forensic analysis was to screen for test results that may have been spurious because valid inferences cannot be made from such test scores. The Jacob and Levitt analyses reported here are just one set of forensic analyses implemented within the Georgia Milestones and should not be considered in isolation.

Note that the schools were flagged within the Jacob and Levitt analyses based on statistical evidence alone. If flagged, that does not necessarily mean that the schools engaged in inappropriate testing activity. However, the statistical evidence does suggest that something aberrant or unusual occurred and, barring a simple explanation, warrants further exploration.

All forensic results should be used with caution, and data for schools and grades within schools and their results may serve as good starting points for the evaluation of potential testing irregularities. The 2017 Jacob and Levitt analyses represent an important step in helping to maintain the integrity of future administrations of the Milestones EOG.

References

Jacob, B. & Levitt, S (2003). Rotten apples: An Investigation of the prevalence and predictors of teacher cheating, *The Quarterly Journal of Economics*, 118 (3), 843-877.

Koenker, R. (2005). *Quantile regression*. Cambridge, UK: Cambridge University Press.

Koenker, R. (2011). *Quantile regression in R: A vignette*. Retrieved November 1, 2011, from http://www.econ.uiuc.edu/~roger/research/rq/vig.pdf.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Retrieved from the Vienna University, Institute for Statistics and Mathematics Web site: http://www.R-project.org.