Minnesota Department of Education

Report on MCA Contractor Performance

Fiscal Year 2016

Report

To the

Legislature

As required by

Minnesota 2015 Special Session Law,

H.F. No. 1, Chapter 3, Article 3, section 14

COMMISSIONER:

Brenda Cassellius, Ed. D.

Report on MCA Contractor Performance

February 10, 2016

FOR MORE INFORMATION CONTACT:

Jennifer Dugan

Statewide Testing

(651) 582-8654

jennifer.dugan@state.mn.us

FY 16

Report to the Legislature

As required by

Minnesota

2015 Special

Session Law

H.F. No. 1, Chapter 3, Article 3, section 14

Cost of Report Preparation

The total cost for the Minnesota Department of Education (MDE) to prepare this report was approximately $400. Most of these costs involved staff time in analyzing data and preparing the written report. Incidental costs include paper, copying and other office supplies.

Estimated costs are provided in accordance with Minnesota Statutes 2015, section 3.197, which requires that at the beginning of a report to the Legislature, the cost of preparing the report must be provided.

Legislative Charge

Under H.F. No.1, Article 3, section 14 of the 2015 Minnesota legislative special session, MDE is required to provide a report regarding Pearson's performance during the spring 2015 online test administration.

REPORT ON MCA CONTRACTOR PERFORMANCE

By February 10, 2016, the commissioner of education must report to the legislative committee with jurisdiction over education finance and policy describing the performance of the contractor providing the Minnesota Comprehensive Assessments to the state, including any payment adjusted to reflect the contractor's failure to perform according to the terms of the state contract, findings from the qualified independent contractor under section 13, and any other information about online administration of the Minnesota Comprehensive Assessments the commissioner wishes to include in the report.

Executive Summary

Technical difficulties, degraded performance, and service outages were experienced during the spring 2015 online testing window. Due to the nature of many of these outages, MDE's and Pearson's actions during the 2015 administration window were in an effort to allow the online test window to come to as smooth of a close as possible. MDE and Pearson adjusted plans in preparation for the 2015-2016 administrations based on lessons learned from the spring 2015 online administration experiences. These efforts included technological as well as training and support adjustments described below. Based on the administration experiences and reports during the first few months of the Optional Local Purpose Assessment (OLPA), the technological and training adjustments have been largely effective and supportive of a relatively smooth online administration.

Overview of 2015 Testing Window and Technical Difficulties

The 2015 online testing window opened March 9, 2015, for all MCA and MTAS assessments. These assessments are administered consistent with federal and state regulations.

From March 9 – May 19, Pearson administered the following online assessments:

- Reading - 438,849 tests
- Math - 437,315 tests
- Science - 183,377 tests

MDE's Division of Statewide Testing monitors the Pearson Help Desk logs. In addition to the widespread interruptions described in greater detail below, MDE received reports of difficulties with online calculators loading or freezing, artwork rendering slowly, and other less pervasive technical difficulties.

Issue 1 – Degraded Performance due to Server Failures

Starting on Tuesday, April 14, PearsonAccess users contacted Pearson and MDE reporting issues with the system availability and performance degradation. Users were experiencing slowness and unavailability of the PearsonAccess system. This included slowness in approving students to test and resuming student testing. Throughout the day, degraded performance

appeared to be a widespread issue. However, the online testing system, TestNav was not impacted and those students who were able to start the assessment were able to continue without issue.

Pearson determined that a failed primary drive on one of two Fatwire servers caused the degraded performance. Fatwire is the content management system for PearsonAccess. It serves the majority of the common artifacts (images, banners, base page layouts) used in the PearsonAccess system. Pearson added two additional servers (for a total of four in production) to add capacity and additional security in the event of future hardware failures in Fatwire.

MDE and Pearson sent email communications to district assessment and technology staff throughout the day to inform the field of the situation.

On Wednesday, April 15, Pearson confirmed that the issues from the previous day were resolved and the system was fully operational. For about an hour on April 15, users reported some degraded performance of slowing or stalling of the system.

The issues that occurred for an hour on April 15 impacted three different servers at three different times. Each event lasted between three and 10 minutes and was characterized by the given server falling into a state where it was unresponsive. These issues were unrelated to the hardware issues experienced the previous day. Pearson increased server capacity to a total of five in production. Pearson also added very sensitive monitoring code to provide an alert within 15 seconds of a server exhibiting signs of becoming unresponsive.

By mid-morning the system was back to fully operational. MDE and Pearson sent email communications to district assessment and technology staff throughout the morning.

At no time during these two days was there any risk to student responses or the transmission of data to Pearson. The components of the system that manage scoring, reporting and On Demand Reports were unaffected.

Issue 2 – DDoS Attack (volume-based)

On Tuesday, April 21, several districts reported students encountering difficulty testing early in the afternoon. Pearson confirmed its servers experienced a Distributed Denial of Service (DDoS) attack from approximately 2:30-3:00 p.m. This malicious attack, from an outside party, targeted the TestNav.com portal and affected the online testing platform.

DDoS attacks are increasing both in frequency and in size and pose a significant risk to service delivery for online testing. DDoS attacks are predicated upon the assumption that resources are finite -- and by exhausting these finite resources, a malicious attacker can successfully deny service for legitimate traffic attempting to use these services. A DDoS attack is not an effort to hack into Pearson's system, and at no point was student data at risk or compromised.

MDE and Pearson sent email communications to district assessment and technology staff that afternoon informing them of the situation. Pearson responded to the DDoS attack by adjusting the system's "approval" process in order for a computer to access the server. They also implemented Akamai Prolexic Services from DDoS vendor, Akamai, to route all traffic to Pearson's servers through Akamai for scrubbing.

That afternoon, MDE determined the best course of action was to suspend testing Wednesday, April 22. This was a result of Pearson requiring additional time to provide sufficient documentation and assurance when testing could resume. During this time, Pearson worked diligently to resolve the issues and establish a plan to manage this type of attack. Since district staff needed sufficient time to make adjustments, testing was delayed for one day.

On Thursday, April 23, testing resumed. With the announcement to resume testing, MDE also added two days to the testing windows.

Issue 3 – DDOS Attack (protocol-based)

The morning of Wednesday, May 13, Pearson experienced a different type of (DDoS) attack on their servers. Unlike the previous attack, this attack was not specific to the volume of traffic to Pearson's servers, but instead the nature of the traffic being sent. In response, Pearson implemented new filtering parameters to scrub offending traffic without affecting that which was legitimate. To improve response times to an attack, Pearson entered into agreement with Akamai to allow them to block content on Pearson's behalf ("Always On" Routing). To date, no legitimate content has been blocked.

MDE suspended testing for the remainder of Wednesday and Thursday.

On Friday, May 15, testing resumed.

Greater detail of the technical difficulties and timelines during the Spring 2015 administration are detailed in Appendix A.

Actions Taken After Testing Window

Independent Statistical Analysis

Due to technical disruptions, MDE requested Human Resources Research Organization (HumRRO) to conduct an independent analysis. This action is consistent with industry best practices as well as past practice in Minnesota when technical difficulties were experienced during the 2013 test administration. There was also a requirement in H.F. 1 Sec. 13 from the 2015 legislative special session.

HumRRO provided a statistical investigation of the potential impact of computer disruptions on student scores. The analysis focused on students testing the afternoon of April 21. Due to the DDoS attack students had limited or no access to TestNav. Statistical analyses used propensity matching, which uses several variables, to match the disrupted students with a group of students who tested on non-affected administration dates.

There was no statistical evidence to suggest that the disruption, on average, negatively affected student scores. For some grades the disruption was beneficial; for others it was not. Most grades showed no impact at all. Any observed effects were small. Based on the statistical analyses in the report, the students observed score is the best estimate of achievement for 2015.

The full report is available in Appendix B.

Payment adjustment and Additional Services

1. Pearson's overall contract price was reduced by $1 million.
2. In the 2015-2016 school year, Pearson will provide funding support for the administration of The ACT for up to 10,000 students to help ensure all eligible students are given the opportunity to participate. This is necessary because the funding allocated for district reimbursement of The ACT may not be sufficient to reimburse all interested and eligible grade 11 and 12 students.
3. Pearson will provide Perspective at no cost to the state for the duration of the current contract and any renewals. Perspective is an online tool with learning resources aligned to Minnesota academic standards. Students can access targeted resources based on performance of Minnesota assessments. Teachers can assign resources to individual students or groups of students.
4. Pearson will provide WritetoLearn to students in 8th and 10th grade at no cost to the state for the duration of the contract and any renewals. WriteToLearn is an online writing tool with student feedback. Students are able to compose and revise a response to prompts.
5. Pearson will pay for an alignment study to determine the testing options for a new high school writing exam required by the 2015 Legislature.
6. Provide additional training and support to staff and districts.
7. Implement various improvements to technical support and reporting.

Table 1: Table of Payment Adjustments and Additional Services

Description	Amount
Reduction	$1,000,000
Administration of The ACT	$565,000
Pearson Perspective (annually)	$111,968
WritetoLearn (annually)	$1,100,000
Options study for writing exam	TBD when study is conducted; ~$50,000

Future Solutions for Spring 2016 MCA Testing

MDE has consulted with MN.IT to prepare for the spring 2016 administration. MDE, MN.IT and Pearson met to discuss the events of the spring 2015 administration as well as discuss plans for spring 2016 administration. During the conversation, the participants agreed that utilizing the cloud-based systems was the best way to minimize vulnerabilities of future DDoS attacks as well as utilize the latest in technology and technology industry best practices.

Pearson built its next-generation systems on Amazon Web Services (AWS) Cloud platform, which provides additional layers of security and built-in protections against DDoS attacks.

- The TestNav8 test engine is cloud-based, app-based, browser-based and supported on Tablets and Chromebooks. This test engine will be housed on a Virtual Private Cloud, a private network within a public cloud infrastructure, which provides significant scalability to scale up during peak testing.
- PearsonAccessNext, the administrative system necessary for starting testing sessions, will also be cloud-based.

The massive scale and built-in protections afforded by Amazon will provide fundamental protection against large-scale attacks. In addition to the inherent protections against DDoS attacks, Pearson will also implement several additional protective services to further harden and protect its assessment delivery platforms from attacks. When designing a DDoS response posture for NextGen assessment services in AWS, Pearson focused on four key initiatives:

1. Minimize the Attack Surface Area
2. Scale to Absorb the Attack
3. Safeguard Exposed Resources
4. Baseline Normal Behavior

Minimize the Attack Surface Area

A key strategy in reducing the exposure of Internet-facing services is to reduce the number of entry points into those services. Pearson will take advantage of two main approaches in order to minimize the attack surface for its assessment delivery services: Virtual Private Cloud (VPC) and Elastic Load Balancers (ELB). The combination of these two key services reduces opportunities for attackers while maintaining the ability to scale these entry points on demand.

Pearson will minimize the opportunities an attacker has to target the applications through the use of Virtual Private Cloud (VPC).

- Effectively “hides” instances from the Internet ensuring nonpublic instances are only available on a private subnet with private DNS entries.
- Minimizes public entry points into assessment platforms through the configuration of security groups and network access control lists (ACLs).
- Allows for the control of inbound and outbound traffic to application instances by specifically allowing communication only on the ports and protocols required for the applications. Access to any other port or protocol is automatically denied.

The attack surface will also be minimized through the widespread implementation of ELBs in Amazon AWS.

- ELB allows Pearson the ability to place all compute instances behind a single ELB tier while only requiring that the ELB itself be exposed to the public Internet.
- ELBs are able to auto-scale to handle fluctuations in demand.
- Amazon has several DDOS protections built into their ELB load balancers that block many traditional DDoS attacks by default. ELB is designed to pass only well-formed connections to the web applications on ports and protocols that are explicitly specified. This provides an additional layer of DDoS resiliency for the assessment delivery services.

Scale to Absorb the Attack

DDoS attacks are about scale. Most attackers achieve their purpose by sending a level of traffic that the application cannot accommodate. By implementing an architecture that can out scale the attack, Pearson can create a barrier that requires more time and resources on the part of the attacker, thereby making its applications more resilient. With AWS, Pearson will take advantage of scaling in three areas:

- Instance Scaling: Individual pieces of the infrastructure have the ability to rapidly scale as independent units. Additional CPU or memory capacity can be added to a server instance very quickly with no impact to the end-user.
- Environment Scaling: Logic is implemented to automatically adjust the number of instances serving a particular application based on fluctuation of load and incoming request rates. Any given environment can rapidly scale from just a few servers to several hundred in a matter of minutes.
- AWS Global Service Scaling: Pearson leverages several global services from AWS that further increase scalability in its infrastructure. Content is served through the CloudFront Content Distribution Network (CDN). With more than 20 unique points of presence in the U.S. CloudFront is able to scale well beyond what any individual application environment would be able to support and provides further buffer against malicious attacks.

Safeguard Exposed Resources

Minimizing the attack surface by “hiding” all instances from the public Internet is not practical. There will always be some number of public entry points necessary in order to provide the required services for assessment delivery. For those points of access, where public exposure to the Internet is necessary, it is crucial to be able to provide additional safeguards against resource exposure. Pearson will add additional safeguards to ensure only legitimate traffic accesses its services through the implementation of Amazon CloudFront and Web Application Firewall (WAF).

Amazon CloudFront is a Content Distribution Network designed to place content closer to the user. This results in lower latency and overall improved web application performance for end user requests. Acting as a proxy for all requests, CloudFront carries with it the additional benefit of providing safeguards to resource exposure by:

- Requiring that all requests to Pearson infrastructure first go through CloudFront. This allows Pearson the ability to govern access of content by only allowing requests for content that first go through CloudFront.
- Allowing for restriction of access by geographical region or origin of request (Geo-Blocking).
- Employing additional filtering capabilities to ensure that only valid TCP connections and HTTP requests are made while dropping invalid requests.

Web Application Firewall (WAF)

DDoS attacks that happen at the application layer often target specific pieces of an application with much lower volumes of traffic. To mitigate these types of attacks and safeguard applications, Pearson will implement Web Application Firewall (WAF) capabilities to include the following capabilities:

- Block common exploits such as cross-site scripting and SQL injection prior to the requests making it to Pearson application services.
- Perform rate limiting of HTTP requests to protect against HTTP GET and POST floods aimed at targeting specific components in a web application.
- Perform HTTP level inspection to inspect requests and identify those that do not conform to normal patterns.

Baseline Normal Behavior

To successfully mitigate and defend against malicious DDoS attacks, it is imperative to know precisely when an application is under attack. While using basic “up/down” monitoring is one way to know when services are impacted, it is more beneficial to take a proactive approach and develop a sound understanding of normal traffic patterns in order to best detect when anomalous behavior is occurring. Pearson will develop a custom enterprise data warehouse capable of processing millions of metrics per minute from the infrastructure and applications. Real-time analysis and correlation identify deviations from expected levels and engineers are able to begin troubleshooting and triage at the first sign of behavior anomalies.

- Benchmarks are created from daily analysis and alerts are configured to alarm when usage patterns are beyond “expected levels”.
- Continuous metric collection and evaluation allows for automated response to changes in observed behavior. For example, alerts and actions can be set when CPU resources move beyond expected values for given time of day -- resulting in an automated response that adds additional CPU resources within minutes.
- Network traffic flow logs are evaluated in real-time for anomalous patterns and are configured to call out to responsible parties for response when significant deviations are measured.

Continued Mitigations to Combat DDoS Attacks

Given the growing number and scale of DDoS attacks, Pearson implemented a multi-service solution for DDoS detection and mitigation in Pearson-controlled data centers. These new protections are aimed at providing the most comprehensive mitigation strategy possible to ensure service availability throughout the online testing cycle. Two distinct protective strategies recently implemented are “Always on” traffic routing through distributed scrubbing centers and advanced on-premise DDoS protection appliances.

Always-On Routing through Distributed Scrubbing Centers
Pearson works with the worldwide leader in content distribution and DDoS protection services to provide cloud-based DDoS mitigation for Pearson’s services. Pearson routes all traffic flowing into the Pearson data center through several geographically distributed scrubbing centers. These scrubbing centers constantly observe all traffic destined for the Pearson data center and look for potential DDoS attack vectors, drop detected attack traffic and forward only clean traffic to the application origin. This solution moves the mitigation away from the Pearson data center and provides more than two Terabytes per second of bandwidth to absorb even the largest DDoS attacks. When unique challenges and protections are required, the security service is able to rapidly adapt and implement mitigations tailored to the attack through their dedicated 24/7 Security Operations Center.

Advanced On-Premise DDoS Protection Appliances
As an additional layer of protection, Pearson implemented on-premise DDoS protection leveraging industry-leading DDoS protection appliances. These appliances (implemented in summer 2015) provide a buffer zone of protection in the event that the scrubbing centers require additional time to put in place customized mitigation to respond to specific DDoS attacks. Positioned at the perimeter of the Pearson data center, these protective appliances allow Pearson to maintain service availability while the distributed scrubbing centers prepare to mitigate even the largest DDoS attacks.

Additional Customer Service and Support Improvements

While there was a great deal of work and emphasis on DDoS attacks prevention from the technical perspective, MDE and Pearson are committed to providing a better overall experience testing online for districts and students. The sections below highlight the improvements to district support in advance of the 2015-2016 OLPA test administration.

Customer Service

- Improved and additional training for helpdesk representatives.
- Implementing new customer support software to increase the efficiency of customer support teams by accelerating customer identification, improving incident routing, creating a central knowledge base, and enhancing customer self-service options.

Communications and Processes

- Developed and published a Minnesota Technology Training and Support Plan.
- Developed and implemented on-site support visits process.
- Developed and implemented Office Hours support process.
- Distributed device survey to Minnesota districts; plan to contact districts with risks identified.

Trainings and Support

In preparation for the 2015-2016 OLPA administration, the following technology trainings were available to Minnesota districts.

Apple and Pearson Joint Trainings: Pearson hosted six regional training events in collaboration with Apple to provide training and support to Minnesota districts planning to test on iPads and Macs in 2015-16.

Pearson Technology Trainings: Pearson delivered four live webinar trainings for MN technology coordinators on technology requirements, including Readiness and Infrastructure Trials, Proctor Caching, SystemCheck, TestNav 8, and the Early Warning System.

Regional Trainings: Pearson hosted three regional trainings October 20-23 to outline processes for online technical readiness and provide hands-on experience in preparing tablets and devices for online testing.

Minnesota System/Product Trainings: Pearson delivered live webinar trainings for PearsonAccess, TestNav 8, Item Samplers, and Perspective, and Online Reporting. Pearson posted recorded trainings to PearsonAccess.

Through the first two months (October 19 - December 21, 2015) of the 2015-2016 OLPA administration, the calls to Pearson's Customer Service are significantly less than during the first two and a half months (September 29 - December 21, 2014) of the 2014-2015 OLPA administration. MDE and Pearson plan to prepare and support districts for the Spring 2016 administration in a consistent approach to that of the 2015-2016 OLPA administration.

Appendix A

Actions Taken During Testing Window

April 14

9:52 a.m. System Status Page updated to reflect degraded performance.

10:51 a.m. – Email to District Assessment Coordinators regarding degraded performance, with expected update by 12 p.m. Pearson began rolling restarts of the PearsonAccess servers. Pearson provided increased monitoring of customer service managements system and the number of accessible phone lines to resolve issues.

4:04 p.m. – Email to District Assessment Coordinators regarding degraded performance.

5 p.m. – Pearson serviced the affected server by rebuilding it from the mirrored drive (which had failed but did not suffer hardware problems). Additionally Pearson added two additional servers to the cluster.

Around 9 p.m. Pearson brought the affected applications back online and began testing against Quality Control (QC) and Production (Prod) environments.

All testing and validations were completed and all affected groups and interested parties were notified at approximately 11:45 p.m.

April 15

12 p.m. – Pearson noted three short periods of degraded performance in the morning lasting about five minutes each. TestNav was not affected. Servers were restarted which rectified the degraded performance.

April 21

2:30 p.m. – Volume-based DDoS attack on TestNav.com.

Pearson implemented Akamai Prolexic anti-DDoS services.

May 13-15

7:09 a.m. - Protocol-based DDoS attack on TestNav.com

Pearson implemented new filtering parameters and acted immediately on any abnormal traffic as if it were a DDoS attack.

To improve response times to an attack, Pearson entered into an agreement with Akamai to allow them to block content on Pearson's behalf (Always On Routing).

Appendix B

A Statistical Investigation of Computer Disruptions on Student and School Scores: 2015 FINAL Report

2015 No. 32

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

A Statistical Investigation of Computer Disruptions on Student and School Scores: 2015

FINAL Report

Prepared for: Minnesota Department of Education
1500 Highway 36 West
Roseville, MN 55113

Prepared under: A85179/A94278

Authors: Bethany H. Bynum
Julianne M. Edwards

Date: July 23, 2015

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | Fax: 703.549.9025 | humrro.org

A Statistical Investigation of Computer Disruptions on Students and School Scores: 2015

Table of Contents

List of Tables

List of Figures

Executive Summary

In May 2015, the Human Resources Research Organization (HumRRO) was asked by the Minnesota Department of Education (MDE) to investigate the effects of computer disruptions for students taking online versions of the Minnesota Comprehensive Assessment Series III (MCA-III). HumRRO has been MDE's independent psychometric quality assurance contractor since 2006. During April 2015, there were a number of issues that occurred with the online testing system that directly and indirectly impacted student testing. On April 14, April 15, and April 21, there were unexpected service interruptions that affected teachers' and test monitors' access to PearsonAccess. During the afternoon of April 21, there was a distributed denial of service (DDoS) attack on the TestNav.com portal (the Pearson test delivery platform), which caused students to have limited or no access to TestNav. There is no way to identify students that were impacted by the disruptions that occurred as a result of the service interruptions to PearsonAccess, thus the investigation in this document focuses on the disruption that occurred as a result of the DDoS attack on April 21. To ensure that reported results are valid and reliable, HumRRO investigated the impact of the computer disruptions on students' scores. We worked with MDE and the Technical Advisory Committee (TAC) to conduct a thorough analysis of the assessment data.

Our investigation is based on the well-accepted premise that students' test scores tend to exhibit consistency over time; that is, students who do well in one year of the test tend to do well the subsequent year. However, there could very well be students who scored higher than expected because they had a very good, productive year in school and were simply better prepared. Likewise, there could very well be students who scored lower than expected due to illness or other disruptions in their personal lives that temporarily lowered their ability to perform and show what they actually knew. We cannot know for certain that an *individual* student's performance was specifically impacted by the disruption and not some other event in the student's life that resulted in them performing better or worse than expected. However, we can use the results of the analyses to make an assessment about whether there are *trends* in the data that suggest the disruption had a systematic impact on student performance.

Because most students did not test on April 14, 15 and 21, the days when there were known disruptions to Pearson's testing system, we used a set of variables to match disrupted and non-disrupted students to help estimate the impact of the disruption on student scores. By matching the samples on variables that are likely to predict student scores, any difference between the two samples can be better attributed to the computer disruptions. We used a number of analyses to examine differences in scores between students who were disrupted and those that were not disrupted. Additionally, we investigated the impact of computer disruption on school scores by considering alternative ways to compute school-level means, taking into consideration the disruptions. We do not have the ability to undo the disruptions and we do not have the ability to know what a student would have scored if there were no disruptions. However, we can consider alternative ways of computing school means and evaluating the differences in those means compared to using students' observed scores.

The results of the analyses show no statistical evidence to suggest that the disruption, on average, adversely impacted students, who were testing when the DDoS attack occurred on April 21. Although there is some evidence to suggest differences in scores when comparing students who were disrupted with those who weren't, the effects of the disruption were not in a consistent direction. This indicates that for some grades and subjects the disruption was beneficial and in other grades the disruption was detrimental. And, for most grades there was no impact at all. Additionally, any observed effects were small suggesting that any adjustment

based on the effect size would be inconsequential. Overall, based on the statistical analyses described in this report, students observed score is the best estimate of achievement for this year. This report is intended to inform MDE and the Minnesota TAC of the statistical impact of computer disruption and to be used as a piece of evidence in considering whether policy actions are appropriate.

A Statistical Investigation of the Effects of Computer Disruptions on Student and School Scores: 2015

In May 2015, the Human Resources Research Organization (HumRRO) was asked by the Minnesota Department of Education (MDE) to investigate the effects of computer disruptions for students taking online versions of the Minnesota Comprehensive Assessment Series III (MCA-III). HumRRO has been MDE's independent psychometric quality assurance contractor since 2006. During April 2015, there were a number of issues that occurred with the online testing system that directly and indirectly impacted student testing. On April 14, April 15, and April 21, there were unexpected service interruptions that affected teachers' and test monitors' access to PearsonAccess. During the afternoon of April 21, there was a distributed denial of service (DDoS) attack on the TestNav.com portal (the Pearson test delivery platform, [TestNav]), which caused students to have limited or no access to TestNav. To ensure that reported results are valid and reliable HumRRO investigated the impact of the computer disruptions on students' scores. We worked with MDE and the Technical Advisory Committee (TAC) to conduct a thorough analysis of the assessment data. This document summarizes the results of the statistical investigation for MCA-III reading and math grades 3 – 8, reading grade 10, and math grade 11.

Defining Computer Disruption

Schools and Districts reported multiple types of disruptions that occurred during MCA-III testing in 2015. The disruptions span from occasional problems with the online calculator during testing to complete shutdown of the testing system. Some disruptions directly impacted students' testing experiences (e.g. students were kicked out of TestNav in the middle of their test) while others did not directly impact students testing experiences (e.g. teachers were unable to use PearsonAccess). In order to investigate the potential impact of any given disruption on a test takers performance, we must have the ability to identify those students who had their testing experience disrupted in some manner. Note, there is no way to identify students that were disrupted from testing due to localized issues or problems with online interfacing that were not a result of system wide problems. Thus, we are unable to investigate the potential impact of these disruptions. There were two system wide interruptions that occurred on specific days in April that were a result of systematic problems with Pearson's testing system. One interruption occurred with PearsonAccess, the system teachers and test administrators use to manage student test-taking, and occurred on several days in the testing window. This interruption did not directly impact students' testing experiences, and there is no way to track or account for this impact. The investigation in this document focuses on the disruption that occurred as a result of the DDoS attack on the TestNav system. This disruption caused limited or no access to TestNav from approximately 2:30 PM to 3:00 PM on April 21. HumRRO received testing data from Pearson for April 21 that included submission times for each item. We also received testing data for April 27 and 28, days that would be considered normal testing days. Using this data, we identified a sample of students that had a high probability of testing during the disruption that began at 2:30 PM on April 21.

On a "normal" testing day there tends to be between 80-150 item responses per second. This was the case on April 21 until 2:29:44 PM when that number dropped to 42 and remained low until 2:30:29 PM at which point the item responses jumped to over 400. The item responses remained above 100 until 2:31:42 PM at which point the item responses dropped to 5 and then went to zero (with a very few exceptions) until 3:01:45 PM when item responses resumed to normal levels.

The assumption is that anyone testing when the first drop in responses occurred (at 2:29:44 PM) was disrupted. We do not have exact information on who was testing at this time, but we can create a range of time for which we can reasonably expect a student could have been working on an item when the disruption occurred. To try to determine a reasonable amount of time between item response submittals, we looked at item response submissions on April 27 and April 28. We took the differences between the timestamp for each item response and removed any item responses where there was more than 10 minutes between item responses. We assumed that if there was a 10 minute lag between item responses then a student may have been on a break. We also removed item responses were there was 0 seconds between item responses. These appear to be instances where multiple items were present on a page and submitted simultaneously. We then took the mean of the remaining item response differences to determine the average amount of time students took to respond to/submit an answer to an item. For reading, the average amount of time was 86 seconds with a standard deviation of 110 seconds, and for math, the average amount of time was 163 seconds with a standard deviation of 148 seconds. The response times were consistent across grades. The average time and the standard deviation were considered when identifying students that were likely disrupted.

We used a number of criteria to identify students who were likely testing when the disruption happened. This is not to say that other students were not affected. However, if we were unable to detect an effect in the concentrated sample, then it is unlikely we would find an effect in a larger sample of students where some of those students were potentially not disrupted. Students had to meet one of three criteria to be included in the disruption sample:

- Criterion 1: students who submitted a response between the first drop in responses at 2:29:44 PM and when the testing system resumed to normal at 3:01:45 PM.
- Criterion 2: students who were testing within 2 minutes of the 2:29:44 PM for reading and within 3 minutes of the 2:29:44 PM for math; These values were determined based on the mean response time on a typical day. The rationale behind including these students is that they were likely working on an item when the disruption happened, but may have been unable to submit the item because of the disruption.
- Criterion 3: students who continued testing after the system resumed at 3:01:45 PM. If a student submitted an item response within 2 standard deviations of the average item response time (8 ½ minutes for Math and 5 ½ minutes for reading) *and* submitted item responses after the testing system resumed then they were also included in the disruption sample.

Propensity Matching

To determine the impact of the disruption, we first matched the disrupted students with a group of students who tested on the non-affected administration dates. The disrupted sample included students who tested on April 21 and met one of the criteria described above. The matched non-disrupted sample was chosen from students who did not test on April 14, 15, or 21, as some type of disruption occurred on these days.

Propensity matching is a statistical approach used to match sample groups on a set of variables that are likely to be related to the outcome of interest when random group assignment is not possible. For this study, propensity matching attempted to match the disrupted students with a group of students who have equivalent ability estimates, using variables that are highly related to ability, but unrelated to the disruption. The result was two groups of students who were as

closely matched as we could make them, except for their experience with computer disruptions and 2015 test scores. By matching groups on these variables, the difference in 2015 test scores between the two groups could be more strongly attributed to disruption.

The following variables were used for matching:

- 2014 math and reading theta scores
- Gender (0 = Female, 1 = Male)
- Ethnicity (Dummy coded for Asian, Hispanic, Black, White)
- Eligibility for Free/Reduced-price Lunch status (0 = No, 1 = Yes)
- Limited English Proficiency (LEP) status (0 = No, 1 = Yes)
- School-level percentage of Free/Reduced-price Lunch students
- School-level percentage of Special Education (SPE) students (0 = No, 1 = Yes)
- School-level achievement (Average of 2014 student-level reading and math scores)

For grade 3, where there are no prior year test scores, current year score for the non-disrupted subject (reading or math) was used.

Prior to matching, we examined differences between the two samples on these variables using Cohen's *d*. There were a number of small to moderate differences (Cohen's $d > .10$ and $< .44$) between the disrupted and non-disrupted samples. These results suggest that propensity matching is necessary to ensure the two samples are equivalent. Appendix A tables report the differences prior to matching.

Next, using logistic regression, we regressed group membership (disrupted or not disrupted) onto the matching variables. The pseudo R^2 values of the logistic regression were small, ranging from .0005 to .006. The small values suggest that overall, the combination of prior year student achievement, demographics, free/reduced-price lunch status, school-level achievement and school-level free/reduced-price lunch percentage had little relationship to the likelihood that a student experienced disruptions.

To match the two samples, the predicted disruption probabilities from the logistic regression analyses were saved. The predicted probabilities represent the probability that a student was in the disrupted sample. We used the nearest neighbor method to match the two samples. The predicted disruption probability for each student in the disrupted sample was matched to the student with the closest predicted disruption probability in the non-disrupted sample. The sampling was done without replacement so that each student in the disrupted sample was matched with a unique student in the non-disrupted sample. The average difference between a disrupted student's predicted probability and the matched non-disrupted student's predicted probability was .0000022 for reading and .0000019 for math. The largest difference was .000006. Differences no larger than .20 have been shown to reduce bias and produce accurate group difference estimates (Austin, 2009; Connelly, Sackett, & Waters, 2015). The results suggest that everyone in the disrupted sample was matched with a student in the non-disrupted sample with a nearly identical predicted disruption probability.

To further evaluate the closeness of the matched sample, we examined the mean difference of the matching variables. The average Cohen's *d* between the two samples was .002 for Reading

and .007 for Math and ranged from 0 to .128 for reading and 0 to .109 for math grades 3 through 8. Cohen's *d* effect sizes near zero suggest that the samples were effectively balanced on prior year achievement, school-level achievement, gender, ethnicity, race, free/reduced-price lunch status, LEP status, school-level free/reduced-price lunch percentage, and school-level SPE percentage. For Math grade 11, the propensity matching resulted in Cohen's *d* differences greater than .15 on several matching variables, suggesting, for a number of variables there were small to moderate differences between the groups. While the differences for the majority of the variables were small (0 to .085), differences of .158 and .176 emerged for the 2014 math theta scores and school achievement, suggesting that the matched groups may have some small achievement differences not attributable to the disruption. To account for this difference, we used the 2014 math theta score differences as the baseline difference between the two groups. If 2015 math scores are no more different than the 2014 math scores then we can reasonably assume that differences are not attributable to the disruption but to some other difference between the group that was not accounted for by the propensity matching. A summary of the mean, standard deviations, and effect sizes before and after matching are found in Appendix A.

2015 Student-Level Analyses

Using the matched samples, we examined whether students' scores from tests with computer disruptions differed from students' scores from tests that were not disrupted. By matching the samples on variables that are likely to predict student scores, any difference between the two samples can be better attributed to the computer disruptions. We used a number of analyses to examine differences in scores. Scores from 2015 were maximum likelihood theta estimates for reading and math.

Differences in Average 2015 Test Scores

If testing disruptions had no overall impact, then the averages of the 2015 test scores for the two matched groups would be expected to be nearly identical. On the other hand, differences in the average 2015 test scores would be evidence that computer disruption did impact test performance. We examined mean differences between the two samples using a *t*-test and Cohen's *d* effect size. Table 1 presents the results.

Table 1. Mean Differences between Disrupted and Non-Disrupted Groups by Grade

Grade	*n*	Non-Disrupted Mean (*SD*)	Disrupted Mean (*SD*)	*t*-value	*d*	*F*-value
Reading						
3	589	0.04 (1.16)	1.16 (0.10)	-1.04	-0.06	**1.21**
4	844	0.00 (1.17)	1.17 (0.10)	-1.85	-0.09	**1.17**
5	698	0.07 (1.13)	1.13 (0.07)	-0.01	0.00	1.05
6	577	0.40 (1.11)	1.11 (0.53)	**-2.08**	-0.12	1.10
7	397	0.07 (1.09)	1.09 (0.07)	-0.03	0.00	1.09
8	359	0.13 (1.16)	1.16 (0.12)	0.08	0.01	1.15
10	371	-0.02 (1.10)	1.10 (-0.09)	0.82	0.06	1.20
Math						
3	306	0.02 (1.14)	1.14 (-0.03)	0.51	0.04	1.23
4	367	0.24 (1.14)	1.14 (0.25)	-0.03	0.00	1.07
5	488	0.43 (1.07)	1.07 (0.40)	0.36	0.02	1.16
6	346	0.41 (1.05)	1.05 (0.37)	0.64	0.05	**1.42**
7	410	0.25 (1.08)	1.08 (0.17)	1.07	0.07	1.21
8	336	0.36 (1.01)	1.01 (0.32)	0.50	0.04	1.20
11	420	0.32 (1.02)	1.02 (0.21)	1.71	0.12	**1.44**

Note. *F*-value = Test of the equality of variance. *t*-value = Test of mean difference. *d* = Cohen's *d*. Bolded values indicate statistical significance at $p < .05$ (two-tailed).

For reading, the effect sizes ranged from -.12 to .06, suggesting no overall directional effect. The mean difference was statistically significant for grade 6, where the disrupted sample had higher theta scores than the non-disrupted sample. However, the overall effect size was small (*d* = -.12). For math, the effect sizes ranged from .00 to .12. With the exception of grade 4, the disrupted group had higher theta scores than the non-disrupted group, but generally the differences were small and none of the differences were significantly different. The largest differences was for grade 11, where the effect size was .12. The effect size was smaller than the 2014 math scores, suggesting that the difference may be due to differences not accounted for by the propensity matching. Of particular note, the standard deviations for the non-disrupted group are systematically larger than the standard deviations for the disrupted group, suggesting the variance in scores is smaller for the disrupted group. We examined the equality of variances between the two groups and despite the systematic difference, there were only significant differences for grade 3 reading, grade 4 reading, grade 6 math, and grade 11 math.

Examining the Predictability of 2015 Test Scores

First, to examine predictability of 2015 test scores, we examined the incremental variance that disruption added to the prediction of the 2015 test scores beyond other known indicators of performance (including prior year achievement, free/reduced-price lunch status, LEP status, race, ethnicity, gender, school-level achievement, school-level SPE, school-level percentage of free/reduced lunch and school size). If the inclusion of disruption in the multiple regression models adds to the estimation of 2015 scores, then this would supply evidence that disruption impacted 2015 test scores. For this model, the two groups were included in the regression models together.

Table 2 reports the R^2 values for each model, the R^2 change between the two models, and the standardized and unstandardized regression coefficient for disruption. With the inclusion of

disruption in the model, the R^2 value changed very little or not at all, indicating that disruption added very little to the prediction of 2015 scores. After controlling for all other variables in the model, disruption was a significant predictor of 2015 scores for grade 4 reading and grade 5 math. The effect size was positive suggesting that disruption had a positive impact on theta scores.

Table 2. Incremental Validity Estimation of Disruption

Grade	*n*	Covariates Only R^2	Covariates + Disruption R^2	ΔR^2	β estimate for disruption (0,1)	*b* estimates for disruption
Reading						
3	1178	0.612	0.612	0.000	0.008	0.018
4	1688	0.693	0.694	0.001	**0.033**	**0.073**
5	1396	0.721	0.721	0.000	0.006	0.014
6	1154	0.658	0.660	0.003	**0.050**	**0.109**
7	794	0.705	0.705	0.000	-0.001	-0.002
8	718	0.728	0.728	0.000	-0.016	-0.035
10	742	0.644	0.645	0.001	-0.030	-0.063
Math						
3	612	0.633	0.633	0.000	0.012	0.026
4	734	0.781	0.781	0.000	0.019	0.042
5	976	0.765	0.765	0.000	0.019	0.038
6	692	0.763	0.763	0.000	0.008	0.015
7	820	0.802	0.802	0.001	-0.025	-0.052
8	672	0.771	0.771	0.000	-0.014	-0.027
11	840	0.683	0.683	0.000	0.005	0.010

Note. ΔR^2 is the difference in the R^2 values between the Covariates Only model and the Covariates + Disruption model. β is standardized regression coefficient and *b* is the raw unstandardized regression coefficient. Disruption is a dichotomous variable where 1 = disrupted and 0 = not disrupted.

Next, using all available data to create equations that predicted 2015 test scores, we can statistically determine whether disrupted students scored differently than expected. Prediction equations were estimated for the disrupted students and separately for the matched sample of non-disrupted students. In addition to prediction equations, this technique gave us multiple regression coefficients (R^2). R^2 values are interpreted like correlation coefficients to tell us how well 2015 test scores can actually be predicted. If students' performance was affected, the strength of the prediction for the disrupted students should be less than for the non-disrupted students as shown by lower multiple regression coefficients. A lower coefficient means that students' performance in the disrupted group was not as predictable as students' performance in the non-disrupted group. This would supply another piece of evidence about the impact of the computer disruptions.

Table 3 presents the multiple R^2 values for the disrupted and non-disrupted groups. Overall, 2015 test scores were well predicted for both samples, with 60% to 75% of the variance accounted for by the predictor variables for reading and 58% to 81% for math. Note that prediction is lower for grade 3 in both subjects because students in grade 3 did not have prior year test scores, the strongest predictor of 2015 test scores. Generally, there were slightly higher R^2 values for the non-disrupted group; although, the difference in variance accounted for was practically small, ranging from .3% to 9%. Because the variance in 2015 scores was systematically larger for the non-disrupted group, the differences in R^2 values could be due to

differences in variance and not wholly attributable to differences in prediction. We computed Root Mean Square Error (RMSE) estimates to account for the differences in variance. The RMSE differences were small, ranging from .001 to .084. The RMSE differences did not show a systematic direction difference. For several grades, the RMSE was higher for the disrupted group.

Table 3. Predictability of 2015 Scores for Non-Disrupted and Disrupted Groups

Grade	*n*	Non-Disrupted R^2	Disrupted R^2	R^2 Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Differences
Reading							
3	589	0.629	0.605	0.024	0.708	0.662	0.045
4	844	0.712	0.679	0.033	0.627	0.613	0.014
5	698	0.723	0.722	0.001	0.594	0.610	-0.015
6	577	0.691	0.632	0.059	0.615	0.639	-0.024
7	397	0.729	0.699	0.031	0.566	0.572	-0.006
8	359	0.749	0.721	0.028	0.582	0.573	0.009
10	371	0.678	0.619	0.059	0.625	0.620	0.006
Math							
3	306	0.679	0.584	0.096	0.644	0.660	-0.017
4	367	0.758	0.814	-0.056	0.559	0.475	0.084
5	488	0.785	0.751	0.034	0.495	0.495	0.001
6	346	0.777	0.749	0.028	0.498	0.443	0.055
7	410	0.806	0.809	-0.003	0.477	0.430	0.047
8	336	0.780	0.776	0.004	0.473	0.435	0.038
11	420	0.697	0.682	0.015	0.560	0.478	0.082

Note. R^2 Difference is the difference between the Non-disrupted R^2 and the Disrupted R^2.

Examine Distributions of Predicted Student Scores

The prediction equations for the non-disrupted students give us a statistical statement about what to normally expect for students testing under non-disrupted conditions. The prediction is not perfect, but given the high R^2 values we can use the prediction equation to calculate how disrupted students might have scored had they not been disrupted. For each disrupted student, we computed their 2015 predicted score using the regression equation computed for the non-disrupted students. Next, we computed the difference between the predicted score and observed score, where positive values indicate higher predicted scores than observed and negative values indicate higher observed scores than predicted. Tables 4 through 7 present the distribution of observed and predicted scores for the non-disrupted and disrupted sample for reading and math, respectively. The difference between observed and predicted scores is also reported.[1]

[1] Technical Note: When a prediction equation is derived on one sample and applied to a second sample, the variance of the residuals is expected to be larger due to shrinkage. Our predictions of performance for the disruption group were slightly weaker than would be expected based on the shrinkage associated with applying the prediction equation to a randomly equivalent sample. Given that our second sample was not randomly equivalent, but differs by the computer disruption, the small difference suggests that our prediction utility is not severely reduced in the disrupted sample. However, we did investigate differences in predicted scores for the two samples further.

Table 4. Distribution of the Difference between Predicted and Observed Reading Scores for the Non-Disrupted Sample

	Mean	*SD*	5th	10th	90th	95th
Grade 3 (*n* = 589)						
Difference	0.00	0.71	-1.13	-0.91	0.88	1.19
Predicted Theta	0.04	0.92	-1.69	-1.29	1.12	1.43
Observed Theta	0.04	1.16	-2.01	-1.39	1.47	1.86
Grade 4 (*n* = 844)						
Difference	0.00	0.63	-1.04	-0.75	0.75	1.00
Predicted Theta	0.00	0.99	-1.77	-1.33	1.21	1.55
Observed Theta	0.00	1.17	-1.94	-1.55	1.46	1.87
Grade 5 (*n* = 698)						
Difference	0.00	0.59	-0.91	-0.72	0.73	1.03
Predicted Theta	0.07	0.96	-1.73	-1.22	1.24	1.51
Observed Theta	0.07	1.13	-1.98	-1.41	1.44	1.81
Grade 6 (*n* = 577)						
Difference	0.00	0.62	-1.00	-0.79	0.79	1.03
Predicted Theta	0.40	0.92	-1.32	-0.77	1.56	1.83
Observed Theta	0.40	1.11	-1.50	-0.97	1.68	2.21
Grade 7 (*n* = 397)						
Difference	0.00	0.57	-0.93	-0.73	0.69	0.91
Predicted Theta	0.07	0.93	-1.62	-1.16	1.27	1.52
Observed Theta	0.07	1.09	-1.83	-1.25	1.35	1.87
Grade 8 (*n* = 359)						
Difference	0.00	0.58	-0.96	-0.76	0.70	0.98
Predicted Theta	0.13	1.01	-1.73	-1.25	1.35	1.62
Observed Theta	0.13	1.16	-1.90	-1.40	1.47	1.87
Grade 10 (*n* = 371)						
Difference	0.00	0.63	-1.00	-0.74	0.76	0.94
Predicted Theta	-0.02	0.91	-1.49	-1.28	1.08	1.32
Observed Theta	-0.02	1.10	-2.00	-1.57	1.31	1.73

Note. For the non-disrupted group, if all variables are normally distributed, the mean difference between the predicted and observed theta should be zero.

Table 5. Distribution of the Difference between Predicted and Observed Reading Scores for the Disrupted Sample

	Mean	*SD*	5th	10th	90th	95th
Grade 3 (*n* = 589)						
Difference	-0.03	0.69	-1.07	-0.83	0.86	1.21
Predicted Theta	0.08	0.85	-1.42	-1.04	1.15	1.36
Observed Theta	0.10	1.05	-1.70	-1.27	1.38	1.73
Grade 4 (*n* = 844)						
Difference	-0.08	0.63	-1.13	-0.85	0.69	0.95
Predicted Theta	0.02	0.90	-1.52	-1.18	1.10	1.38
Observed Theta	0.10	1.08	-1.71	-1.30	1.49	1.82
Grade 5 (*n* = 698)						
Difference	-0.01	0.62	-1.07	-0.79	0.78	1.03
Predicted Theta	0.05	0.95	-1.63	-1.18	1.18	1.54
Observed Theta	0.07	1.16	-1.96	-1.40	1.43	1.89
Grade 6 (*n* = 577)						
Difference	-0.11	0.65	-1.23	-0.92	0.61	0.89
Predicted Theta	0.42	0.86	-1.02	-0.80	1.43	1.78
Observed Theta	0.53	1.05	-1.17	-0.80	1.85	2.52
Grade 7 (*n* = 397)						
Difference	0.02	0.61	-0.95	-0.71	0.76	1.01
Predicted Theta	0.09	0.89	-1.49	-1.09	1.12	1.65
Observed Theta	0.07	1.04	-1.83	-1.24	1.34	1.62
Grade 8 (*n* = 359)						
Difference	0.04	0.61	-0.91	-0.75	0.82	1.07
Predicted Theta	0.16	0.90	-1.39	-1.09	1.34	1.52
Observed Theta	0.12	1.08	-1.60	-1.19	1.52	2.06
Grade 10 (*n* = 371)						
Difference	0.07	0.64	-0.95	-0.75	0.85	1.10
Predicted Theta	-0.01	0.83	-1.35	-1.02	1.01	1.26
Observed Theta	-0.09	1.00	-1.73	-1.32	1.16	1.47

Table 6. Distribution of the Difference between Predicted and Observed Math Scores for the Non-Disrupted Sample

	Mean	*SD*	5th	10th	90th	95th
Grade 3 (*n* = 306)						
Difference	0.00	0.64	-1.06	-0.83	0.86	1.00
Predicted Theta	0.02	0.94	-1.69	-1.28	1.12	1.53
Observed Theta	0.02	1.14	-1.87	-1.51	1.40	1.74
Grade 4 (*n* = 367)						
Difference	0.00	0.56	-0.91	-0.73	0.66	0.93
Predicted Theta	0.24	0.99	-1.49	-1.03	1.47	1.76
Observed Theta	0.24	1.14	-1.76	-1.23	1.65	2.06
Grade 5 (*n* = 488)						
Difference	0.00	0.49	-0.80	-0.61	0.60	0.81
Predicted Theta	0.43	0.95	-1.29	-0.79	1.57	1.84
Observed Theta	0.43	1.07	-1.43	-0.89	1.62	2.03
Grade 6 (*n* = 346)						
Difference	0.00	0.50	-0.81	-0.59	0.58	0.78
Predicted Theta	0.41	0.93	-1.26	-0.76	1.53	1.82
Observed Theta	0.41	1.05	-1.55	-0.92	1.71	2.01
Grade 7 (*n* =410)						
Difference	0.00	0.48	-0.74	-0.55	0.61	0.74
Predicted Theta	0.25	0.97	-1.56	-0.97	1.49	1.69
Observed Theta	0.25	1.08	-1.53	-1.22	1.64	1.98
Grade 8 (*n* =336)						
Difference	0.00	0.47	-0.75	-0.63	0.59	0.75
Predicted Theta	0.36	0.89	-1.26	-0.78	1.46	1.72
Observed Theta	0.36	1.01	-1.57	-0.96	1.47	1.88
Grade 11 (*n* =420)						
Difference	0.00	0.56	-0.84	-0.68	0.68	0.91
Predicted Theta	0.32	0.85	-1.08	-0.66	1.44	1.68
Observed Theta	0.32	1.02	-1.42	-1.02	1.53	1.90

Note. For the non-disrupted group, if all variables are normally distributed, the mean difference between the predicted and observed theta should be zero.

Table 7. Distribution of the Difference between Predicted and Observed Math Scores for the Disrupted Sample

	Mean	*SD*	5th	10th	90th	95th
Grade 3 (*n* = 306)						
Difference	-0.02	0.67	-1.13	-0.84	0.84	1.13
Predicted Theta	-0.05	0.84	-1.65	-1.25	1.03	1.28
Observed Theta	-0.03	1.02	-1.67	-1.41	1.22	1.62
Grade 4 (*n* = 367)						
Difference	-0.05	**0.50**	-0.87	-0.66	0.60	0.74
Predicted Theta	0.19	0.98	-1.76	-1.22	1.34	1.61
Observed Theta	0.25	1.10	-1.80	-1.15	1.60	1.87
Grade 5 (*n* = 488)						
Difference	-0.04	0.51	-0.89	-0.68	0.59	0.79
Predicted Theta	0.36	0.88	-1.09	-0.79	1.46	1.76
Observed Theta	0.40	0.99	-1.19	-0.86	1.64	1.89
Grade 6 (*n* = 346)						
Difference	-0.02	**0.45**	-0.70	-0.56	0.55	0.73
Predicted Theta	0.35	0.80	-1.19	-0.70	1.33	1.44
Observed Theta	0.37	0.88	-1.14	-0.78	1.44	1.69
Grade 7 (*n* =410)						
Difference	0.04	0.45	-0.69	-0.53	0.62	0.80
Predicted Theta	0.21	0.88	-1.33	-0.96	1.23	1.47
Observed Theta	0.17	0.98	-1.62	-1.14	1.40	1.57
Grade 8 (*n* =336)						
Difference	0.01	0.46	-0.74	-0.52	0.57	0.81
Predicted Theta	0.33	0.90	-1.18	-0.83	1.54	1.76
Observed Theta	0.32	0.92	-1.24	-0.79	1.49	1.88
Grade 11 (*n* =420)						
Difference	-0.03	**0.51**	-0.95	-0.71	0.60	0.79
Predicted Theta	0.18	0.76	-1.01	-0.79	1.18	1.45
Observed Theta	0.21	0.85	-1.13	-0.87	1.29	1.61

Note. Bolded values indicate significant difference in standard deviation at $p < .05$.

Compare Predictions for Disrupted Students to Non-Disrupted Students

Large numbers of students with *notable differences* between observed and predicted scores provides another piece of evidence about the impact of the computer disruptions. We defined large number and notable differences by comparing the difference in observed and predicted scores between the non-disrupted and disrupted groups. The non-disrupted group represented the baseline: what would be expected under normal testing conditions.

First, we examined the difference in the standard deviation for the disrupted and non-disrupted group using the F-test to examine the equality of the two variances. This provides information on whether the spread of differences in predicted and observed scores is statistically different for the two groups. The variance of the differences were significantly different for math grades 4, 6, and 11. For grade 4, the disrupted group had a larger variance of differences. For grade 6 and 11, the non-disrupted group had a larger variance of differences. Results suggest

that the generally the variance of the differences were similar for the two groups and for grades where there were differences, the differences were not systematic.

Next, we compared the distribution of differences using P-P plots. The P-P plots provide an evaluation of whether the differences between observed and predicted scores are normally distributed. We would expect the differences between observed and predicted scores to be normally distributed for the non-disrupted sample. That is, generally speaking, most of the differences should be near zero and there should be equal number of differences where the observed score is greater than the predicted score and the predicted score is greater than the observed score. If the disruption impacted scores, then the difference between predicted and observed would be larger for the disrupted group and deviate from both the normal distribution and the disrupted group distribution. We compared P-P plots for the non-disrupted and disrupted groups. Appendix B provides the P-P plots. Overall, the differences between predicted and observed scores did not vary from the normal distribution for either the non-disrupted or the disrupted group. As such, there were no systematic differences between the two groups.

Finally, we computed the difference in observed and predicted scores at the 5th, 10th, 90th and 95th percentile for the non-disrupted group and determined the number of students in the disrupted group who were at or below the same cut point for the 5th and 10th percentile and those that were at or above the cut point for the 90th and 95th percentile. If more than 5% and 10% of the disrupted students were below the 5th and 10th percentile cuts, respectively, then more students in the disrupted group scored higher than expected. If more than 10% and 5% of the disrupted students were above the 90th and 95th percentile cut, respectively, then more students in the disrupted group scored lower than expected. Either case would provide evidence that the computer disruption had an impact on scores. Table 8 presents the percent of students in the disrupted group below the 5th and 10th percentile cuts and above the 90th and 95th percentile cuts.

Table 8. Percent of Disrupted Students with Predicted and Observed Score Differences at the 5th, 10th, 90th and 95th Percentile of Non-Disrupted Students

Grade	*n*	5th	10th	90th	95th
Reading					
3	589	4.41%	8.15%	9.68%	5.26%
4	844	6.52%	13.63%	8.65%	4.62%
5	698	8.02%	11.89%	11.46%	4.73%
6	577	7.80%	12.82%	7.11%	3.99%
7	397	5.29%	9.57%	12.34%	6.80%
8	359	4.18%	9.75%	11.98%	6.96%
10	371	4.31%	10.24%	14.02%	8.09%
Math					
3	306	5.88%	10.13%	9.80%	6.86%
4	367	4.36%	7.63%	8.45%	1.91%
5	488	6.97%	11.68%	9.84%	4.71%
6	346	2.89%	8.09%	9.54%	4.34%
7	410	3.90%	9.27%	10.00%	5.61%
8	336	3.87%	6.55%	9.52%	6.55%
11	420	6.43%	10.71%	7.38%	3.10%

Note. Percentages larger than 5% or 10% at the 5th and 5^{10h} percentile, respectively, indicates an advantage. Percentages larger than 5% and 10% at the 95th and 90th percentile, respectively indicates a disadvantage.

For reading, the results show a slightly higher percent of disrupted students above the 90th and/or 95th cut point for grades 5, 7, 8 and 10, than would be expected. These results suggest that for grade 7, for example, approximately 6.80% of disrupted students had higher predicted scores than observed score; 1.8% higher than what would have been expected based on the non-disrupted sample. In contrast, there were a higher percentage of disrupted students below the 5th and 10th percentile cut point for reading grades 4, 5, and 6. The largest differences were in grade 4, where 6.52% of the disrupted sample was below the 5th percentile cut and nearly 14% of the disrupted sample was below the 10th percentile cut. This suggests that nearly 4% more disrupted students had higher observed scores than predicted scores; providing evidence that the disruption did not always have a negative impact. It is important to note, that given our sample sizes, these discrepancies amount to unexpected differences between predicted and observed scores for 1 to 14 students depending on the grade.

For Math grades 5 and 11, there were higher percentages of disrupted students below the 5th and 10th percentile cuts, indicating higher observed than predicted scores. For grades 3, 7 and 8 there were higher percentages of disrupted students above the 95th percentile cut. The differences above what would have been expected ranged from .61% to 1.86%. The percent of disrupted students in the 90th percentile were all at or below 10%. Overall, for math, there were only small differences between what would have been expected and what was observed.

Theta Score Differences

The disruption occurred at a definable point in time. As such, we were able to identify the last item a student submitted prior to the disruption. Since there is an identifiable point on the test in which students were disrupted, the test can be divided into two based on whether an item was completed prior to the disruption or after the disruption. If the disruption had an impact on students, items taken after the disruption may have been impacted, leading to lower (or higher) scores than what would be expected if no disruption occurred. To evaluate whether there were differences in responding prior to and after the disruption, we computed two scores, one using items prior to the disruption and one using items after the disruption. The differences in the scores were compared.

All operational items that were given to a student prior to the disruption were included in the calculation of the *before disruption* theta score. All operational items given to a student after the disruption, even if it was on a subsequent day, were included in the *after disruption* theta score. The item that was identified as the "point of disruption" was included in the *after disruption* theta score. The point of disruption was identified by the response submittal times. Before disruption and after disruption theta scores were estimated using Maximum Likelihood Estimation (MLE), using the same scoring algorithm that is used to produce operational scores. Through propensity matching we matched, one-to-one, students in the disrupted sample to students in the non-disrupted sample. The point on the test at which the disrupted student was interrupted was used to define the before and after disruption point for the matched student in the non-disrupted sample. That is, if a student in the disrupted sample was disrupted at the 18th item, the matched non-disrupted student test was split so that the before disrupted score included all items prior to the 18th item and the after disruption score included the 18th item and all items after the 18th item.

For students that were disrupted while taking a field test item, all operational items prior to the field test item were included in the before disruption score and all operational items after the field test item were included in the after disruption score. Finally, to ensure stable theta

estimates were computed, we excluded students who completed less than 10 items prior to the disruption or students that completed less than10 items after the disruption.

Table 9 includes the mean theta scores before the disruption and after the disruption for the non-disrupted and disrupted groups. For both groups, the after disruption score was lower than the before disruption score. This is not surprising given that students tend to do less well on items presented at the end of the test because of fatigue.

To determine if the disrupted group was adversely effected, we compared the differences in the before and after scores for the two groups (see Table 10). We examined mean differences between the two samples using a *t*-test and Cohen's *d* effect size. Overall, the differences were small, with effect sizes ranging from -0.31 to 0.12 for reading and -.18 to .18 for Math. The direction of the differences was not always consistent. For several grades, the disrupted group had smaller differences between before and after scores and for other grades, the non-disrupted group had smaller differences. The difference between the disrupted and non-disrupted groups was statistically significant for reading grade 5 and math grade 5. For reading grade 5 the differences between before and after disruption scores were larger for the disrupted group. However, for math the differences were larger for the non-disrupted group.

Table 9. Mean Theta Scores Before the Disruption and After the Disruption for Non-Disrupted and Disrupted Groups

Grade	*n*	Non-Disrupted Before Disruption Mean (*SD*)	Non-Disrupted After Disruption Mean (*SD*)	Disrupted Before Disruption Mean (*SD*)	Disrupted After Disruption Mean (*SD*)
Reading					
3	444	0.11 (1.25)	-0.06 (1.29)	0.26 (1.14)	0.12 (1.17)
4	585	0.14 (1.20)	0.06 (1.36)	0.23 (1.10)	0.15 (1.29)
5	525	0.12 (1.22)	0.02 (1.28)	0.17 (1.20)	-0.02 (1.29)
6	477	0.57 (1.20)	0.27 (1.24)	0.70 (1.13)	0.48 (1.13)
7	259	0.15 (1.12)	0.00 (1.25)	0.24 (1.09)	0.19 (1.33)
8	254	0.21 (1.23)	0.08 (1.34)	0.30 (1.15)	0.18 (1.11)
10	331	0.14 (1.23)	-0.15 (1.21)	0.10 (1.05)	-0.20 (1.20)
Math					
3	167	0.11 (1.03)	0.11 (1.23)	0.05 (1.02)	-0.01 (1.00)
4	198	0.32 (1.18)	0.29 (1.25)	0.18 (1.15)	0.12 (1.19)
5	246	0.46 (1.15)	0.34 (1.11)	0.45 (0.96)	0.43 (1.04)
6	193	0.45 (1.09)	0.43 (1.21)	0.46 (0.95)	0.40 (0.90)
7	147	0.35 (1.14)	0.26 (1.17)	0.11 (0.94)	0.03 (1.01)
8	119	0.41 (1.05)	0.31 (1.18)	0.31 (0.97)	0.33 (1.06)
11	286	0.36 (1.06)	0.23 (1.14)	0.27 (0.88)	0.18 (0.99)

Note. Students with less than 10 items before the disruption or less than 10 items after the disruption were not include in the analyses.

Table 10. Mean Differences among Theta Scores Before the Disruption and After the Disruption for Non-Disrupted and Disrupted Groups

Grade	*n*	Non-Disrupted Mean (*SD*)	Disrupted Mean (*SD*)	*t*-value	*d*
Reading					
3	444	0.18 (0.88)	0.14 (0.87)	0.67	-0.04
4	585	0.08 (0.90)	0.09 (0.86)	-0.09	0.01
5	525	0.10 (0.80)	0.20 (0.86)	**-1.96**	0.12
6	477	0.29 (0.81)	0.22 (0.78)	1.50	-0.10
7	259	0.16 (0.86)	0.05 (0.83)	1.48	-0.13
8	254	0.14 (0.90)	0.12 (0.80)	0.23	-0.02
10	331	0.28 (0.84)	0.30 (0.80)	-0.31	0.02
Math					
3	167	0.00 (0.61)	0.06 (0.60)	-0.90	0.10
4	198	0.03 (0.51)	0.06 (0.56)	-0.58	0.06
5	246	0.12 (0.56)	0.01 (0.58)	**2.00**	-0.18
6	193	0.01 (0.55)	0.06 (0.52)	-0.89	0.09
7	147	0.09 (0.57)	0.07 (0.50)	0.20	-0.02
8	119	0.10 (0.66)	-0.01 (0.60)	1.40	-0.18
11	286	0.13 (0.71)	0.09 (0.64)	0.74	-0.06

Note. d = Cohen's *d*. Positive Cohen's *d* values indicates a larger difference between scores before the disruption and after the disruption for the disrupted group. Bolded values indicate significant differences at $p < .05$ (two-tailed).

Student-Level Summary

A number of analyses were conducted to examine the potential impact of computer disruptions on student scores. The statistical evidence provided in this report is intended to inform MDE and the TAC about whether the computer disruption systematically impacted student test scores. The evidence shows that there were potentially some disruption effects; however, the effects on students' scores were neither widespread nor large. In addition, the evidence shows that disruptions were not consistently detrimental, but at times were beneficial. Among all the statistical analyses, reading grade 6 was the only grade to show consistent difference between the non-disrupted and disrupted sample and the results suggested that the disrupted sample had slightly higher scores than the non-disrupted sample.

2015 School-Level Analyses

School-level accountability is based on the aggregation of student-level scores. While computer disruptions did not have a large impact on student-level scores, there is concern that once aggregated the disruption could impact school-level accountability results. We investigated the impact of computer disruption on school scores using the following set of analyses.

Distribution of School Disruptions

We defined schools as potentially disrupted or not disrupted. Not disrupted schools were those that tested online but did not test on April 14, 15, or 21. Because there were multiple days on which disruption occurred and the type of disruption differed, we defined several samples of potentially disrupted schools. The first sample included schools that tested online on April 14,

April 15, or April 21. The second sample included schools that tested on April 21. The third sample included schools with students that were included in the student-level disrupted analyses, testing within a window of time on April 21 when the disruption occurred. Table 11 provides the number of schools in each sample. Note that the schools listed in the potentially disrupted sample were combined to make up the sample that tested online on April 14, 15, and 21.

Table 11. Sample Distribution of Schools

Grade	Total # Schools	Schools Tested on April 14 & 15 but not on April 21[a]	Schools Tested on April 21 and No Students Included in Disrupted Sample[a]	Schools Tested on April 21 with Students in Disrupted Sample[b]	Schools that did not Test on April 14, 15 or 21
Reading					
3	956	217	295	88	356
4	959	169	291	113	386
5	948	185	285	96	382
6	725	133	184	53	355
7	678	144	172	43	319
8	721	147	175	36	363
10	718	132	136	34	416
Math					
3	955	78	231	52	594
4	954	80	192	54	628
5	945	71	187	62	625
6	723	65	126	42	490
7	678	78	130	39	431
8	719	77	129	31	482
11	726	152	138	35	401

[a] Potentially disrupted sample, may or may not have had disrupted students, but none were identified based on the available data. [b] Known disrupted sample, a proportion of students were included in the disrupted sample for the student-level analyses.

Table 12 provides a summary of the average percentage of students who tested on April 14, 15, or 21. On average, for schools that tested on these days, 43% to 55% of students within a grade tested. More notable is the range of percentages. Some schools tested less than one percent on those days while other schools tested every student within a grade on those days.

Of the students who tested on those days, a proportion was actually disrupted. Since all students included in the disrupted sample are students with the highest likelihood of being disrupted, the percentage of disrupted students within a school can be considered a lower bound estimate of the actual percent of students within a school that were disrupted. On average, for a school that tested on April 21, 7% to 14% of students were included in the disrupted sample, ranging from 0% to 83%.

Table 12. Average Percent of Students who Tested on April 14, 15 and 21 for All Schools that Tested on April 14, 15, and 21

Grade	*n*	Mean	*SD*	Min	Max
Reading					
3	569	51.58%	34.84%	0.45%	100%
4	548	51.70%	34.25%	0.63%	100%
5	533	48.40%	35.81%	0.34%	100%
6	330	47.25%	38.63%	0.29%	100%
7	309	48.09%	40.27%	0.20%	100%
8	297	51.55%	39.37%	0.26%	100%
10	237	51.46%	39.05%	0.19%	100%
Math					
3	340	43.87%	32.26%	0.45%	100%
4	306	45.76%	33.24%	0.65%	100%
5	299	43.87%	32.52%	0.28%	100%
6	207	47.38%	37.72%	0.17%	100%
7	208	48.88%	39.38%	0.27%	100%
8	191	50.38%	39.89%	0.26%	100%
11	275	55.06%	36.74%	0.37%	100%

Note. Schools with less than 10 students testing in a grade were excluded.

School-Level Score Differences

To evaluate the impact of the disruption on school-level means it is informative to consider possible ways that the school-level means could be computed, taking into consideration the disruptions. We do not have the ability to undo the disruptions and we do not have the ability to know what a student would have scored if there were no disruptions. However, we can consider alternative ways of computing school means and evaluating the differences in those means compared to using students' observed scores. Thus, we computed school-level means using four alternatives.

First, the students who were identified in the disruption sample were excluded from the school-level means. For some schools, this effectively excluded less than one percent of students and for others it excluded more than 80% of students. Again, we did not exclude all students who may have been disrupted, but we excluded a concentrated set of students who were. Second, all students who tested on April 21 were excluded from the school-level mean calculation. Considering there were two types of disruptions on April 21, this approach takes into consideration that we do not have complete data to determine whether students were directly or indirectly affected on April 21, so we removed all students potentially affected. Next, we computed scores removing students who tested on April 14, 15, or 21, as these are the days that were identified as having disruptions. For all three of these days, we are unable to identify students that were potentially disrupted due to the interruption to PearsonAccess; thus, we computed scores excluding any student who tested on these days.

Finally, we computed school-level means using the predicted score from the student-level analyses described above for students who were disrupted. It is important to understand that differences in the observed and predicted scores do not *only* capture the disruption effect. There are other uncertainties in test scores (e.g., some students may not do as well as they should

because they didn't eat breakfast or they feel sick) that can yield differences between observed and predicted scores. If school-level means differ when using the predicted scores, we cannot solely attribute the difference to the disruption. However, if the scores are not different, then evidence is provided that the disruptions did not adversely impact school-level means.

Tables13, 14, and 15 present the results comparing the school-level means using observed scores and school-level means removing students who tested on April 14, 15, and 21 (Table 13), removing students who tested on April 21 (Table 14), removing students who were identified in the disruption sample (Table 15), and using the predicted score in place of the observed score for students in the disrupted sample (Table 15).

Table 13. Mean Scores for Schools that Tested on April 14, 15 or 21

Grade	Group A *n*	Group A Mean θ_1	Group A *SD*	Group B *n*	Group B Mean θ_1	Group B *SD*	Group C *n*	Group C Mean θ_1	Group C *SD*	Avg. % of Students Removed	$\theta_2 - \theta_3$	*d*[a]
Reading												
3	569	0.02	0.51	519	0.01	0.50	519	-0.06	0.7	51.6%	-0.08	-0.12
4	548	0.02	0.51	498	0.02	0.51	498	-0.11	0.72	51.7%	-0.13	-0.21
5	533	-0.01	0.52	494	-0.01	0.52	494	-0.11	0.72	48.4%	-0.10	-0.17
6	330	0.04	0.53	297	0.02	0.54	297	-0.11	0.73	47.2%	-0.13	-0.21
7	309	-0.16	0.53	274	-0.15	0.53	274	-0.25	0.75	48.1%	-0.10	-0.16
8	297	-0.13	0.51	273	-0.13	0.52	273	-0.28	0.72	51.6%	-0.16	-0.25
10	237	-0.39	0.55	221	-0.40	0.56	221	-0.58	0.67	51.5%	-0.17	-0.28
Math												
3	340	-0.09	0.53	318	-0.10	0.53	318	-0.13	0.64	43.9%	-0.03	-0.05
4	306	-0.03	0.56	283	-0.02	0.57	283	-0.08	0.76	45.8%	-0.06	-0.09
5	299	-0.02	0.58	288	-0.01	0.57	288	-0.11	0.73	43.9%	-0.10	-0.15
6	207	-0.05	0.62	191	-0.06	0.63	191	-0.18	0.81	47.4%	-0.12	-0.16
7	208	-0.20	0.58	192	-0.19	0.57	192	-0.36	0.81	49.1%	-0.17	-0.24
8	191	-0.21	0.61	177	-0.21	0.62	177	-0.36	0.76	50.4%	-0.15	-0.21
11	275	-0.31	0.66	255	-0.33	0.66	255	-0.45	0.76	55.1%	-0.11	-0.16

Note. Group A = All schools that tested on April 14, 15 or 21. Group B = Removing schools that tested everyone on April 14, 15 or 21. Group C = Removing students who tested on April 14, 15 or 21. *d* = Cohen's *d*. Schools with less than 10 students testing in a grade were excluded.

[a] Cohen's *d* comparing the average school scores removing schools that tested everyone on April 14, 15, and 21 and the average school scores removing students who tested on April 14, 15 and 21. These two samples were compared because the sample included the same schools. Positive values indicate that the scores removing students who tested on April 14, 15, and 21 were higher than the observed scores.

Table 14. Mean Scores for Schools that Tested on April 21

Grade	Group A *n*	Group A Mean θ_1	Group A *SD*	Group B *n*	Group B Mean θ_1	Group B *SD*	Group C *n*	Group C Mean θ_1	Group C *SD*	Avg. % of Students Removed	$\theta_2 - \theta_3$	*d*[a]
Reading												
3	370	0.06	0.48	344	0.05	0.48	344	-0.01	0.69	55.2%	-0.06	-0.10
4	391	0.04	0.48	364	0.04	0.48	364	-0.08	0.68	54.6%	-0.13	-0.21
5	367	0.01	0.50	342	0.01	0.49	342	-0.09	0.69	51.3%	-0.10	-0.17
6	218	0.01	0.53	199	0.00	0.54	199	-0.18	0.75	52.7%	-0.18	-0.27
7	195	-0.17	0.54	179	-0.17	0.53	179	-0.28	0.73	53.2%	-0.11	-0.17
8	190	-0.12	0.53	176	-0.13	0.54	176	-0.31	0.73	56.4%	-0.18	-0.28
10	138	-0.38	0.53	131	-0.38	0.53	131	-0.55	0.63	55.4%	-0.17	-0.29
Math												
3	275	-0.09	0.53	257	-0.09	0.54	257	-0.13	0.65	44.8%	-0.03	-0.06
4	241	0.01	0.54	228	0.01	0.54	228	-0.06	0.71	43.2%	-0.07	-0.11
5	241	0.02	0.55	234	0.03	0.54	234	-0.08	0.69	44.8%	-0.11	-0.18
6	159	-0.06	0.60	150	-0.05	0.61	150	-0.16	0.78	47.6%	-0.11	-0.16
7	151	-0.14	0.57	144	-0.14	0.58	144	-0.31	0.84	48.6%	-0.17	-0.24
8	137	-0.17	0.56	129	-0.17	0.57	129	-0.36	0.77	52.1%	-0.19	-0.29
11	155	-0.25	0.60	147	-0.27	0.61	147	-0.36	0.72	53.3%	-0.10	-0.15

Note. Group A = All Schools that Tested on April 21. Group B = Removing Schools that Tested Everyone on April 21. Group C = Removing Students Who Tested on 21. *d* = Cohen's *d*. Schools with less than 10 students testing in a grade were excluded.

[a] Cohen's *d* comparing the average school scores removing schools that tested everyone on April 21 and the average school scores removing students who tested on April 21. These two samples were compared because the sample included the same schools. Positive values indicate that the scores removing students who tested on April 21 were higher than the observed scores.

Table 15. Mean Scores for Schools with Students in the Disrupted Sample

Grade	*n*	Score A Mean θ (*SD*)	Score B Mean θ (*SD*)	*d*[a]	Score C Mean θ (*SD*)	Avg. % of Students Removed	*d*[b]
Reading							
3	87	0.13 (0.49)	0.13 (0.49)	0.00	0.13 (0.51)	8.2%	0.00
4	111	0.09 (0.45)	0.09 (0.45)	-0.01	0.09 (0.47)	9.1%	0.01
5	94	0.10 (0.41)	0.09 (0.41)	0.00	0.09 (0.42)	10.4%	-0.01
6	53	0.06 (0.50)	0.04 (0.50)	-0.03	0.05 (0.51)	12.4%	-0.01
7	42	-0.15 (0.52)	-0.14 (0.51)	0.03	-0.16 (0.53)	11.5%	-0.02
8[c]	34	-0.05 (0.35)	-0.04 (0.35)	0.02	-0.07 (0.36)	9.8%	-0.04
10	31	-0.40 (0.57)	-0.40 (0.60)	0.00	-0.41 (0.59)	9.3%	-0.03
Math							
3	51	-0.07 (0.57)	-0.07 (0.57)	0.00	-0.08 (0.58)	7.3%	-0.01
4	53	0.05 (0.55)	0.06 (0.54)	0.00	0.05 (0.56)	9.1%	-0.01
5	62	0.18 (0.46)	0.18 (0.45)	0.00	0.17 (0.47)	9.2%	-0.02
6	41	0.11 (0.39)	0.11 (0.41)	-0.02	0.08 (0.45)	9.9%	-0.07
7	38	-0.04 (0.54)	-0.04 (0.54)	0.02	-0.06 (0.54)	8.4%	-0.02
8	31	-0.21 (0.64)	-0.21 (0.65)	0.01	-0.22 (0.66)	10.9%	-0.01
11	34	-0.02 (0.47)	-0.04 (0.46)	-0.04	-0.02 (0.53)	14.1%	-0.01

Note. Score A = School scores for schools with students in the disrupted sample. Score B = School scores replacing observed scores with predicted scores for students in the disrupted sample. Score C = school scores removing students in the disrupted sample. *d* = Cohen's *d*. Schools with less than 10 students testing in a grade were excluded.

[a] Cohen's *d* comparing the average school scores for schools with students in the disrupted sample and the average school scores replacing observed scores with predicted scores for students in the disrupted sample. Positive values indicate that the scores using the predicted probability were higher than the observed scores.

[b] Cohen's d comparing the average school scores for schools with students in the disrupted sample and the average school scores removing students in the disrupted sample. Positive values indicate that the scores removing students with disruptions were higher than the observed scores.

The results, in Tables 13 and 14, suggest that there are small to moderate differences depending on the alternative approach for calculating school means. The largest differences in school-level means were seen when excluding all students who tested on April 14, 15, or 21. The Cohen's *d*'s ranged from -.05 to -.28. When these students were excluded, the school-level means dropped on average by .12 theta points and the differences ranged from .03 to .17. This suggests removing all students who tested on April 14, 15, and 21 adversely affects school-level scores. Similar results were seen when removing students that tested on April 21. School-level scores dropped on average .12 theta points with differences ranging from .03 to .19, when all students who tested on April 21 were removed from the school-level score.

From Table 15, the difference in school-level means, after removing students in the disrupted sample, were small for most grades. The Cohen's *d* ranged from -.07 to.01. For most grades in math and reading, the mean theta score was lower after removing students in the disrupted sample. There were two exceptions. For reading grade 3 and reading grade 4, there was a slight increase in school-level scores after removing the students in the disrupted sample (d = .0001 and .01, respectively). Finally, when the predicted scores were substituted for the observed scores the differences in school-level means varied by grade and subject. For reading

grades 4, 5, 6 and 10 and math grades 3, 5, 6 and 11, there was a very small decrease in school means, with an average decrease of .006 theta points. For the remaining grades, the means increased. The effect size for all of the grades and subjects were small ranging from -.04 to .03.

School-Level Classification Differences

Lastly, we examined the impact of disruptions on the percent of students classified as proficient (Meets Expectations) for schools by grade. We calculated percent proficient using the observed scores and replacing observed scores with predicted scores for students in the disrupted sample. Again, it is important to interpret these results with the understanding that differences in these estimates are not *only* a result of disruptions. The predicted scores account for the general uncertainties in the estimation of a student's ability. Therefore, differences cannot be attributed solely to the disruption. However if the percentage of proficient students are not different, then evidence is provided that the disruptions did not adversely impact classification. We also calculated percent proficient removing students in the disrupted sample. Table 16 presents the results.

Table 16. Difference in School Classification for Schools with Students in the Disrupted Sample

Grade	*n*	Observed % Proficient Mean (*SD*)	Predicted % Proficient Mean (*SD*)	*d*[a]	% Proficient Removing Students in Disruption Mean (*SD*)	*d*[b]
Reading						
3	87	61.83% (18.43%)	62.16% (18.76%)	0.02	62.05% (18.84%)	0.01
4	111	57.14% (16.74%)	57.08% (17.00%)	0.00	57.42% (17.41%)	0.02
5	94	67.29% (14.47%)	67.26% (14.79%)	0.00	67.27% (15.24%)	0.00
6	53	60.36% (17.79%)	58.23% (19.66%)	-0.11	60.34% (18.00%)	0.00
7	42	48.36% (18.85%)	48.62% (18.83%)	0.01	47.51% (19.98%)	-0.04
8[c]	34	51.89% (13.20%)	51.32% (13.41%)	-0.04	51.46% (12.90%)	-0.03
10	31	47.40% (19.76%)	47.26% (20.11%)	-0.01	47.04% (19.73%)	-0.02
Math						
3	51	68.84% (20.38%)	69.04% (20.92%)	0.01	68.60% (20.89%)	-0.01
4	52	68.74% (19.36%)	68.65% (19.25%)	0.00	68.53% (19.78%)	-0.01
5	62	60.95% (17.11%)	61.20% (16.60%)	0.02	61.06% (17.57%)	0.01
6	41	55.65% (15.76%)	54.86% (17.11%)	-0.05	54.35% (18.25%)	-0.08
7	38	51.06% (19.68%)	50.99% (19.81%)	0.00	50.68% (19.61%)	-0.02
8	30	48.54% (21.73%)	48.29% (21.82%)	-0.01	48.67% (22.18%)	0.01
11	33	46.54% (18.78%)	44.46% (18.60%)	-0.11	46.66% (21.00%)	0.01

Note. *d* = Cohen's *d*. Schools with less than 10 students testing in a grade were excluded.
[a] Cohen's *d* comparing the observed percent proficient to the predicted percent proficient. Positive values indicate that the scores using the predicted probability were higher than the observed scores.
[b] Cohen's d comparing the observed percent proficient to the percent proficient removing students in the disruption sample. Positive values indicate that the scores removing students in the disruption sample were higher than the observed scores
[c] One school was removed from grade 8 reading because no school level information was available to compute predicted theta scores.

The results are inconsistent across grades. For several grades the percent of students classified as proficient increased when the predicted score was used or when the disrupted students were removed from the calculation. For other grades, the percent of students classified as proficient decreased when the predicted scores were used or when the disrupted students were removed. Overall the differences were very small with Cohen's *d* estimates ranging from -.11 to .02.

School-Level Summary

School-level accountability is based on the aggregation of student-level scores. Our investigation examined the effect of removing students that were potentially disrupted from school-level scores. Overall, our results indicate excluding students that were potentially disrupted result in negligible change to school-level scores. If the state were to take a broader approach to adjusting scores by excluding all students who tested on April 21 or all students that tested on April 14, 15, or 21, then schools would see lower overall test scores.

Conclusions

This report provides a statistical investigation of the potential impact of computer disruptions on student- and school-level scores for the MCA-III. The investigation is based on the well-accepted premise that students' test scores tend to exhibit consistency over time; that is, students who do well in one year of the test tend to do well the subsequent year. However, there could very well be students who scored higher than expected because they had a very good, productive year in school and were simply better prepared. Likewise, there could very well be students who scored lower than expected due to illness or other disruptions in their personal lives that temporarily lowered their ability to perform and show what they actually knew. We cannot know for certain that an *individual* student's performance was specifically impacted by the disruption and not some other event in the student's life that resulted in them performing better or worse than expected. However, we can use the results of the analyses to make an assessment about whether there are *trends* in the data that suggest the disruption had a systematic impact on student performance.

There is no statistical evidence to suggest that the disruption, on average, adversely impacted students, who were testing when the DDoS attack occurred on April 21. Although there is some evidence to suggest differences in scores when comparing students who were disrupted with those who weren't, the effects of the disruption were not in a consistent direction. This indicates that for some grades and subjects the disruption was beneficial and in other grades the disruption was detrimental. And, for most grades there was no impact at all. Additionally, any observed effects were small suggesting that any adjustment based on the effect size would be inconsequential. Overall, based on the statistical analyses described in this report, students observed score is the best estimate of achievement for this year. This report is intended to inform MDE and the Minnesota TAC of the statistical impact of computer disruption and to be used as a piece of evidence in considering whether policy actions are appropriate.

References

Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulation. *Biometrical Journal, 51,* 171-184. doi: 10.1002/bimj.200810488

Connelly, B. S., Sackett, P. R., & Waters, S. D. (2015). Balancing treatment and control groups in quasi-experiments: An Introduction to propensity scoring. *Personnel Psychology, 66,* 407-442. doi: 10.1111/peps.12020

Appendix A

Mean Differences between Non-Disrupted and Disrupted Samples before and after Propensity Matching

Table A.1. Mean Covariate Differences Before Matching for Reading Grade 3

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.014	0.003	0.015	0.003	**-0.205**
2015 Math Scale Scores	-0.019	1.108	-0.010	1.029	-0.009
School Free/Reduced Lunch	0.413	0.240	0.387	0.208	**0.109**
School Achievement	0.048	0.440	0.104	0.409	**-0.129**
School SPE	0.124	0.053	0.120	0.041	0.075
Indian	0.026	0.159	0.027	0.163	-0.008
Asian	0.077	0.267	0.085	0.279	-0.028
Hispanic	0.095	0.293	0.100	0.300	-0.018
Black	0.127	0.333	0.105	0.307	0.066
White	0.674	0.469	0.683	0.466	-0.020
Male	0.509	0.500	0.458	0.499	**0.104**
Free/Reduced Lunch	0.420	0.494	0.376	0.485	0.088
Limited English Proficiency	0.126	0.332	0.127	0.333	-0.004

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 41,683 – 42,249 for the non-disrupted group and 589 – 590 for the disrupted group.

Table A.2. Mean Covariate Differences After Matching for Reading Grade 3

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.015	0.003	0.015	0.003	0.000
2015 Math Scale Scores	-0.063	1.108	-0.010	1.029	-0.049
School Free/Reduced Lunch	0.391	0.225	0.386	0.209	0.020
School Achievement	0.118	0.423	0.104	0.410	0.034
School SPE	0.120	0.035	0.120	0.041	-0.009
Indian	0.017	0.129	0.027	0.163	-0.070
Asian	0.097	0.296	0.085	0.279	0.041
Hispanic	0.105	0.307	0.100	0.300	0.017
Black	0.102	0.303	0.105	0.307	-0.011
White	0.679	0.467	0.683	0.466	-0.007
Male	0.474	0.500	0.457	0.499	0.034
Free/Reduced Lunch	0.416	0.493	0.375	0.485	0.083
Limited English Proficiency	0.131	0.337	0.127	0.334	0.010

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 589 after matching.

Table A.3. Mean Covariate Differences Before Matching for Reading Grade 4

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.021	0.005	0.022	0.005	**-0.219**
2014 Math Scale Scores	0.017	1.103	-0.050	1.050	0.061
2014 Reading Scale Scores	0.018	1.123	-0.061	1.078	0.070
School Free/Reduced Lunch	0.408	0.235	0.406	0.247	0.009
School Achievement	0.047	0.437	0.058	0.433	-0.025
School SPE	0.127	0.056	0.123	0.030	0.072
Indian	0.023	0.151	0.016	0.126	0.049
Asian	0.079	0.270	0.096	0.295	-0.064
Hispanic	0.090	0.286	0.115	0.319	-0.086
Black	0.122	0.327	0.143	0.350	-0.065
White	0.685	0.465	0.630	0.483	**0.118**
Male	0.507	0.500	0.463	0.499	0.088
Free/Reduced Lunch	0.407	0.491	0.409	0.492	-0.004
Limited English Proficiency	0.101	0.301	0.143	0.350	**-0.140**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 39.649 – 41,280 for the non-disrupted group and 845 – 873 for the disrupted group.

Table A.4. Mean Covariate Differences After Matching for Reading Grade 4

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.022	0.005	0.022	0.005	0.000
2014 Math Scale Scores	-0.059	1.138	-0.038	1.041	-0.019
2014 Reading Scale Scores	-0.094	1.149	-0.061	1.079	-0.030
School Free/Reduced Lunch	0.404	0.239	0.404	0.247	0.002
School Achievement	0.063	0.448	0.061	0.431	0.004
School SPE	0.123	0.037	0.123	0.030	0.002
Indian	0.012	0.108	0.015	0.123	-0.031
Asian	0.100	0.300	0.095	0.293	0.016
Hispanic	0.115	0.319	0.115	0.319	0.000
Black	0.153	0.360	0.135	0.342	0.051
White	0.621	0.485	0.640	0.480	-0.039
Male	0.492	0.500	0.462	0.499	0.059
Free/Reduced Lunch	0.403	0.491	0.403	0.491	0.000
Limited English Proficiency	0.127	0.333	0.139	0.346	-0.035

Note. Sample size is 844 after matching.

Table A.5. Mean Covariate Differences Before Matching for Reading Grade 5

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.017	0.003	0.017	0.003	**-0.193**
2014 Math Scale Scores	0.070	1.108	0.090	1.088	-0.017
2014 Reading Scale Scores	0.019	1.150	-0.005	1.137	0.021
School Free/Reduced Lunch	0.405	0.235	0.396	0.212	0.036
School Achievement	0.051	0.437	0.093	0.380	-0.098
School SPE	0.129	0.056	0.123	0.033	0.104
Indian	0.022	0.148	0.021	0.143	0.011
Asian	0.082	0.274	0.077	0.267	0.016
Hispanic	0.090	0.287	0.095	0.294	-0.018
Black	0.122	0.327	0.136	0.343	-0.042
White	0.683	0.465	0.671	0.470	0.026
Male	0.506	0.500	0.481	0.500	0.050
Free/Reduced Lunch	0.400	0.490	0.394	0.489	0.012
Limited English Proficiency	0.090	0.286	0.094	0.292	-0.014

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 41,724 – 43,354for the non-disrupted group and 701 – 723 for the disrupted group.

Table A.6. Mean Covariate Differences After Matching for Reading Grade 5

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.017	0.003	0.017	0.003	0.000
2014 Math Scale Scores	0.089	1.113	0.104	1.073	-0.014
2014 Reading Scale Scores	0.004	1.134	-0.007	1.137	0.010
School Free/Reduced Lunch	0.396	0.238	0.396	0.213	0.001
School Achievement	0.093	0.411	0.094	0.383	-0.003
School SPE	0.124	0.033	0.123	0.033	0.018
Indian	0.020	0.140	0.021	0.145	-0.010
Asian	0.085	0.278	0.079	0.270	0.021
Hispanic	0.103	0.304	0.095	0.293	0.029
Black	0.117	0.322	0.130	0.337	-0.039
White	0.675	0.469	0.675	0.469	0.000
Male	0.424	0.495	0.479	0.500	**-0.109**
Free/Reduced Lunch	0.405	0.491	0.388	0.488	0.035
Limited English Proficiency	0.097	0.297	0.092	0.289	0.020

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 698 after matching.

Table A.7. Mean Covariate Differences Before Matching for Reading Grade 6

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.014	0.009	0.020	0.012	**-0.741**
2014 Math Scale Scores	0.160	1.097	0.475	1.021	**-0.287**
2014 Reading Scale Scores	0.099	1.130	0.339	1.066	**-0.213**
School Free/Reduced Lunch	0.389	0.212	0.295	0.213	**0.444**
School Achievement	0.068	0.414	0.244	0.397	**-0.425**
School SPE	0.136	0.061	0.118	0.036	**0.294**
Indian	0.022	0.148	0.017	0.130	0.034
Asian	0.077	0.267	0.101	0.301	-0.088
Hispanic	0.082	0.274	0.058	0.234	0.087
Black	0.112	0.315	0.091	0.288	0.065
White	0.706	0.456	0.733	0.443	-0.059
Male	0.507	0.500	0.505	0.500	0.005
Free/Reduced Lunch	0.389	0.487	0.261	0.440	**0.263**
Limited English Proficiency	0.071	0.256	0.030	0.170	**0.159**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 42,381– 45,105 for the non-disrupted group and 599 – 636 for the disrupted group.

Table A.8. Mean Covariate Differences After Matching for Reading Grade 6

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.020	0.012	0.020	0.012	0.000
2014 Math Scale Scores	0.464	1.055	0.504	1.002	-0.039
2014 Reading Scale Scores	0.315	1.130	0.342	1.057	-0.025
School Free/Reduced Lunch	0.283	0.185	0.293	0.215	-0.051
School Achievement	0.263	0.345	0.250	0.398	0.033
School SPE	0.116	0.034	0.115	0.031	0.017
Indian	0.019	0.137	0.017	0.131	0.013
Asian	0.118	0.323	0.099	0.299	0.061
Hispanic	0.047	0.211	0.061	0.239	-0.062
Black	0.078	0.268	0.088	0.284	-0.038
White	0.738	0.440	0.735	0.442	0.008
Male	0.506	0.500	0.510	0.500	-0.007
Free/Reduced Lunch	0.251	0.434	0.262	0.440	-0.024
Limited English Proficiency	0.038	0.192	0.028	0.164	0.058

Note. Sample size is 577 after matching.

Table A.9. Mean Covariate Differences Before Matching for Reading Grade 7

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.010	0.004	0.012	0.008	**-0.557**
2014 Math Scale Scores	0.122	1.073	0.141	1.074	-0.018
2014 Reading Scale Scores	0.071	1.093	0.092	1.092	-0.020
School Free/Reduced Lunch	0.364	0.207	0.416	0.256	**-0.247**
School Achievement	0.072	0.411	0.028	0.442	**0.107**
School SPE	0.138	0.063	0.135	0.038	0.056
Indian	0.023	0.149	0.030	0.172	-0.052
Asian	0.070	0.255	0.103	0.304	**-0.129**
Hispanic	0.081	0.272	0.086	0.281	-0.021
Black	0.110	0.313	0.096	0.295	0.045
White	0.715	0.451	0.685	0.465	0.067
Male	0.511	0.500	0.514	0.500	-0.006
Free/Reduced Lunch	0.377	0.485	0.421	0.494	-0.089
Limited English Proficiency	0.066	0.249	0.126	0.332	**-0.239**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 41,439 – 44,740 for the non-disrupted group and 399 – 428 for the disrupted group.

Table A.10. Mean Covariate Differences After Matching for Reading Grade 7

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.012	0.008	0.012	0.008	0.000
2014 Math Scale Scores	0.106	1.091	0.139	1.075	-0.031
2014 Reading Scale Scores	0.097	1.120	0.100	1.089	-0.003
School Free/Reduced Lunch	0.407	0.233	0.412	0.255	-0.024
School Achievement	0.037	0.438	0.028	0.445	0.019
School SPE	0.132	0.040	0.135	0.038	-0.058
Indian	0.040	0.197	0.033	0.178	0.040
Asian	0.111	0.314	0.106	0.308	0.016
Hispanic	0.076	0.265	0.076	0.265	0.000
Black	0.113	0.317	0.096	0.295	0.058
White	0.660	0.474	0.690	0.463	-0.064
Male	0.499	0.501	0.511	0.501	-0.025
Free/Reduced Lunch	0.441	0.497	0.423	0.495	0.036
Limited English Proficiency	0.134	0.341	0.121	0.326	0.038

Note. Sample size is 397 after matching.

Table A.11. Mean Covariate Differences Before Matching for Reading Grade 8

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.009	0.006	0.012	0.009	**-0.562**
2014 Math Scale Scores	0.120	1.071	0.123	1.052	-0.002
2014 Reading Scale Scores	0.103	1.092	0.043	1.038	0.055
School Free/Reduced Lunch	0.366	0.208	0.459	0.228	**-0.443**
School Achievement	0.068	0.416	-0.024	0.377	**0.220**
School SPE	0.139	0.066	0.138	0.033	0.019
Indian	0.023	0.148	0.023	0.148	0.000
Asian	0.072	0.259	0.103	0.304	**-0.117**
Hispanic	0.078	0.268	0.118	0.322	**-0.149**
Black	0.106	0.308	0.120	0.325	-0.045
White	0.719	0.449	0.638	0.481	**0.182**
Male	0.512	0.500	0.480	0.500	0.063
Free/Reduced Lunch	0.371	0.483	0.450	0.498	**-0.163**
Limited English Proficiency	0.064	0.245	0.133	0.339	**-0.279**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 41,607– 44,549 for the non-disrupted group and 373 – 400 for the disrupted group.

Table A.12. Mean Covariate Differences After Matching for Reading Grade 8

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.012	0.009	0.012	0.009	0.000
2014 Math Scale Scores	0.031	1.089	0.129	1.047	-0.092
2014 Reading Scale Scores	0.045	1.161	0.064	1.019	-0.017
School Free/Reduced Lunch	0.449	0.224	0.447	0.221	0.008
School Achievement	-0.011	0.437	-0.010	0.373	-0.001
School SPE	0.137	0.041	0.134	0.030	0.083
Indian	0.039	0.194	0.025	0.157	0.079
Asian	0.100	0.301	0.103	0.304	-0.009
Hispanic	0.145	0.352	0.117	0.322	0.083
Black	0.075	0.264	0.095	0.293	-0.070
White	0.641	0.480	0.660	0.474	-0.041
Male	0.440	0.497	0.468	0.500	-0.056
Free/Reduced Lunch	0.418	0.494	0.429	0.496	-0.023
Limited English Proficiency	0.111	0.315	0.125	0.332	-0.043

Note. Sample size is 359 after matching.

Table A.13. Mean Covariate Differences Before Matching for Reading Grade 10

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.009	0.004	0.011	0.005	**-0.371**
2014 Math Scale Scores	0.169	1.054	0.315	1.004	**-0.138**
2014 Reading Scale Scores	0.065	1.067	0.070	0.997	-0.005
School Free/Reduced Lunch	0.323	0.196	0.371	0.187	**-0.246**
School Achievement	0.009	0.445	-0.122	0.414	**0.296**
School SPE	0.129	0.071	0.132	0.045	-0.046
Indian	0.018	0.133	0.012	0.109	0.047
Asian	0.072	0.259	0.062	0.241	0.040
Hispanic	0.068	0.251	0.088	0.284	-0.081
Black	0.096	0.294	0.100	0.300	-0.014
White	0.745	0.436	0.738	0.440	0.015
Male	0.509	0.500	0.500	0.501	0.018
Free/Reduced Lunch	0.329	0.470	0.360	0.480	-0.065
Limited English Proficiency	0.043	0.203	0.081	0.273	**-0.186**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 41,282 – 46,774 for the non-disrupted group and 376– 420 for the disrupted group.

Table A.14. Mean Covariate Differences After Matching for Reading Grade 10

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.011	0.005	0.011	0.005	0.000
2014 Math Scale Scores	0.375	0.990	0.332	0.998	0.043
2014 Reading Scale Scores	0.068	1.109	0.085	0.992	-0.017
School Free/Reduced Lunch	0.342	0.195	0.348	0.149	-0.031
School Achievement	-0.044	0.437	-0.071	0.309	0.071
School SPE	0.135	0.073	0.135	0.043	0.000
Indian	0.003	0.052	0.013	0.115	**-0.129**
Asian	0.070	0.256	0.065	0.246	0.021
Hispanic	0.059	0.237	0.078	0.269	-0.075
Black	0.062	0.241	0.067	0.251	-0.022
White	0.806	0.396	0.776	0.417	0.073
Male	0.504	0.501	0.491	0.501	0.027
Free/Reduced Lunch	0.310	0.463	0.326	0.469	-0.035
Limited English Proficiency	0.046	0.209	0.043	0.203	0.013

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 371 after matching.

Table A.15. Mean Covariate Differences Before Matching for Math Grade 3

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.006	0.002	0.007	0.003	**-0.415**
2015 Reading Scale Scores	0.066	1.171	-0.009	1.106	0.064
School Free/Reduced Lunch	0.408	0.236	0.410	0.301	-0.011
School Achievement	0.056	0.431	-0.010	0.544	**0.152**
School SPE	0.125	0.049	0.125	0.050	-0.012
Indian	0.023	0.151	0.039	0.194	-0.104
Asian	0.075	0.264	0.075	0.263	0.002
Hispanic	0.090	0.286	0.156	0.363	**-0.229**
Black	0.125	0.330	0.081	0.274	**0.131**
White	0.686	0.464	0.649	0.478	0.079
Male	0.510	0.500	0.458	0.499	**0.103**
Free/Reduced Lunch	0.412	0.492	0.445	0.498	-0.067
Limited English Proficiency	0.121	0.326	0.166	0.372	**-0.137**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 52,076 – 52,364 for the non-disrupted group and 307 – 308 for the disrupted group.

Table A.16. Mean Covariate Differences After Matching for Math Grade 3

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.007	0.003	0.007	0.003	0.000
2015 Reading Scale Scores	0.092	1.250	-0.012	1.107	0.088
School Free/Reduced Lunch	0.412	0.254	0.410	0.300	0.007
School Achievement	0.000	0.468	-0.006	0.541	0.012
School SPE	0.122	0.034	0.126	0.050	-0.094
Indian	0.039	0.194	0.039	0.194	0.000
Asian	0.049	0.216	0.075	0.264	**-0.109**
Hispanic	0.173	0.379	0.154	0.361	0.053
Black	0.072	0.259	0.082	0.274	-0.037
White	0.667	0.472	0.650	0.478	0.034
Male	0.451	0.498	0.458	0.499	-0.013
Free/Reduced Lunch	0.412	0.493	0.444	0.498	-0.066
Limited English Proficiency	0.147	0.355	0.163	0.370	-0.045

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 306 after matching.

Table A.17. Mean Covariate Differences Before Matching for Math Grade 4

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.007	0.003	0.009	0.004	**-0.425**
2014 Math Scale Scores	0.056	1.080	0.135	1.137	-0.073
2014 Reading Scale Scores	0.055	1.103	0.068	1.140	-0.012
School Free/Reduced Lunch	0.408	0.235	0.357	0.252	**0.215**
School Achievement	0.062	0.430	0.131	0.474	**-0.161**
School SPE	0.126	0.052	0.116	0.030	**0.204**
Indian	0.023	0.151	0.026	0.159	-0.018
Asian	0.076	0.266	0.130	0.337	**-0.203**
Hispanic	0.088	0.284	0.060	0.238	**0.100**
Black	0.120	0.325	0.120	0.325	0.000
White	0.691	0.462	0.664	0.473	0.059
Male	0.502	0.500	0.466	0.500	0.072
Free/Reduced Lunch	0.403	0.490	0.349	0.477	**0.110**
Limited English Proficiency	0.101	0.302	0.141	0.348	**-0.131**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 49,278 – 51,402 for the non-disrupted group and 370 – 384 for the disrupted group.

Table A.18. Mean Covariate Differences After Matching for Math Grade 4

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.009	0.004	0.009	0.004	0.000
2014 Math Scale Scores	0.182	1.124	0.146	1.126	0.033
2014 Reading Scale Scores	0.106	1.115	0.069	1.138	0.033
School Free/Reduced Lunch	0.332	0.221	0.351	0.247	-0.083
School Achievement	0.187	0.394	0.144	0.464	**0.101**
School SPE	0.116	0.033	0.116	0.030	-0.007
Indian	0.014	0.116	0.027	0.163	-0.098
Asian	0.117	0.322	0.128	0.335	-0.033
Hispanic	0.074	0.261	0.054	0.227	0.078
Black	0.084	0.278	0.112	0.315	-0.092
White	0.711	0.454	0.678	0.468	0.071
Male	0.469	0.500	0.460	0.499	0.016
Free/Reduced Lunch	0.319	0.467	0.335	0.473	-0.035
Limited English Proficiency	0.139	0.346	0.125	0.332	0.040

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 367 after matching.

Table A.19. Mean Covariate Differences Before Matching for Math Grade 5

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.010	0.005	0.012	0.006	**-0.465**
2014 Math Scale Scores	0.094	1.082	0.288	1.026	**-0.180**
2014 Reading Scale Scores	0.035	1.126	0.189	1.088	**-0.137**
School Free/Reduced Lunch	0.395	0.227	0.338	0.228	**0.248**
School Achievement	0.069	0.418	0.214	0.374	**-0.346**
School SPE	0.129	0.054	0.117	0.029	**0.219**
Indian	0.024	0.152	0.014	0.117	0.064
Asian	0.070	0.256	0.136	0.343	-**0.254**
Hispanic	0.089	0.285	0.080	0.271	0.033
Black	0.116	0.320	0.098	0.297	0.056
White	0.700	0.458	0.671	0.470	0.065
Male	0.505	0.500	0.477	0.500	0.056
Free/Reduced Lunch	0.392	0.488	0.320	0.467	**0.148**
Limited English Proficiency	0.082	0.275	0.098	0.298	-0.057

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 50,063 – 52,191 for the non-disrupted group and 489 – 501 for the disrupted group.

Table A.20. Mean Covariate Differences After Matching for Math Grade 5

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.012	0.006	0.012	0.006	0.000
2014 Math Scale Scores	0.375	1.099	0.295	1.017	0.075
2014 Reading Scale Scores	0.285	1.088	0.195	1.082	0.083
School Free/Reduced Lunch	0.319	0.205	0.338	0.228	-0.086
School Achievement	0.230	0.352	0.214	0.373	0.044
School SPE	0.114	0.032	0.117	0.029	-0.086
Indian	0.012	0.110	0.014	0.119	-0.018
Asian	0.113	0.317	0.135	0.342	-0.068
Hispanic	0.068	0.251	0.080	0.271	-0.047
Black	0.105	0.306	0.096	0.295	0.027
White	0.703	0.457	0.674	0.469	0.062
Male	0.502	0.501	0.480	0.500	0.045
Free/Reduced Lunch	0.270	0.445	0.316	0.465	-0.099
Limited English Proficiency	0.076	0.265	0.096	0.295	-0.073

Note. Sample size is 488 after matching.

Table A.21. Mean Covariate Differences Before Matching for Math Grade 6

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.007	0.004	0.010	0.006	**-0.555**
2014 Math Scale Scores	0.163	1.094	0.433	0.976	**-0.247**
2014 Reading Scale Scores	0.099	1.130	0.217	0.996	**-0.104**
School Free/Reduced Lunch	0.381	0.219	0.445	0.231	**-0.292**
School Achievement	0.078	0.421	-0.011	0.342	**0.212**
School SPE	0.135	0.059	0.140	0.030	-0.077
Indian	0.023	0.151	0.048	0.214	**-0.163**
Asian	0.076	0.265	0.061	0.240	0.056
Hispanic	0.082	0.275	0.093	0.291	-0.040
Black	0.112	0.315	0.115	0.319	-0.008
White	0.705	0.456	0.680	0.467	0.055
Male	0.507	0.500	0.496	0.501	0.022
Free/Reduced Lunch	0.384	0.486	0.441	0.497	**-0.118**
Limited English Proficiency	0.074	0.261	0.037	0.190	**0.138**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 47,703 – 50,904 for the non-disrupted group and 347 – 375 for the disrupted group.

Table A.22. Mean Covariate Differences After Matching for Math Grade 6

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.010	0.006	0.010	0.006	0.000
2014 Math Scale Scores	0.526	1.075	0.460	0.943	0.065
2014 Reading Scale Scores	0.278	1.062	0.221	0.994	0.055
School Free/Reduced Lunch	0.429	0.230	0.440	0.227	-0.046
School Achievement	-0.002	0.442	-0.002	0.336	0.001
School SPE	0.138	0.040	0.141	0.030	-0.093
Indian	0.032	0.176	0.049	0.216	-0.088
Asian	0.058	0.234	0.058	0.234	0.000
Hispanic	0.078	0.269	0.087	0.282	-0.032
Black	0.107	0.309	0.104	0.306	0.009
White	0.725	0.447	0.702	0.458	0.051
Male	0.497	0.501	0.497	0.501	0.000
Free/Reduced Lunch	0.422	0.495	0.425	0.495	-0.006
Limited English Proficiency	0.012	0.107	0.020	0.141	-0.070

Note. Sample size is 346 after matching.

Table A.23. Mean Covariate Differences Before Matching for Math Grade 7

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.009	0.002	0.010	0.002	**-0.250**
2014 Math Scale Scores	0.136	1.069	0.184	1.011	-0.044
2014 Reading Scale Scores	0.071	1.090	0.115	1.018	-0.041
School Free/Reduced Lunch	0.362	0.210	0.367	0.173	-0.026
School Achievement	0.081	0.408	0.054	0.371	0.067
School SPE	0.137	0.063	0.139	0.034	-0.024
Indian	0.023	0.149	0.034	0.182	-0.079
Asian	0.073	0.260	0.080	0.272	-0.028
Hispanic	0.080	0.271	0.078	0.268	0.008
Black	0.107	0.309	0.121	0.327	-0.046
White	0.716	0.451	0.684	0.465	0.071
Male	0.511	0.500	0.467	0.499	0.089
Free/Reduced Lunch	0.377	0.485	0.335	0.472	0.087
Limited English Proficiency	0.074	0.261	0.057	0.233	0.063

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 45,430 – 49,001 for the non-disrupted group and 413 – 437 for the disrupted group.

Table A.24. Mean Covariate Differences After Matching for Math Grade 7

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.010	0.002	0.010	0.002	0.000
2014 Math Scale Scores	0.253	1.118	0.207	0.988	0.044
2014 Reading Scale Scores	0.178	1.121	0.115	1.019	0.059
School Free/Reduced Lunch	0.358	0.209	0.364	0.172	-0.032
School Achievement	0.081	0.426	0.061	0.371	0.052
School SPE	0.137	0.071	0.139	0.033	-0.034
Indian	0.020	0.138	0.029	0.169	-0.064
Asian	0.093	0.290	0.085	0.280	0.026
Hispanic	0.059	0.235	0.076	0.265	-0.068
Black	0.132	0.339	0.120	0.325	0.037
White	0.698	0.460	0.690	0.463	0.016
Male	0.451	0.498	0.468	0.500	-0.034
Free/Reduced Lunch	0.302	0.460	0.320	0.467	-0.037
Limited English Proficiency	0.056	0.230	0.051	0.221	0.022

Note. Sample size is 410 after matching.

Table A.25. Mean Covariate Differences Before Matching for Math Grade 8

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.007	0.006	0.012	0.013	**-0.813**
2014 Math Scale Scores	0.134	1.064	0.268	0.994	**-0.126**
2014 Reading Scale Scores	0.097	1.087	0.122	1.066	-0.023
School Free/Reduced Lunch	0.360	0.208	0.437	0.196	**-0.370**
School Achievement	0.074	0.407	-0.033	0.343	**0.264**
School SPE	0.139	0.066	0.127	0.065	**0.190**
Indian	0.023	0.148	0.017	0.129	0.038
Asian	0.069	0.254	0.141	0.348	**-0.281**
Hispanic	0.080	0.271	0.054	0.225	0.098
Black	0.107	0.309	0.096	0.295	0.035
White	0.720	0.449	0.690	0.463	0.066
Male	0.510	0.500	0.490	0.501	0.040
Free/Reduced Lunch	0.366	0.482	0.395	0.490	-0.062
Limited English Proficiency	0.068	0.251	0.082	0.275	-0.057

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 47,174 – 50,473 for the non-disrupted group and 340 – 355 for the disrupted group.

Table A.26. Mean Covariate Differences After Matching for Math Grade 8

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.012	0.013	0.012	0.013	0.000
2014 Math Scale Scores	0.296	0.965	0.285	0.979	0.011
2014 Reading Scale Scores	0.202	1.008	0.156	1.035	0.045
School Free/Reduced Lunch	0.439	0.241	0.434	0.193	0.022
School Achievement	-0.034	0.443	-0.031	0.342	-0.008
School SPE	0.127	0.050	0.125	0.046	0.038
Indian	0.018	0.133	0.018	0.133	0.000
Asian	0.122	0.328	0.137	0.344	-0.044
Hispanic	0.051	0.219	0.054	0.226	-0.013
Black	0.086	0.281	0.080	0.272	0.022
White	0.723	0.448	0.711	0.454	0.026
Male	0.488	0.501	0.485	0.501	0.006
Free/Reduced Lunch	0.405	0.492	0.384	0.487	0.043
Limited English Proficiency	0.039	0.193	0.063	0.242	-0.109

Note. Sample size is 336 after matching.

Table A.27. Mean Covariate Differences Before Matching for Math Grade 11

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.012	0.006	0.014	0.004	**-0.304**
2014 Math Scale Scores	0.197	0.969	0.301	0.839	**-0.107**
2014 Reading Scale Scores	-0.006	1.079	0.129	0.957	**-0.126**
School Free/Reduced Lunch	0.340	0.210	0.318	0.145	**0.105**
School Achievement	-0.044	0.488	0.026	0.328	**-0.144**
School SPE	0.131	0.076	0.139	0.039	**-0.105**
Indian	0.018	0.134	0.032	0.177	**-0.104**
Asian	0.075	0.264	0.063	0.243	0.047
Hispanic	0.063	0.242	0.055	0.228	0.033
Black	0.098	0.297	0.045	0.206	**0.179**
White	0.745	0.436	0.806	0.396	**-0.139**
Male	0.508	0.500	0.500	0.501	0.015
Free/Reduced Lunch	0.318	0.466	0.251	0.434	**0.144**
Limited English Proficiency	0.047	0.212	0.012	0.110	**0.165**

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample sizes before matching ranges from for 34,910 – 39,813 for the non-disrupted group and 444 – 494 for the disrupted group.

Table A.28. Mean Covariate Differences After Matching for Math Grade 11

	Non-Disrupted Mean	Non-Disrupted *SD*	Disrupted Mean	Disrupted *SD*	*d*
Predicted Probability	0.014	0.004	0.014	0.004	0.000
2014 Math Scale Scores	0.451	0.930	0.312	0.837	**0.158**
2014 Reading Scale Scores	0.171	1.026	0.099	0.945	0.074
School Free/Reduced Lunch	0.323	0.183	0.331	0.135	-0.048
School Achievement	0.091	0.402	0.027	0.323	**0.176**
School SPE	0.125	0.041	0.144	0.034	**-0.496**
Indian	0.024	0.153	0.036	0.186	-0.070
Asian	0.062	0.241	0.069	0.254	-0.029
Hispanic	0.038	0.192	0.040	0.197	-0.012
Black	0.029	0.167	0.029	0.167	0.000
White	0.848	0.360	0.826	0.379	0.058
Male	0.545	0.499	0.533	0.499	0.024
Free/Reduced Lunch	0.210	0.407	0.245	0.431	-0.085
Limited English Proficiency	0.005	0.069	0.007	0.084	-0.031

Note. Bolded values indicate Cohen's d values greater than .10 or -.10. Sample size is 420 after matching.

Appendix B. P-P Plots of the Difference between Predicted and Observed Theta

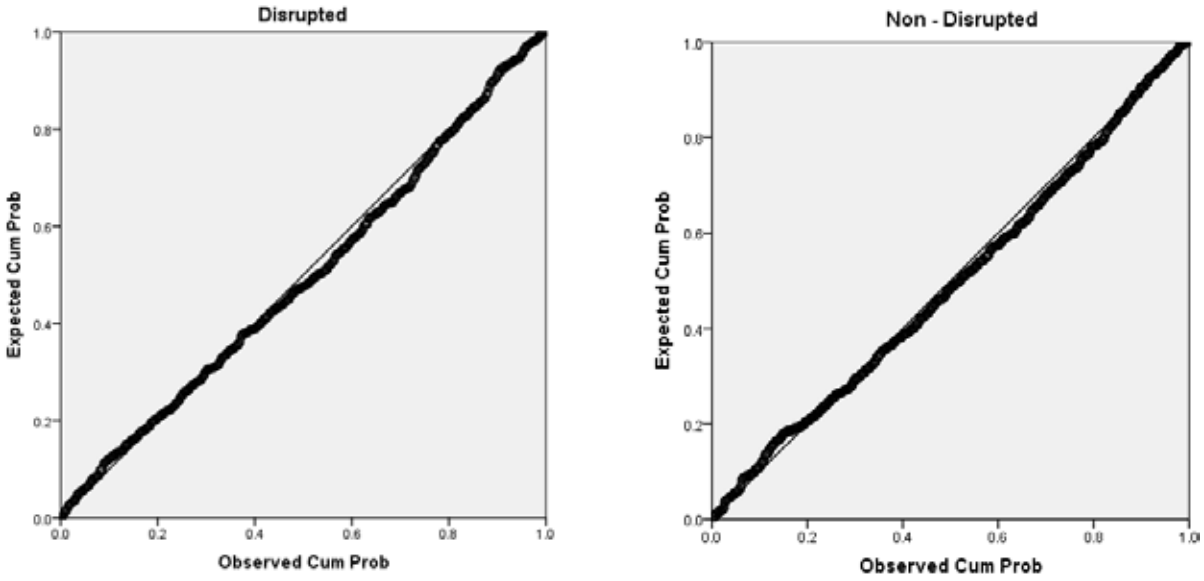

Figure B.1. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 3 Reading.

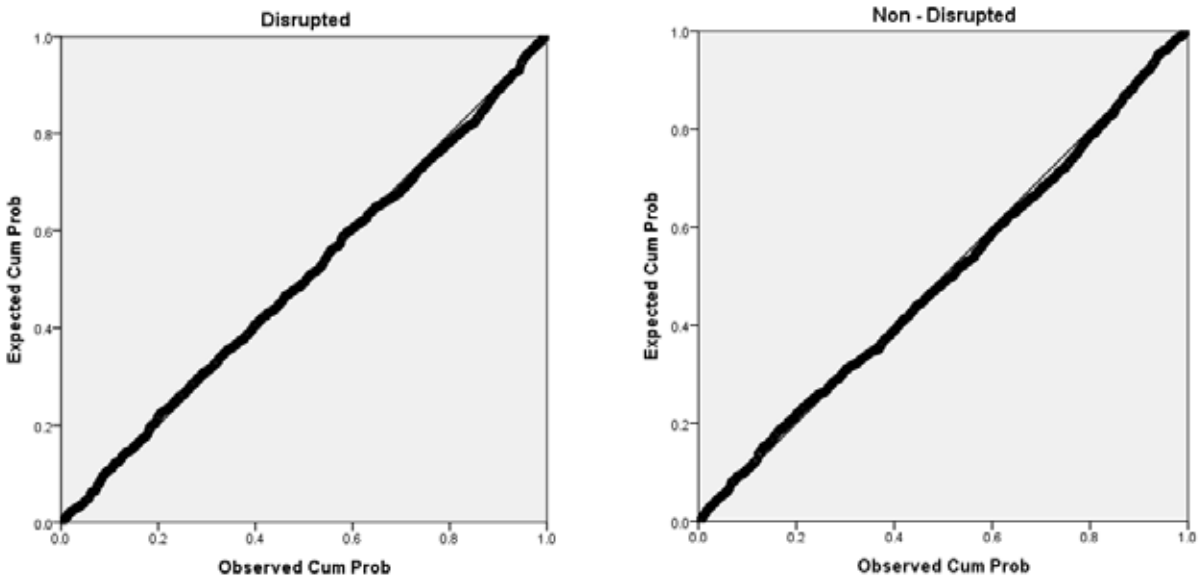

Figure B.2. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 4 Reading.

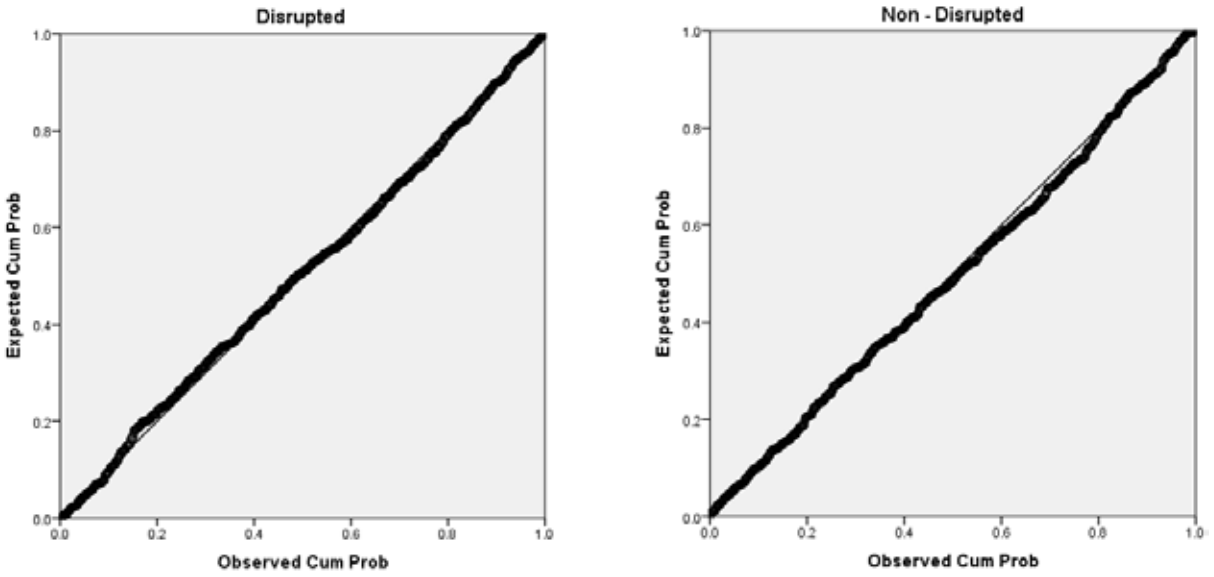

Figure B.3. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 5 Reading.

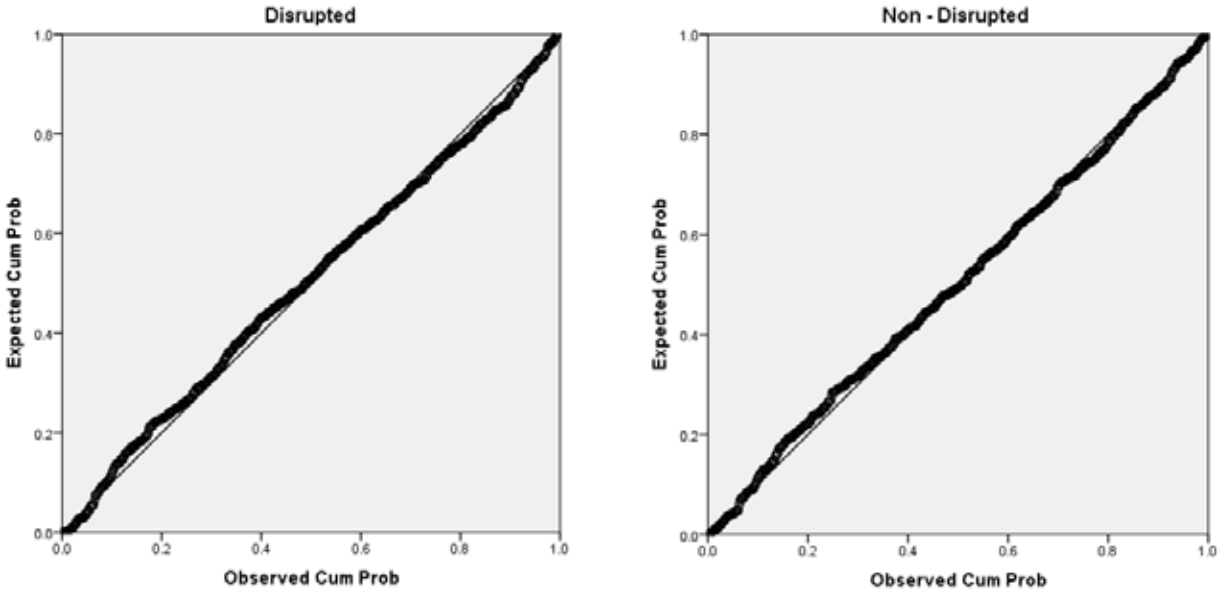

Figure B.4. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 6 Reading.

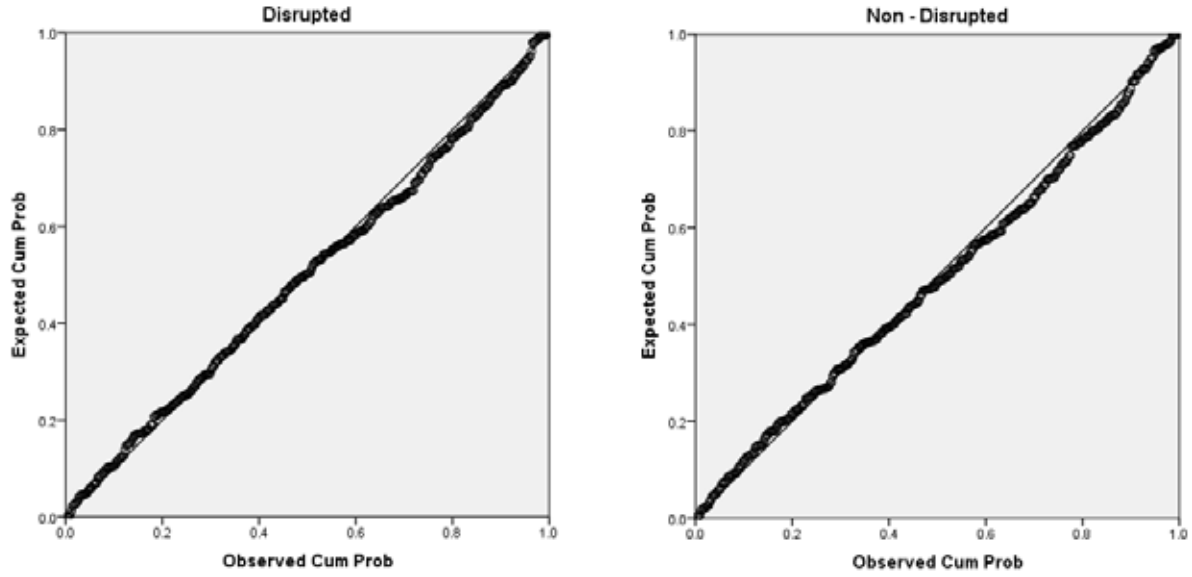

Figure B.5. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 7 Reading.

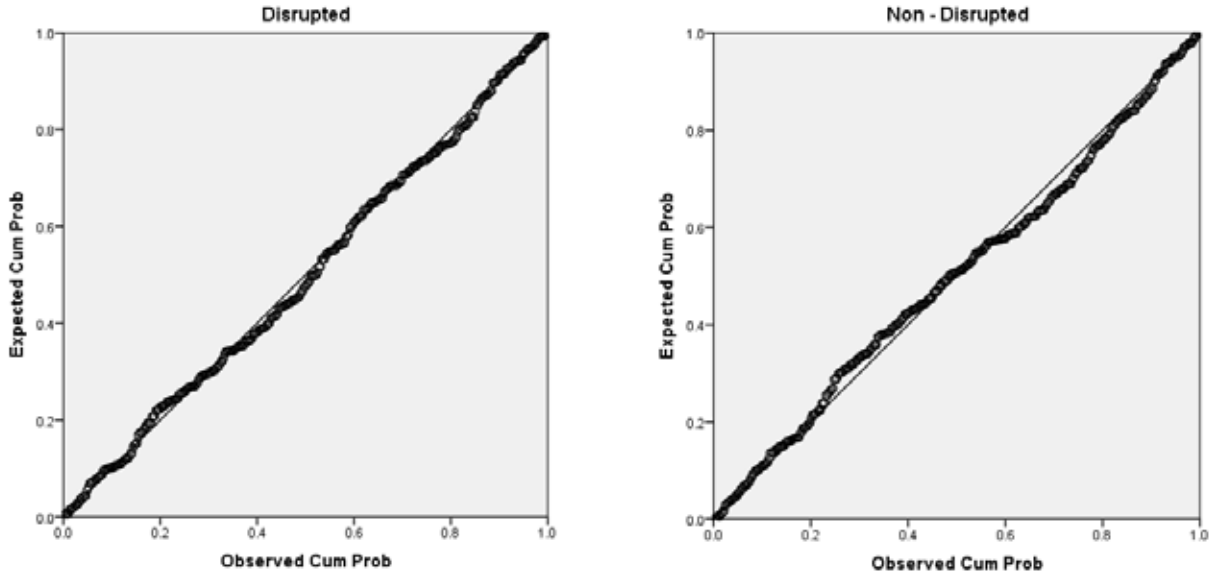

Figure B.6. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 8 Reading.

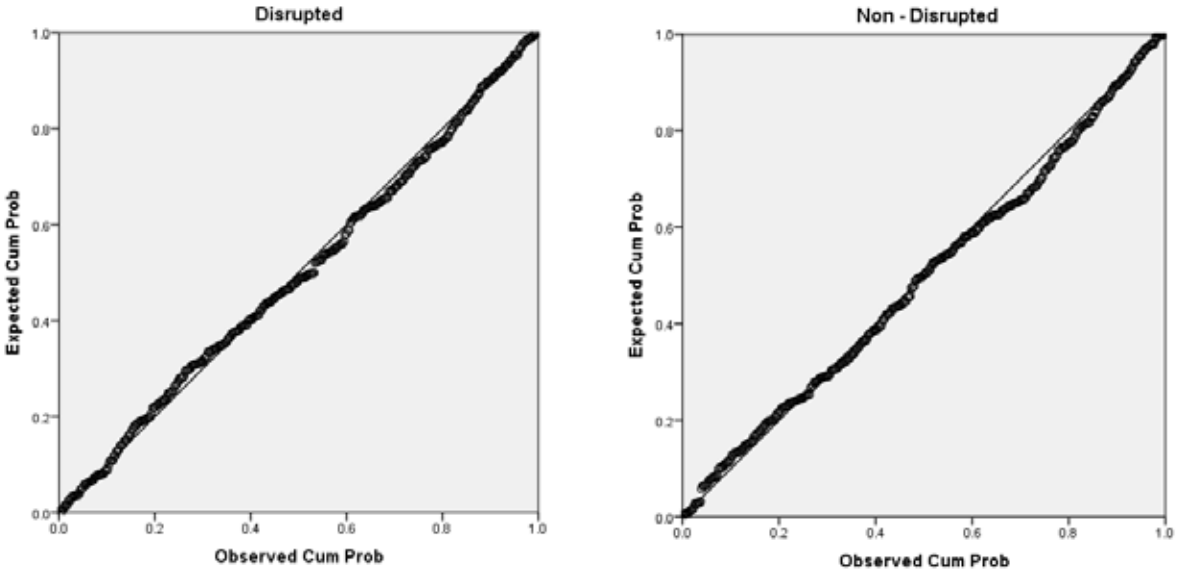

Figure B.7. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 10 Reading.

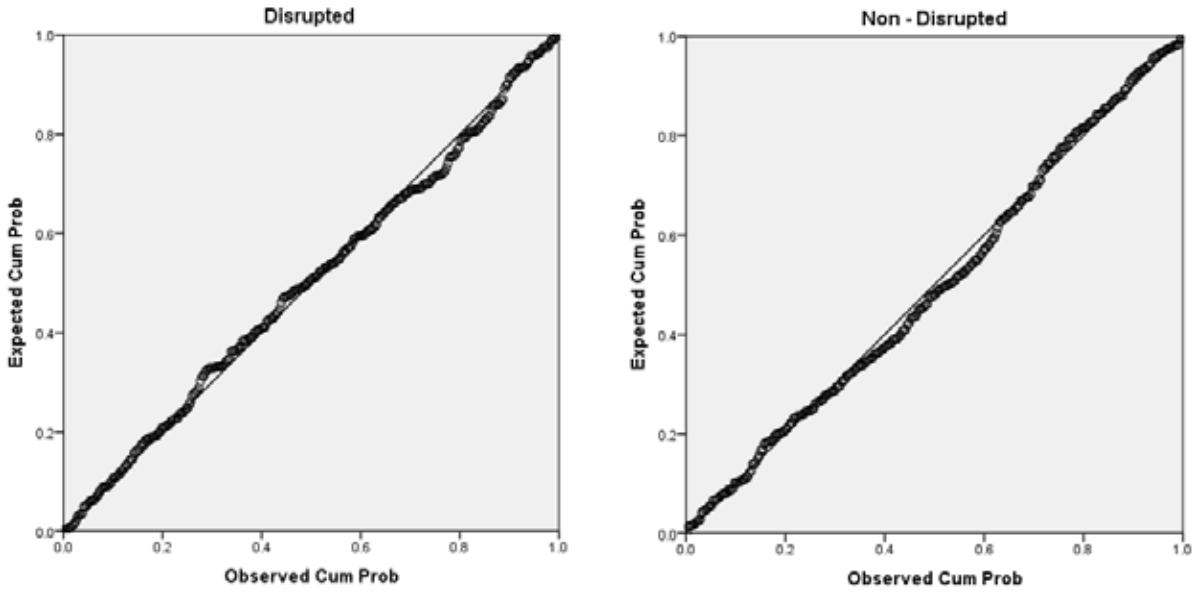

Figure B.8. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 3 Math.

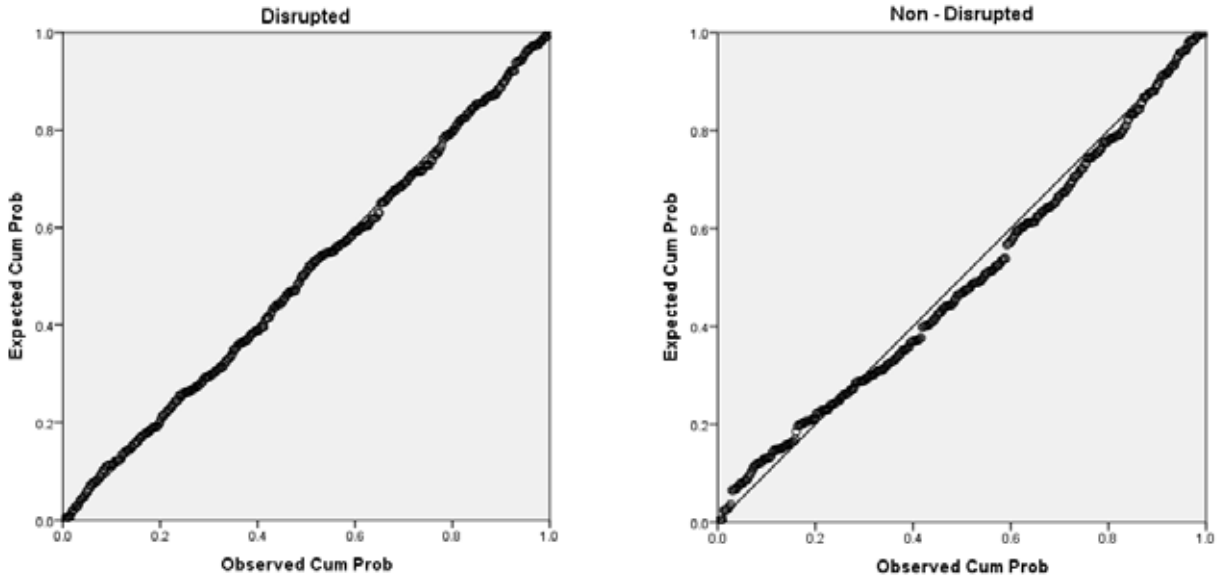

Figure B.9. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 4 Math.

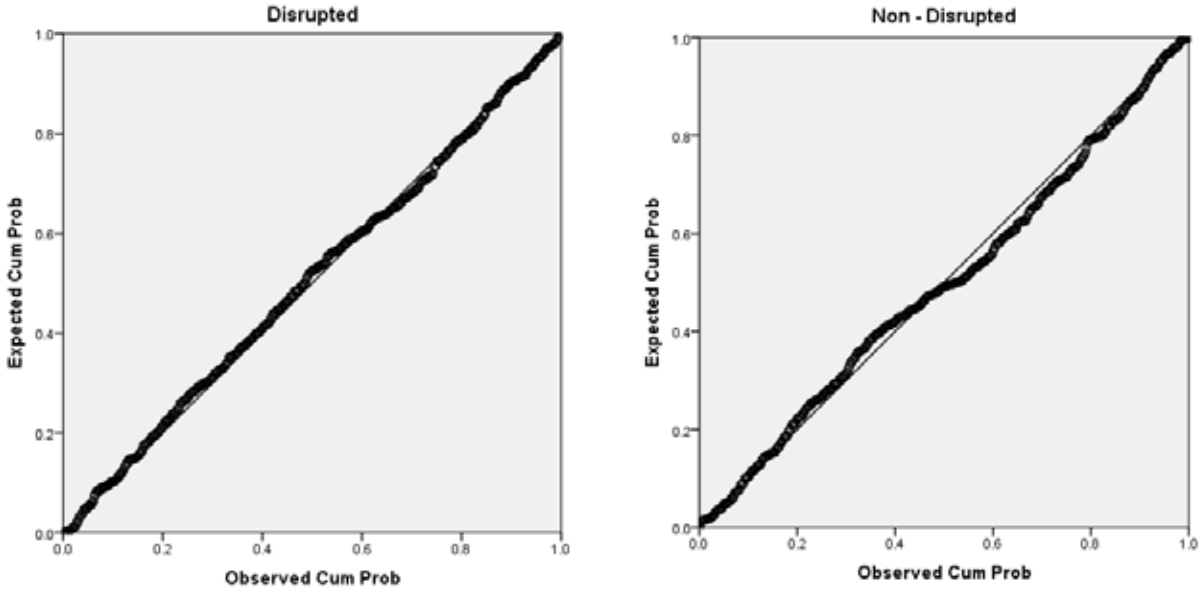

Figure B.10. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 5 Math.

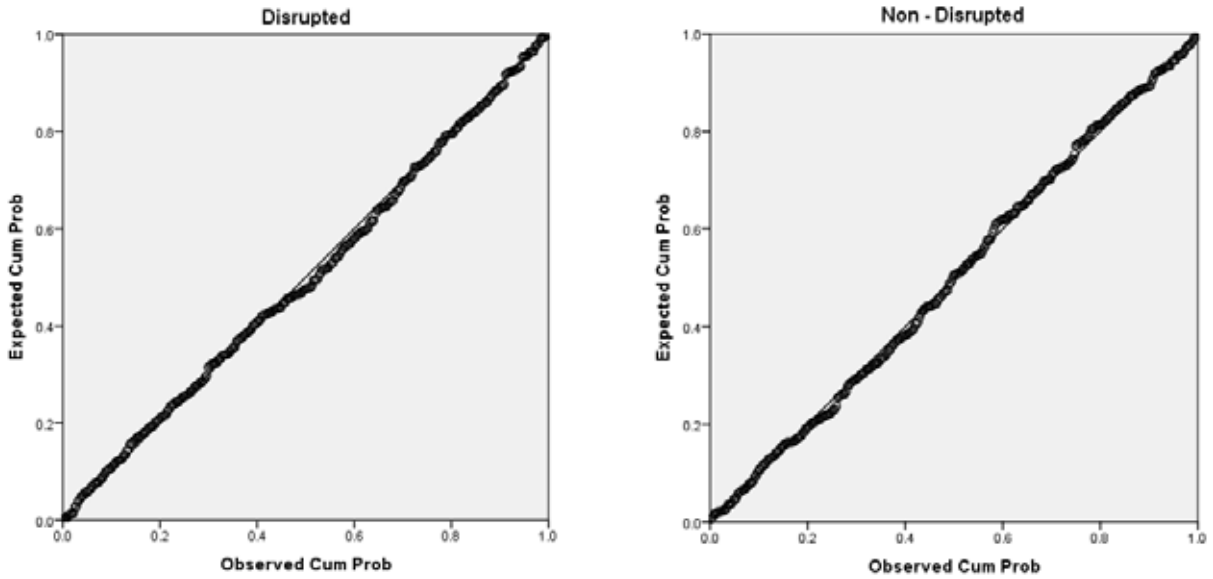

Figure B.11. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 6 Math.

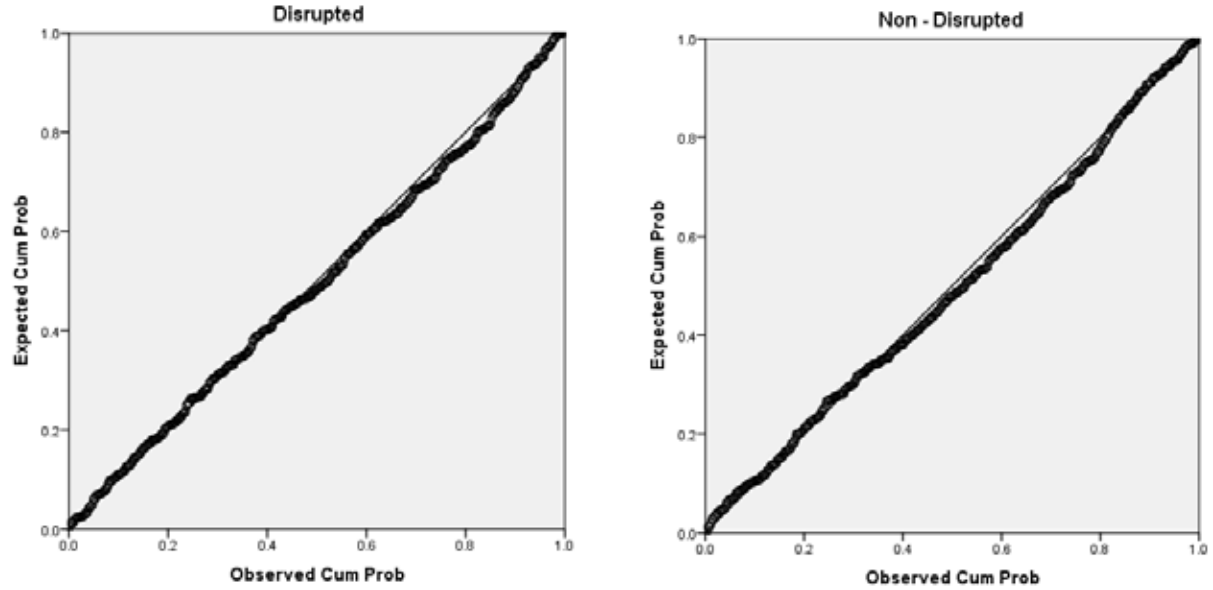

Figure B.12. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 7 Math.

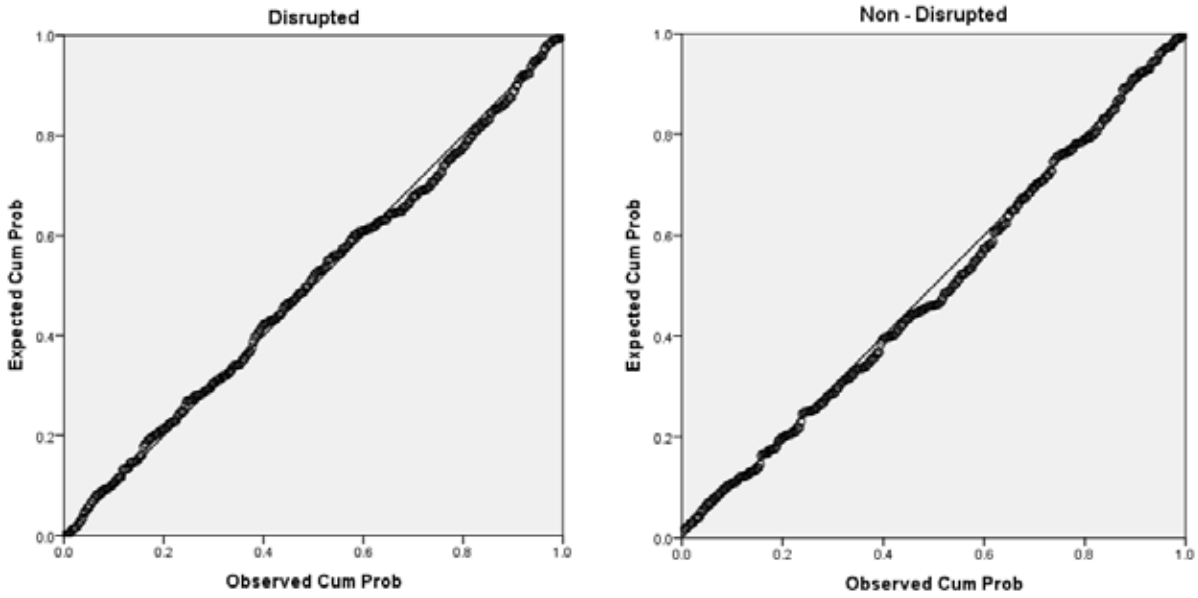

Figure B.13. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 8 Math.

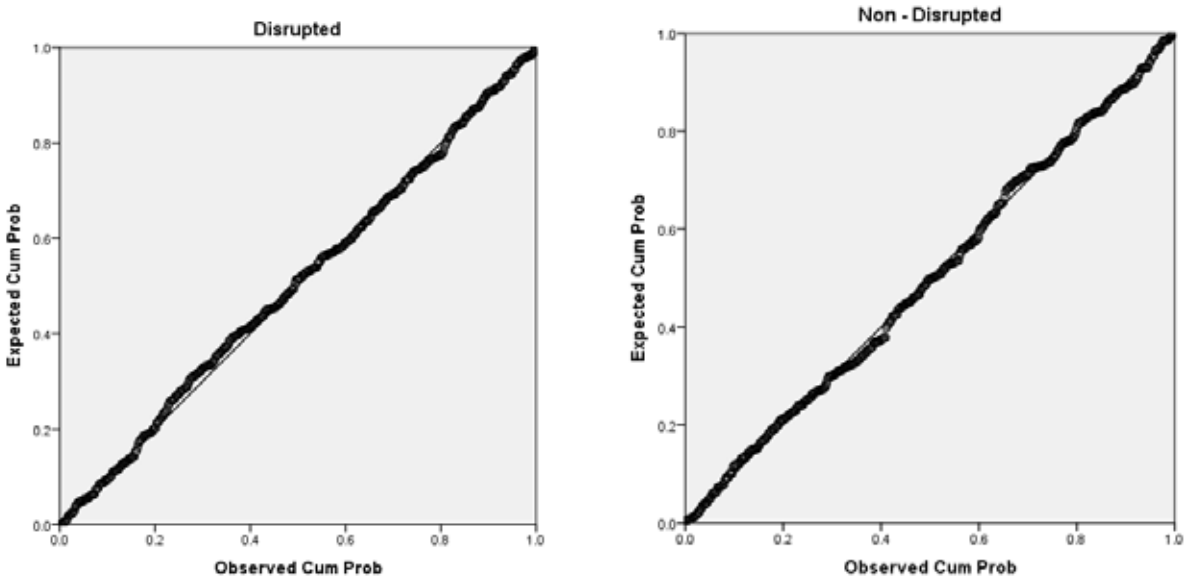

Figure B.14. Normal P-P plot of the Difference between Predicted and Observed Theta for Grade 11 Math.