J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 3, Number 6 · February 2005

# A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program

John Poggio, Douglas R. Glasnapp,
Xiangdong Yang, and Andrew J. Poggio

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 3, Number 6

# A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program

John Poggio, Douglas R. Glasnapp, Xiangdong Yang, and Andrew J. Poggio

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Design and Layout: Thomas Hoffmann

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.



**Preferred citation:**

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). Available from http://www.jtla.org

**Note:**

Portions of this paper were presented at the annual meeting of the National Council on Measurement in Education, San Diego, California, April 2004.


Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). Available from http://www.jtla.org

**Abstract:**

The present study reports results from a quasi-controlled empirical investigation addressing the impact on student test scores when using fixed form computer based testing (CBT) versus paper and pencil (P&P) testing as the delivery mode to assess student mathematics achievement in a state's large scale assessment program. Grade 7 students served as the target population. On a voluntary basis, participation resulted in 644 students being "double" tested: once with a randomly assigned CBT test form, and once with another randomly assigned and equated P&P test form. Both the equivalency of total test scores across different student groupings and the differential impact on individual items were examined.

Descriptively there was very little difference in performance between the CBT and P&P scores obtained (less than 1 percentage point). Results make very clear that there existed no meaningful statistical differences in the composite test scores attained by the same students on a computerized fixed form assessment and an equated form of that assessment when taken in a traditional paper and pencil format. While a few items (9 of 204) were found to behave differently based on mode, close review and inspection of these items were not able to identified factors accounting for the differences.

J·T·L·A

# A Comparative Evaluation of Score Results from Computerized and Paper and Pencil Mathematics Testing in a Large Scale State Assessment Program

John Poggio, Douglas R. Glasnapp, & Xiangdong Yang
University of Kansas

Andrew J. Poggio, University of Iowa

This article reports on what is among the first quasi-controlled systematic empirical investigations addressing the impact on student scores when using fixed form computerized testing (CBT) versus paper and pencil (P&P) to assess student mathematics achievement in a large-scale state assessment program.

The past decade in educational measurement has been a time of considerable and eventful change. Consider the following: resolution and better understanding of standard setting, ready reliance on standards-based criterion referenced assessment, wide acceptance of Item Response Theory as a basis for developing assessments and use of test scores, and generalizability theory as a means to better analyze test component characteristics. Further, assessment has emerged as an important tool for educational reform and change with the No Child Left Behind (NCLB) Act placing assessment as a cornerstone toward gauging school success and, most recently, the emergence in K–12 settings of formal, standardized computerized testing. The first collection of major initiatives has yielded much information, and the jury on the impact of NCLB will be silent for a few years, but it is with respect to computerized based testing that we are in dire need of systematic study and investigation to inform decisions and direction at this time.

In the years ahead, there will be more movement toward and acceptance of computer-based testing (CBT) as the dominant approach to school testing. The presence of technology alone in schools will cause this to occur.

Several states (Kansas, Indiana, North Carolina, and Virginia) are in the process of putting in place major initiatives to move in this direction and others are soon to follow, especially with NAEP conducting field trails

of its CBT design and application. With respect to computer-based testing and large scale state assessment, much has been learned from these major initiatives as well as from pioneering studies conducted to measure the effect of administration mode. The reason for the move to computer-based delivery is clear as there is little doubt that the CBT modality offers advantages. CBTs have become desirable because of immediate score reporting on student performance, the reduction in cost related to printing, shipping, and administering paper and pencil (P&P) assessments, several test security improvements, as well as the continuous testing of students (Parshall, Spray, Kalohn, & Davey, 2002; Wise & Plake, 1990). Further, new, innovative item formats can be utilized in assessments through the use of technology (Jodoin, 2003; Parshall et al., 2002; Huff & Sireci, 2001). However, many challenges exist, including the question of whether the scores produced from tests administered in both the CBT and P&P formats are comparable (Wang & Kolen, 2001; Gallagher, Bridgeman, & Cahalan, 2002). Thus, there is a clear need of systematic study and exhaustive investigation to inform decisions and direction at this time. When a CBT system is implemented, it is paramount that examinee responses are affected only by test content, not administration mode.

A growing body of research exists examining the interchangeability of scores obtained from CBTs and traditional P&P tests. Mazzeo and Harvey (1988) conducted an early literature review and found mixed results. While some studies provide evidence of score equivalence across the two modes, computerized assessments tended to be more difficult than P&P versions of the same test. Pommerich (2004) concludes that the more difficult it is to present a P&P test on a computer, the greater the likelihood of mode effects to occur. Hetter, Segall, and Bloxom (1997) found that administration mode effects are typically small when a CBT is a literal transfer of the fixed number of items from a P&P test to a computer screen in a static manner. That is, mode effects are typically not found for tests where items are presented in their entirety on a single computer screen (Bergstrom, 1992; Spray, Ackerman, Reckase, & Carlson, 1989). For tests where items cannot be fit onto the screen in their entirety without scrolling such as those with reading passages, more significant mode effects have been found (Bridgeman, Lennon, & Jackenthal, 2003). Mead and Drasgow (1993) concluded that scores being measured across the two administration modes are similar for untimed (power) tests but not for speeded tests. Other research has found that CBT and P&P versions of tests yield similar scores (Wise, Barnes, Harvey, & Plake, 1989; Taylor, Jamieson, Eignor, & Kirsch, 1998; Puhan & Boughton, 2004). The previous literature in this area seems to indicate that mode differences typically result from the extent to which the presentation of the test and the process of taking the test differ across modes, and not to differences in content (Pommerich, 2004). CBTs

should be constructed to minimize differences between modes, but comparability cannot simply be assumed as evidenced by the somewhat inconsistent findings in the literature evaluating the comparability of CBTs and P&P tests. Pommerich (2004) offers that it is therefore important for testing programs to conduct comparability studies based on their own tests and technology since findings from previous studies cannot be generalized to similar situations.

Technology has become commonplace in schools and assessments in recent years as evidenced by the literature. In all these efforts we have yet to acquire a sense of the consequence of these significant changes. There is little doubt that the CBT modality will offer advantages. But what of the transition between CBT and the now P&P in-place assessments? Is it possible, defensible and necessary for states and schools to operate "dual" programs based on the untested assumption that there may be differences in performance based on the mode of testing; do CBT advantages always outweigh what may be its deficiencies? Does CBT equally meet the needs of all students, or are some advantaged while others disadvantaged by the methodology, and if differences are found, how might their performance using a P&P approach been impacted? These are a few of the pending and crucial questions. The investigation reported in this paper was designed and implemented to address one of these fundamental needs by carrying out an experimental study of the affect of CBT on test scores in comparison to P&P testing results in the live context of the state of Kansas' large scale assessment program. Information on this central issue is crucial toward advising policy and guiding practice: must dual programs involving both CBT and P&P be operational in a state assessment system to assure equity and fairness, or can change be enacted gradually across schools that are ready to implement CBT or must all change be postponed until all schools are ready to move to the CBT mode for assessment.

## Methods and Procedures

For the spring 2003 federally approved administration of the Kansas large scale state assessment program, all necessary software applications were developed to provide, on a voluntary basis, opportunities for the schools to participate in the implementation of the state's mathematics assessment at grade 7 following a fixed-form computerized testing approach.

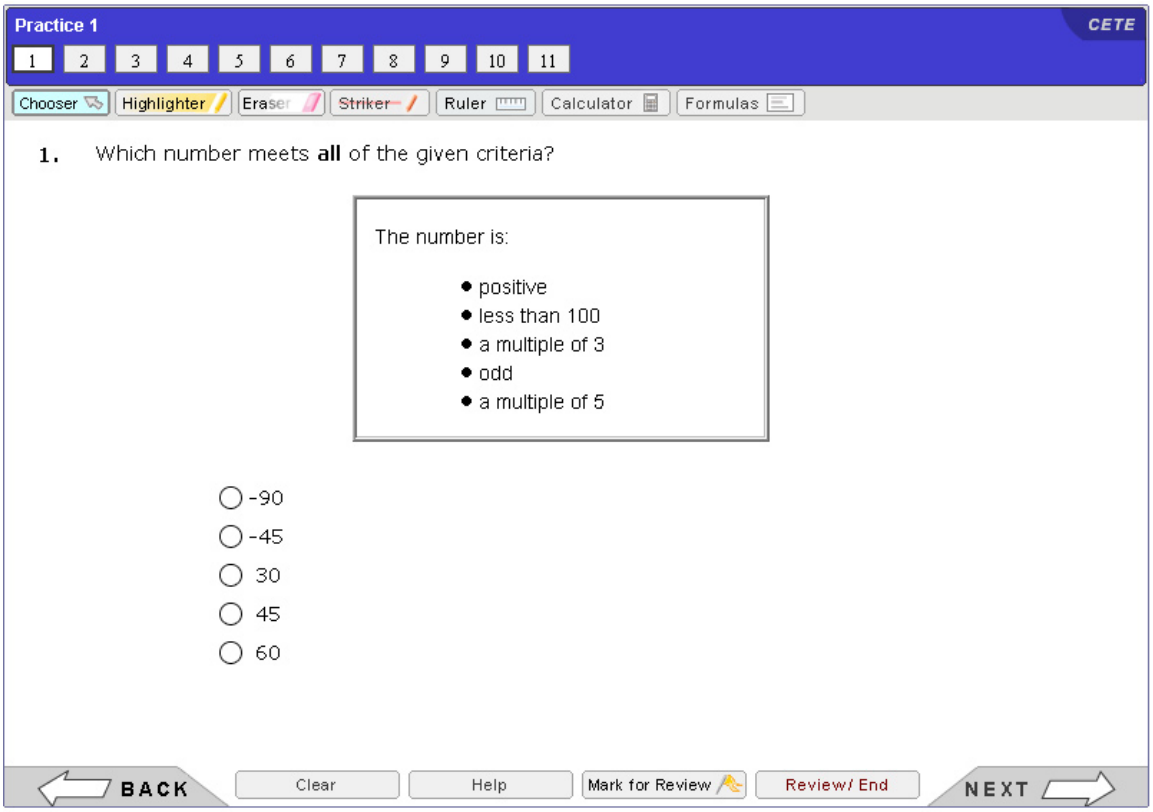



Figure 1. The mathematics test template for the Kansas Computerized Assessment

The image above provides a representation of the Kansas Computerized Assessment (KCA) software application for mathematics assessment. This computerized assessment system is referred to as the "CBT" throughout the remainder of this paper. The image illustrates the essential Kansas CBT template for testing. Inspection shows that one item is presented at a time in the center of the basic template, then supporting features "surround" the item which students use to navigate and respond, mark, etc. an item. The primary application includes features for navigation to specific items, the ability to use tools such as formulas, a calculator, ruler, highlighter, a response-choice striker, and marking an item for later review. This is the template which "presents" all items in this CBT. The four (4) active P&P forms of the grade 7 mathematics multiple choice assessment were made available on the computer platform with one of the four forms randomly assigned a student at the time of testing. The CBT was delivered via the internet in real time and provided complete security features (local registrations, test session tickets, redundant backup systems, load managing software, scheduling, reactivation options, etc.) to assure a proper and standardized offering of this assessment. During actual implementation,

no events occurred or were reported to suggest that problems or issues compromised the score data captured. Backup protocols and quality control procedures confirmed the accuracy and quality of all data captured via CBT administration. To see and learn more about the software application and Kansas CBT procedures, go to: www.kca.cete.us to access links and downloads demonstrating the Kansas Computerized Assessments system. At this site, readers can obtain student training tutorials, practice tests, and directions for local implementation of this CBT program can be reviewed and studied.

Approximately 32,000 Kansas students were eligible and required to sit for the state's grade 7 mathematics assessment. On a voluntary basis, 48 schools agreed to test some or all of their grade 7 students using CBT and in the final sample, 2861 students sat for the CBT. This was the first opportunity for schools to be involved in the state assessments online, and only two of the participating 48 schools reported having previously done any online, formal computerized testing.

The Kansas CBT was supported by an instructional tutorial (including audio) in English and Spanish. All staff and students were required to sit though the tutorial at least once. In addition to the tutorial, two practice tests are available to give students real experiences using the software testing application. Field surveys and observations documented that training on the tool through use of the tutorial and practice testing did occur for all student participants. Solicited self-report survey information confirmed that students were comfortable and largely at ease within the testing CBT environment (Glasnapp, Poggio, Poggio, & Yang, 2005).

Implementation of the study design occurred in the following manner. Four equated parallel forms of the 7$^{th}$ grade test existed, each resulting in three equated percent correct scores, a score based on knowledge items, a score based on application items and a Total score. These forms were constructed to be parallel using common test specifications to assure consistent representative content objective coverage, and all forms were equated based on analyses using linear and IRT methods following a random/equivalent groups design. Standard errors of the equating algorithms were less than 1.1 score units. Total score reliabilities for all forms exceed .92. The CBT randomly administered one of the four forms to a student. Shortly after a school volunteered to participate in the CBT testing, the school was asked if they also would be willing to have their students sit for and complete a second form of the grade 7 mathematics test administered in the P&P format, but the form administered would be randomly assigned from one of the three remaining parallel and equated forms not taken by the student online.

Twelve schools agreed to the additional testing of their students with a parallel P&P form. This level of participation resulted in 646 students being "double" tested: once with a randomly assigned CBT test form, and once with another randomly assigned and equated P&P test form. In some cases, students did not have valid total test scores on one of the assessments and were included or excluded based on the analysis conducted. Because of the voluntary participation, implementing a strictly randomized counterbalanced design was not possible. Rather, schools selected, based on the convenience of their schedules, which test format was given first. Only three schools tested first with the P&P form ($n$ = 102). As unexpected events can occur when procedures such as these are put in place, it turned out that two schools who tested first with the CBT went out of their way to assure that students received the same identical test forms under both modes of testing ($n$ = 102). Though not the design intended, these latter data allowed for the evaluation of the impact of repeating the same test under both conditions, albeit though not controlling for the order effect if one were to exist. To summarize, the design as implemented for the investigation allowed for the study of four groups:

1. Two parallel and equated forms administered under CBT and P&P modes ($n$ = 515)
2. Same form taken under both modes: CBT and P&P ($n$ = 102)
3. Administration of the P&P first: ($n$ = 57)
4. Administration of the CBT first: ($n$ = 480)

# Results

## Descriptive Statistics

The tables that follow report findings showing performance on the major components of the state's grade 7 mathematics assessment: the equated percent correct Total Score (52 items), the equated percent correct Knowledge Information Score (26 items), and the equated percent correct Application/Problem Solving Score (26 items). Tables 1 through 4 present descriptive statistical results (means and standard deviations) based on performance under the different studied conditions (CBT versus P&P as well as the order effect). A review of these data demonstrates that descriptively there is very little difference in performance between the scores obtained (less than 1 percentage point), whether the assessment was taken as a computerized administration or in a traditional paper and pencil mode. In addition, Table 1 provides normative data to demonstrate that students participating in the double testing were not an aberrant sample. Their performance on both CBT and P&P were very comparable to the mean performances of all students in the state taking P&P ($n$=32,518) and those students taking only CBT ($n$=2244).

**Table 1: Percent correct means for students participating in both the CBT and P&P administrations, all students in the state, and students only in the CBT option for state assessment**

| | **Students in BOTH CBT and P&P administrations** *n*=617 | | **All students in the State** *n*=32,518 | | **Student ONLY taking the CBT** *n*=2,244 | |
|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Mean | Std. Deviation | Mean | Std. Deviation |
| Total CBT Score | 53.95 | 16.9 | – | – | 55.84 | 17.2 |
| Total P&P Score | 54.21 | 17.2 | 54.9 | 18.1 | – | – |
| Knowledge CBT Score | 54.05 | 18.8 | – | – | 56.93 | 18.8 |
| Knowledge P&P Score | 54.90 | 18.8 | 55.9 | 19.5 | – | – |
| Application CBT Score | 53.82 | 17.3 | – | – | 54.59 | 18.0 |
| Application P&P Score | 53.49 | 18.0 | 53.8 | 18.9 | – | – |

**Table 2: Descriptive Statistics for students taking both the CBT and a DIFFERENT form of the P&P, and those taking both the CBT but the SAME form of the P&P**

| | **Students taking DIFFERENT forms of the CBT and the P&P** *n*=515 | | **Students taking the SAME form of the CBT and the P&P** *n*=102 | |
|---|---|---|---|---|
| | Mean | Std. Deviation | Mean | Std. Deviation |
| Total CBT Score | 54.21 | 17.1 | 52.65 | 16.2 |
| Total P&P Score | 54.21 | 17.2 | 54.21 | 17.9 |
| Knowledge CBT Score | 54.23 | 19.1 | 53.14 | 17.5 |
| Knowledge P&P Score | 54.63 | 18.9 | 56.25 | 18.6 |
| Application CBT Score | 54.12 | 17.4 | 52.29 | 17.0 |
| Application P&P Score | 53.74 | 17.7 | 52.23 | 19.3 |

**Table 3: Descriptive Statistics based on test order irrespective of Mode**

| | **Total** | | **Knowledge** | | **Application** | |
|---|---|---|---|---|---|---|
| | First | Second | First | Second | First | Second |
| **Mean** | 52.75 | 53.15 | 52.73 | 53.63 | 52.79 | 52.64 |
| **SD** | 18.05 | 17.95 | 19.71 | 19.76 | 18.45 | 18.23 |
| **N** | 538 | 539 | 538 | 539 | 538 | 539 |

**Table 4: Descriptive Statistics for students whose FIRST test was a P&P Form, and those where the FIRST test was a CBT Form**

| | First Assessment was the P&P *n*=57 | | First Assessment was the CBT *n*=480 | |
|---|---|---|---|---|
| | Mean | Std. Deviation | Mean | Std. Deviation |
| Total CBT Score | 52.65 | 16.2 | 52.09 | 18.4 |
| Total P&P Score | 53.21 | 17.9 | 52.27 | 18.3 |
| Knowledge CBT Score | 53.14 | 17.5 | 51.88 | 20.1 |
| Knowledge P&P Score | 53.25 | 18.6 | 52.51 | 20.1 |
| Application CBT Score | 52.29 | 17.0 | 52.31 | 18.7 |
| Application P&P Score | 52.23 | 19.3 | 52.04 | 18.5 |

Tables 5 through 7 present descriptive information on the differential performance on the CBT versus P&P in relation to gender, SES (using Free/Reduced lunch status as the surrogate), and academic placement category (regular education, gifted education, or special education-learning disabled). These latter tables configure and report results only for variables for which there were reasonable sample sizes (i.e., only LD and gifted student categories are reported as there were some data for these groups, but not for other groups such as mobility, etc.).

With respect to demographic/classifications, no gender differences are observed, performance based on academic classification shows main effect separations as would be expected. The same pattern occurred for the SES breakdown. However, comparison between scores attained under the CBT and P&P tested modes do not evidence any meaningful differences. It should be noted that the correlation between Total Scores attained by students taking both the CBT and the P&P was .96. Thus, not only were there small to non-existent differences in group averages, but student also maintained their rank position regardless of the testing mode.

**Table 5: Performance Associated With Gender**

| | | Total | | Knowledge | | Application | |
|---|---|---|---|---|---|---|---|
| | | CBT | P&P | CBT | P&P | CBT | P&P |
| **Female** | Mean | 53.45 | 53.64 | 53.93 | 54.14 | 52.90 | 53.09 |
| | SD | 16.81 | 17.09 | 19.26 | 18.71 | 16.90 | 17.53 |
| | N | 263 | 264 | 263 | 264 | 263 | 264 |
| **Male** | Mean | 52.33 | 52.44 | 52.04 | 52.69 | 52.65 | 52.24 |
| | SD | 19.07 | 18.87 | 20.23 | 20.64 | 19.61 | 19.14 |
| | N | 275 | 275 | 275 | 275 | 275 | 275 |

## Table 6: Performance Associated With Academic Placement

| | | Total | | Knowledge | | Application | |
|---|---|---|---|---|---|---|---|
| | | CBT | P&P | CBT | P&P | CBT | P&P |
| **Gen. Ed.** | Mean | 53.98 | 54.29 | 53.95 | 54.57 | 53.95 | 53.92 |
| | SD | 17.03 | 16.88 | 19.15 | 18.78 | 17.20 | 17.22 |
| | N | 474 | 476 | 474 | 476 | 474 | 476 |
| **Gifted** | Mean | 81.92 | 81.31 | 80.85 | 81.85 | 83.00 | 80.92 |
| | SD | 8.36 | 7.96 | 11.36 | 10.46 | 8.54 | 7.11 |
| | N | 13 | 13 | 13 | 13 | 13 | 13 |
| **SPED-LD** | Mean | 34.53 | 33.06 | 35.81 | 33.74 | 33.66 | 33.23 |
| | SD | 13.17 | 13.35 | 15.15 | 15.49 | 12.57 | 14.47 |
| | N | 32 | 31 | 32 | 31 | 32 | 31 |

## Table 7: Performance Associated With SES

| | | Total | | Knowledge | | Application | |
|---|---|---|---|---|---|---|---|
| | | CBT | P&P | CBT | P&P | CBT | P&P |
| **No Lunch Support** | Mean | 55.13 | 55.62 | 55.24 | 56.14 | 54.90 | 55.01 |
| | SD | 17.77 | 17.33 | 19.55 | 19.27 | 18.24 | 17.58 |
| | N | 377 | 378 | 377 | 378 | 377 | 378 |
| **Free** | Mean | 46.01 | 45.82 | 45.97 | 45.27 | 46.36 | 46.72 |
| | SD | 18.06 | 18.85 | 19.55 | 19.81 | 18.38 | 19.79 |
| | N | 100 | 100 | 100 | 100 | 100 | 100 |
| **Reduced** | Mean | 50.23 | 48.75 | 50.34 | 49.77 | 50.15 | 47.80 |
| | SD | 16.22 | 16.97 | 19.28 | 18.26 | 16.03 | 17.77 |
| | N | 61 | 61 | 61 | 61 | 61 | 61 |

To further examine the differences between students' performances on CBT versus P&P, effect sizes were calculated across various conditions and are reported in Table 8. A positive value in Table 8 indicates a higher mean on P&P. Although scores on CBT and P&P are correlated, effect sizes were calculated as the standardized mean differences between the two scores using the pooled standard deviation *without* taking into account the correlation between them, as convincely argued by Dunlap et al (1996). As expected, the effect sizes between performances on CBT and P&P are generally small. The largest effect size is found between performances on CBT and P&P on the application subscale for the students that are gifted (-.265), which still represents about 80% overlap between the two distributions (Cohen, 1988).

**Table 8: Effect Sizes Between Students' Performances on CBT Versus P&P**

| | | Total | Knowledge | Application |
|---|---|---|---|---|
| **Overall** | | 0.015 | 0.064 | -0.019 |
| **Order** | CBT first | 0.010 | 0.044 | -0.015 |
| | P&P first | 0.033 | 0.006 | -0.003 |
| **Gender** | Female | 0.011 | 0.011 | 0.002 |
| | Male | 0.006 | 0.032 | -0.021 |
| **Academic Placement** | SPED | -0.111 | -0.135 | -0.032 |
| | Gen. Ed. | 0.018 | 0.033 | -0.002 |
| | Gifted | -0.075 | 0.102 | -0.265 |
| **SES** | No lunch | 0.028 | 0.046 | 0.006 |
| | Reduced | -0.089 | -0.030 | -0.139 |
| | Free | -0.010 | -0.036 | 0.018 |

# Statistical Inferential Analysis

Data in this study have a hierarchical structure at three different levels: within-subject, between-subject, and between-school district (USD). At the first level (within-subject), each subject received two tests (CBT vs. P&P) in a particular order (CBT first or P&P first). Therefore, there are two variables at this level: Test Mode (CBT vs. P&P) and Test Order (CBT first vs. P&P first). There are three variables at the second level (between-subject): Gender (male vs. female), SES (no lunch support, reduced vs. free) and SPED (regular education, gifted, or SPED-LD student). The only variable at third level is USD (14 district attendance centers).

To more precisely study the impact of mode of testing across group classification, intraclass coefficients were computed. Table 9 presents the intraclass coefficients for the second and third level variables. Also studied in this analysis are the school district effects, that is, are participants in the study systematically different based on their school district attendance center? The intraclass coefficients reported are partial coefficients having controlled for the other student characteristics in the analysis. Results of this analysis reinforces the observed result: no meaningful main effect in mean differences between grouping conditions based on gender or SES, but large observed differences as one would anticipate across groups for the academic placement factor (SPED) and attendance center.

**Table 9: Intraclass Correlations for Different Student Characteristics**

| | Total Score | Intraclass Correlation | Knowledge | Intraclass Correlation | Application | Intraclass Correlation |
|---|---|---|---|---|---|---|
| Total Variance | 318.086 | | 386.293 | | 327.601 | |
| Gender | 0.586 | 0.002 | 0.271 | 0.001 | 1.030 | 0.003 |
| SES | 2.965 | 0.009 | 3.388 | 0.009 | 2.333 | 0.007 |
| USD | 31.249 | 0.098 | 46.591 | 0.121 | 19.307 | 0.059 |
| SPED | 311.128 | 0.978 | 294.069 | 0.761 | 324.898 | 0.992 |

A statistical analysis was conducted using a hierarchical linear model formulation. The first model fit to the data is given as:

**(1)**

$$Y_{jik} = \beta_{0ik} + \beta_1 mode_{jik} + \beta_2 order_{jik} + \varepsilon_{jik},$$
$$\beta_{0ik} = \gamma_{00k} + \gamma_{01} gender_{ik} + \gamma_{02} ses_{ik} + \gamma_{03} sped_{ik} + \mu_{0ik},$$
$$\gamma_{00k} = \lambda_{000} + \tau_{00k}$$
$$j = 1, 2;\ i = 1, 2, ....., n_k;\ k = 1, 2, ....14.$$

In this model, $Y_{jik}$ is the test score for the $j^{th}$ measurement of subject $i$, which is nested within the USD $k$ with sample size $n_k$. The effects of Mode and Order are treated as fixed. Because Mode and Order are not crossed over within a subject, the interaction between Mode and Order is confounded with the between-subject effects. No interaction term between Mode and Order is included in the model.

The intercept at the first level is treated as random and further regressed on the set of subject-level variables, i.e. Gender, SES and SPED, whose effects are treated as fixed as well. There are 14 USDs that are in the current study, which are treated as a random sample from the state. Therefore, the effect of USD is treated as random. The resulting model was fitted to the data using SAS PROC MIXED. The corresponding results are shown in Table 10. It can be seen from Table 9 that main effects for Test Order and Test Mode are not significant. Both SES and SPED significantly change the subject's test score ($F_{1,516}$ = 10.78, P =.0015; $F_{2,516}$ = 43.82, P =.0000). The estimated variances and their standard errors for random effects are also presented in Table 10. Each of the three random effects shows substantial variation.

**Table 10: Parameter Estimation from the Model Without Interaction**

| | | | Estimate | SE | P |
|---|---|---|---|---|---|
| Fixed Effects | Intercept | | 40.825 | 3.3269 | |
| | Gender | female | -1.156 | 1.3011 | 0.23 |
| | | male | 0 | – | – |
| | SES | | -2.845 | 0.866 | 0.011 |
| | SPED | regular | 16.7099 | 2.7738 | 0.0001 |
| | | gifted | 45.3235 | 4.8827 | 0.0000 |
| | | disability | 0 | – | – |
| | Order | first | 5.8024 | 5.0909 | 0.2901 |
| | | second | 0 | – | – |
| | Mode | CBT | 6.7903 | 5.0909 | 0.5077 |
| | | P&P | 0 | – | – |
| Random Effects | Intercept | | 186.84 | 13.322 | |
| | USD | | 38.963 | 18.756 | |
| | Residual | | 46.801 | 2.914 | |

## A Model With Cross-level Interactions

In fitting the multilevel model without cross-level interactions, the effects of test mode and test order are hypothesized to be constant across different levels of the subject level variables and across different USDs. To test this hypothesis, the coefficients of Mode and Order were set to be random at either person-level or USD-level. The fitted models, however, lead to a non-positive definite Hessian matrix. Therefore, instead of setting the coefficients to be random, only fixed interaction effects between subject-level variables and Mode or Order were specified in the model. Table 11 gives the statistical tests for the cross-level interactions. None of the interaction terms are statistically significant.

**Table 11: Significant Tests of Cross-Level Interactions**

| Effect | Numerator DF | Denominator DF | F Value | P |
|---|---|---|---|---|
| Order*SES | 1 | 509 | 0.41 | 0.5212 |
| Mode*SES | 1 | 509 | 0.01 | 0.9345 |
| Order*SPED | 2 | 509 | 0.91 | 0.3397 |
| Mode*SPED | 2 | 509 | 1.54 | 0.2147 |
| Order*Gender | 1 | 509 | 0.00 | 0.9766 |
| Mode*Gender | 1 | 509 | 0.01 | 0.9344 |

# Evaluation Based on IRT Ability Analyses

Test total scores as reported in the preceding tables and the analyses presented only take into account the number of items being answered correctly, not the properties of the set of items being answered. On the other hand, IRT ability estimates incorporate such information if a 2PL or 3PL model is applied. Using BILOG-MG, three sets of student ability estimates based on a 3PL model were obtained. We selected the 3PL model as these are challenging mathematics tests which do result in considerable variation in performance, and guessing is observed. By pooling all of the items across test modes, CBT and P&P, taken by a student, a total ability estimate was obtained. Ability estimates based on items from only CBT or P&P were also calculated for each student, respectively. The ability estimates were then transformed to a common scale using either CBT or P&P items as the common items. Table 12 gives the descriptive statistics for ability distributions across different test modes after the rescaling. As shown in the histograms (Figure 2), all of the ability distributions obtained from different test modes are approximately normal and largely equivalent. These results confirm the results based on observed scores, i.e., little or no meaningful (or statistical) difference in performance based on mode. The correlation between CBT and P&P ability (theta) estimates was .95.

**Table 12: Descriptive Statistics for Ability Distributions Across Different Test Modes**

| | 3PL | SD | Variance | N |
|---|---|---|---|---|
| Total | 0.0023 | 1.0531 | 1.11091 | 646 |
| CBT | 0.566 | 1.0909 | 1.1902 | 646 |
| P&P | 0.0175 | 0.9688 | 0.9386 | 646 |

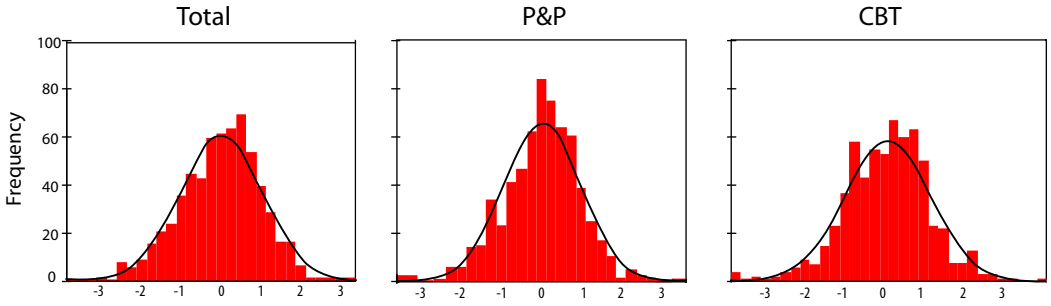



Figure 2. Distributions of ability estimates from different test modes

Inferential analyses bases on ability estimates were carried out using the model specified in equation (1). The observed test scores were replaced by the estimated ability for each student through a 3PL IRT model. The results are presented in Table 13. The identical conclusion as when using observed equated test scores was obtained. That is, main effects for Test Order and Test Mode are not statistically significant. Both SES and SPED significantly change the subject's test score ($F_{1,518}$ = 11.79, P =.0006; $F_{2,516}$ = 47.76, P =.0000).

**Table 13: Fixed and Random Effects of the Model With IRT Ability as Dependent Variable**

| | | | Estimate | SE | P |
|---|---|---|---|---|---|
| Fixed Effects | Intercept | | -0.8804 | 0.2009 | |
| | Gender | female | -0.0579 | 0.0786 | 0.46 |
| | | male | 0 | – | – |
| | SES | | -0.1984 | 0.0577 | 0.006 |
| | SPED | regular | 1.1503 | 0.1675 | 0.0000 |
| | | gifted | 2.8231 | 0.2953 | 0.0000 |
| | | disability | 0 | – | – |
| | Order | first | 0.0582 | 0.0487 | 0.2321 |
| | | second | 0 | – | – |
| | Mode | CBT | 0.0594 | 0.0487 | 0.2231 |
| | | P&P | 0 | – | – |
| Random Effects | Intercept | | 0.6488 | 0.0488 | – |
| | USD | | 0.1461 | 0.0694 | – |
| | Residual | | 0.2367 | 0.0147 | – |

Interactions across the first and second level variables were also tested and the corresponding results were shown in Table 14. None of these interactions were found to be statistically significant ($p$ > .01).

## Table 14: Tests of Cross-Level Interactions When IRT Score as Outcome Measure

| Effect | Numerator DF | Denominator DF | F Value | P |
|---|---|---|---|---|
| Order*SES | 1 | 511 | 0.00 | 0.9889 |
| Mode*SES | 1 | 511 | 0.15 | 0.7033 |
| Order*SPED | 2 | 511 | 3.53 | 0.0300 |
| Mode*SPED | 2 | 511 | 3.54 | 0.0297 |
| Order*Gender | 1 | 511 | 0.11 | 0.7366 |
| Mode*Gender | 1 | 511 | 1.13 | 0.2889 |

To examine the accuracy of measurement (reliability) between the two different test modes, test information was calculated for each comparative pair (P&P vs. CBT) of the four test forms. More information on the CBT than the P&P was observed consistently in the range from the middle to the high end of the ability distribution. A comparable amount of information between the two modes was observed at the low end of the achievement distribution. In Figure 3 which follows we report the Test Information functions for the same form administered via CBT and the P&P modes. The test form chosen for display is that one that evidenced the greatest separation between modes. It can be seen that the CBT test form provides more information across the upper 75 percent of the ability distribution being evaluated by the test. As the standard error of the score estimates, i.e., the degree of measurement precision at an given ability level, is an inverse function of the amount of test information that the test provides at a particular score point, the conditional standard errors are also shown in Figure 3. In effect the same set of items when offered via the CBT mode resulted in more accurate student estimates of scores than the P&P presented test. While the differences in modes (CBT versus P&P) appear large, the differences are in actuality not substantial.

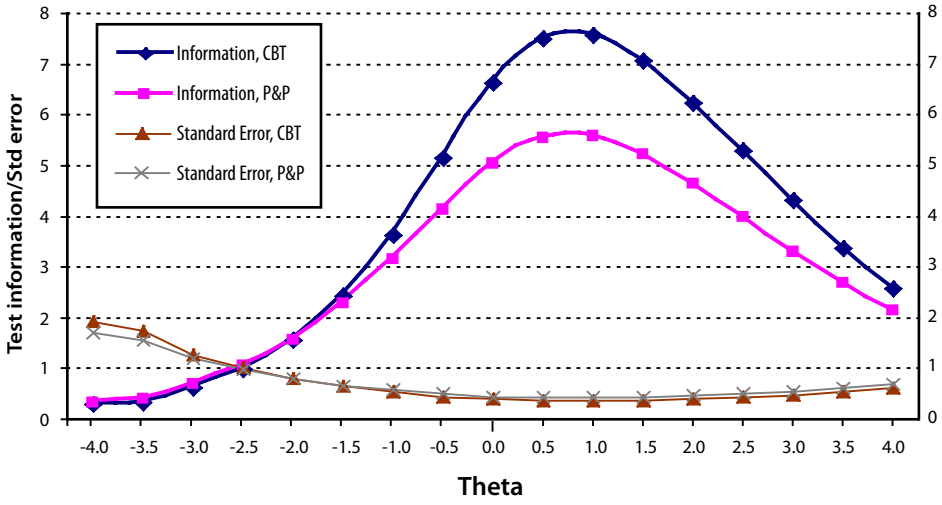



Figure 3. Test Information for Test Form 66: CBT vs. P&P

## Item Level Analysis: DIF over Test Modes

All the previous results reported have addressed *Total Test* performance. A critically important evaluation would be to inspect the conditions of the individual test items between the tested modalities, CBT versus P&P. In this section analyses studying impact at the item level are reported. Specifically, the analyses were conducted to address the following question: Do items perform differently given the mode of administration? This central empirical question was addressed relying on DIF procedures.

Using BILOG-MG, DIF analyses were carried out for each of the four test forms to detect if the same item functions differently across different test modes at the same ability level. All of the DIF analyses used the groups of subject who took the CBT as the reference group. All of the subjects were used in the DIF analysis. The specific number of students taking the different form item combinations under the CBT and P&P modes was presented in Table 15. There were four test forms used in this study. Each of the fours form was randomly assigned to students within both test modes (CBT versus P&P).

**Table 15: Number of Students Across Test Modes and Forms**

| | | P&P | | | | |
|---|---|---|---|---|---|---|
| | | A | B | C | D | **Total** |
| **CBT** | A | 26 | 45 | 42 | 46 | 159 |
| | B | 48 | 28 | 44 | 42 | 162 |
| | C | 40 | 47 | 28 | 48 | 163 |
| | D | 46 | 46 | 45 | 25 | 162 |
| **Total** | | 160 | 166 | 159 | 161 | 646 |

## Differential Item Functioning

To detect uniform and nonuniform DIF, a general IRT-LR method (Thissen, Steinberg & Wainer, 1993) was applied. One-, two- and three-parameter IRT models were sequentially fit to the data. Model selection was based on the relative magnitude of the differences of -2*log likelihood of the two models to the differences of their numbers of parameters. Across all 4 forms, the 3PL model is the model of choice. Since in the current study, sample sizes in both reference and focal groups are relatively small, a 3PL model with a common guessing parameter was fit to each of the samples and the corresponding fit was found statistically to be stronger than the 3PL model with different guessing parameters across items. Using the 3PL model with a common guessing parameter as the baseline model, DIF analyses were conducted as follows.

Step 1: fit a 3PL model with common guessing parameter to the specific test form with *K* items, obtain the -2*log likelihood, denoted as -2*ln(Lc).

Step 2: choose one item as the study item, say, item 5.

Step 3: recode item 5 into two items, item 5R and item 5F. Code item 5R as answered by the Reference group and not reached by the Focal group, code item 5F the other way around.

Step 4: Re-estimate parameters and obtain -2*log likelihood for the test form with *K*+1 items, denoted as -2*ln(La).

Step 5: compute $\chi^2(m)$ = -2*ln(Lc) –[-2*ln(La)], with m = # of parameters in model A - # of parameters in model C.

Step 6: evaluate the presence of DIF for item 5 using standard hypothesis testing procedure.

Table 16 presents the results from the aforementioned DIF procedure based on an analysis of the 204 tested items across the four test forms under the two comparative conditions: CBT vs. P&P. ONLY "flagged" DIF

items are presented in the table. Both discrimination (**a**) and difficulty parameters (**b**) are shown. A probability of $p < .01$ was used to identify an item as behaving "differentially" between the modes.

**Table 16: DIF Analysis Through Likelihood Ratio Method (Reference Group: CBT)**

| Form | Items | –2*Log Likelihood | Difference | *Prob.* | CBT | | P&P | |
|---|---|---|---|---|---|---|---|---|
| | | | | | *b* | *a* | *b* | *a* |
| **A** | 3pl model | 16631.4 | | | | | | |
| | 25 | 16616.1 | 15.3 | 0.0005 | 0.34 | 0.44 | -0.66 | 1.13 |
| | 28 | 16620.7 | 10.7 | 0.0047 | -0.59 | 0.61 | 0.40 | 1.02 |
| **B** | 3pl model | 17776.8 | | | | | | |
| | 13 | 17761.6 | 15.2 | 0.0005 | 1.81 | 0.92 | 0.41 | 0.92 |
| | 18 | 17764.6 | 12.2 | 0.0022 | -0.59 | 1.28 | -1.23 | 0.42 |
| | 33 | 17765.4 | 11.3 | 0.0034 | -3.98 | 1.10 | -4.01 | 0.55 |
| | 49 | 17767.0 | 9.7 | 0.0076 | 0.46 | 1.40 | -0.28 | 0.82 |
| **C** | 3pl model | 18229.5 | | | | | | |
| | 16 | 18212.1 | 17.4 | 0.0002 | 0.50 | 1.64 | -0.03 | 2.78 |
| **D** | 3pl model | 18054.1 | | | | | | |
| | 33 | 18030.0 | 24.1 | 0.0000 | -1.26 | 1.00 | -2.04 | 1.91 |
| | 47 | 18041.8 | 12.3 | 0.0063 | 1.01 | 2.99 | 1.89 | 2.22 |

Nine of the 204 items were identified as performing differentially between the modes. In most cases, these items were observed to be more difficult in the CBT mode, while discrimination tended to be equally diverse. A careful and detailed study of these items was not able to uncover the factor(s) that might account for differential performance. There is a tendency for those flagged items in the CBT mode to be "large," that is, big on the printed page (taking three-quarters up to the entire 8.5x11 inch page) and which therefore require scrolling by the student to see the entire question on a computer screen. However, it is important to note that there are many such items as these on these tests (items requiring scrolling to see the entire question) and only these few were flagged. For the reader's benefit, the item characteristic curves of two flagged items are presented in Figures 4. We are unable to share publicly the flagged items as the tests remain active and security must be assured.

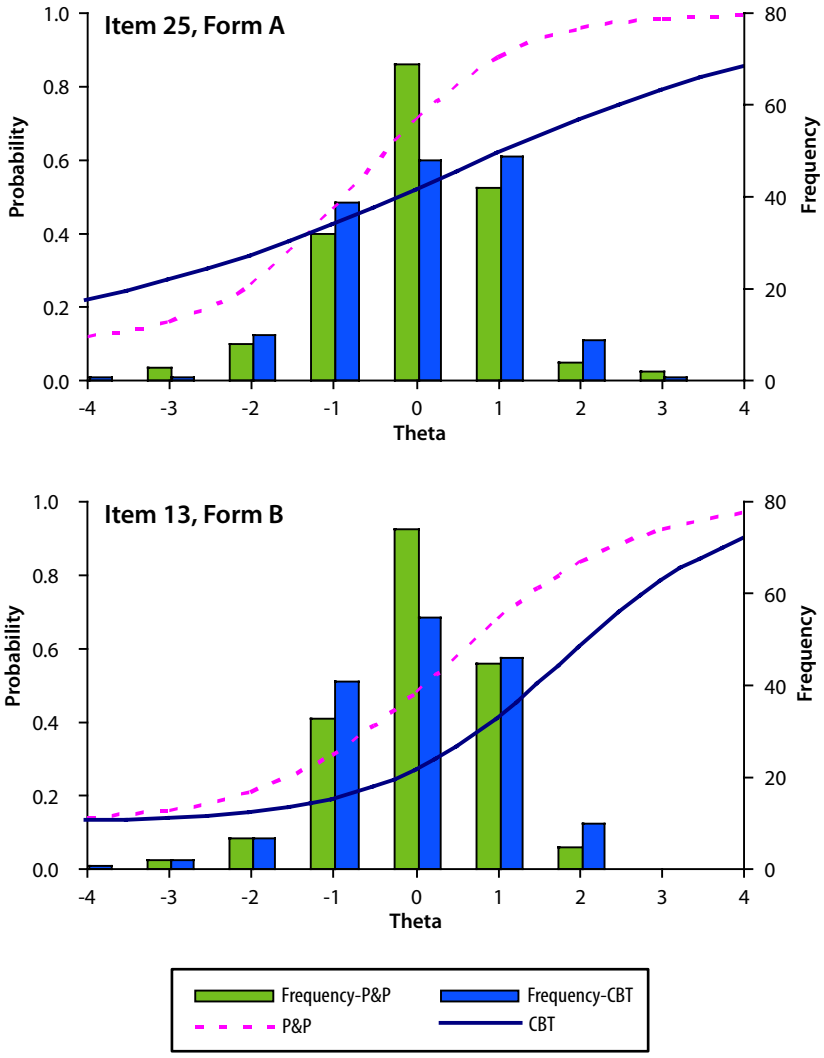



Figure 4. Differential Item Functioning of Item 25 from Form A (upper panel) and Item 13 from Form B (lower panel)

## Differential Item Category (Distractor) Functioning (DICF or DDF)

To further investigate the nature of the presence of DIF for an item across different test modes, differential item category functioning (DICF) or differential distractor functioning (DDF) was conducted for each item that showed DIF. This analysis evaluates whether the response distribution to the item's choices (or distractors) is similar across the modes being examined. That is, for persons of comparable ability, is the likelihood of responses to all of the item choices similar? The DICF was evaluated using a generalized logistic regression procedure. The set of response options for

an item needs to be treated as nominal response categories since there is no order relation among choices. For an item *j* with *k* response categories, the counts at the *k* categories for item *j* can then be assumed to follow a multinomial distribution with probabilities $\{\pi_1, \pi_2, \ldots \pi_k\}$, where $\sum_{k=1}^{K} \pi_k = 1$.

Using the baseline-category Logit model, the odds of selecting response option *k* against the key option across some group *g*, after controlling the effects of ability, θ, can be modeled as:

**(2)**

$$\log\left(\frac{\pi_k(\theta, g)}{\pi_{Key}(\theta, g)}\right) = \alpha_k + \beta_{1k}\theta + \beta_{2k} g, \quad k = 1, 2, \ldots . K - 1,$$

where $\pi_{Key}(\theta, g)$ is the probability of selecting the key option and the key option serves as the baseline category in the DICF case. To fully describe the effect of group or ability, *K–1* logits are needed, as shown in equation 4. The presence of DICF for option *k* was evaluated as follows:

- First, fit a model like equation 4 without the group variable, obtain the corresponding magnitude of the –2*log likelihood, denoted as –2*ln(Lc),
- Second, fit a model like equation 4 including the group variable and record the corresponding –2*log likelihood, denoted as –2*Ln(La),
- Third, calculate $\chi^2(m)$ = –2*ln(Lc) –[–2*ln(La)], with m = *K*–1 for two group comparison.
- Fourth, the presence of DICF for option *k* is evaluated through standard hypothesis testing procedure. Information about which particular category shows the differential effect can be obtained by inspecting the corresponding coefficients $\beta_{2k}$, *k =1,2, ... K–1*.
- Equation 4 detects uniform DICF, the detection of non-uniform DICF can be formulated as:

**(3)**

$$\log\left(\frac{\pi_k(\theta, g)}{\pi_{Key}(\theta, g)}\right) = \alpha_k + \beta_{1k}\theta + \beta_{2k} g + \beta_{3k}\theta * g, \quad k = 1, 2, \ldots . K - 1,$$

using the similar procedure as evaluating uniform DICF.

The DICF analysis results are shown in Table 17. The charts that immediately follow are only intended to illustrate the focus of this analysis: are response distributions across a multiple-choice item's response options comparable. For there to be evidence of "response distribution" differences, it would be the interaction terms given in Table 17 that would have

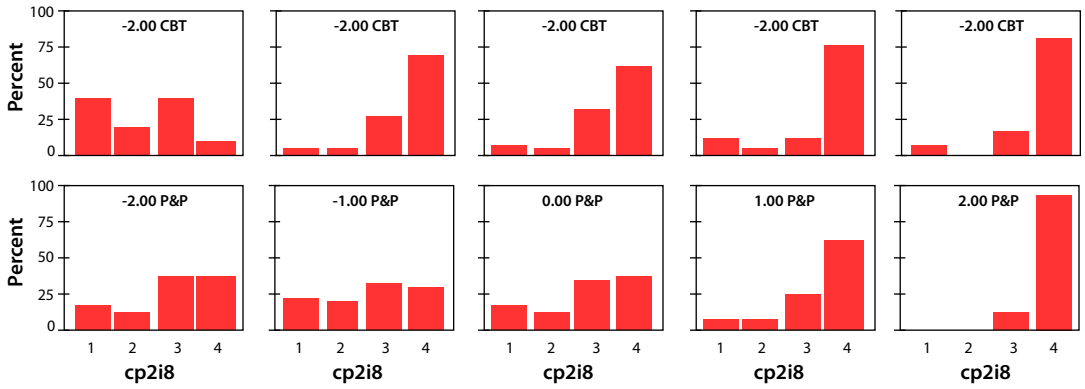
to be statistically significant (the "*int+theta+mode+mode*theta*" term). Based on our DICF analyses, we did not discern that item's "choice distributions" behaved differently between the modes ($p > .01$ for all analyses).

## Table 17: Results of the DICF for Flagged Items

| Form | Item | Model | –2*Log Likelihood | 2 | df | *p* |
|---|---|---|---|---|---|---|
| A | 28 | *int+theta* | 712.7 | | | |
| | | *int+theta+mode* | 700.7 | 12 | 3 | 0.007 |
| | | *int+theta+mode+mode*theta* | 700.5 | 0.2 | 3 | 0.978 |
| B | 33 | *int+theta* | 173.7 | | | |
| | | *int+theta+mode* | 162.1 | 11.6 | 3 | 0.009 |
| | | *int+theta+mode+mode*theta* | 160.6 | 1.5 | 3 | 0.682 |
| | 13 | *int+theta* | 858.7 | | | |
| | | *int+theta+mode* | 840.2 | 18.5 | 4 | 0.001 |
| | | *int+theta+mode+mode*theta* | 838.3 | 1.9 | 4 | 0.754 |
| | 49 | *int+theta* | 777.3 | | | |
| | | *int+theta+mode* | 760.7 | 16.6 | 4 | 0.002 |
| | | *int+theta+mode+mode*theta* | 756.2 | 4.5 | 4 | 0.343 |
| | 43 | *int+theta* | 798.3 | | | |
| | | *int+theta+mode* | 736 | 62.3 | 4 | 0.000 |
| | | *int+theta+mode+mode*theta* | 732.9 | 3.1 | 4 | 0.541 |
| C | 41 | *int+theta* | 412.1 | | | |
| | | *int+theta+mode* | 400.3 | 11.8 | 3 | 0.008 |
| | | *int+theta+mode+mode*theta* | 398.3 | 2 | 3 | 0.572 |
| | 16 | *int+theta* | 738.1 | | | |
| | | *int+theta+mode* | 713.5 | 24.6 | 4 | 0.000 |
| | | *int+theta+mode+mode*theta* | 708.5 | 5 | 4 | 0.287 |
| D | 27 | *int+theta* | 518.1 | | | |
| | | *int+theta+mode* | 505.9 | 12.2 | 3 | 0.007 |
| | | *int+theta+mode+mode*theta* | 502.8 | 3.1 | 3 | 0.376 |
| | 33 | *int+theta* | 287.9 | | | |
| | | *int+theta+mode* | 260.8 | 27.1 | 3 | 0.000 |
| | | *int+theta+mode+mode*theta* | 255.2 | 5.6 | 3 | 0.133 |

The findings from this analysis are illustrated in the images below (Figure 5). For the two items evidencing the greatest DIF among the more than 200 items evaluated, shown are the percent of examines selecting the items' response choices (1, 2, 3, or 4) under the CBT and the P&P modality for examinees at five distinct ability levels (thetas of: -2.0, -1.0, 0.0, 1.0, and 2.0). For example, with reference to student responses to the four (4) choices associated with item 28, approximately 35 percent of the students taking the CBT with thetas of -2.00 chose option 1, about 20 percent option 2, another 35 percent option 3, and about 10 percent selected option 4, etc. CBT response patterns across thetas are then compared using the DCIF model to the response profile for examinees in the P&P test mode. Review of the illustrations reveals that the response mode (CBT or P&P) does not appear to impact the selection or attractiveness of the items' distractors.

**Item 28, Form A** (4 response choices)




**Item 13, Form B** (5 response choices)

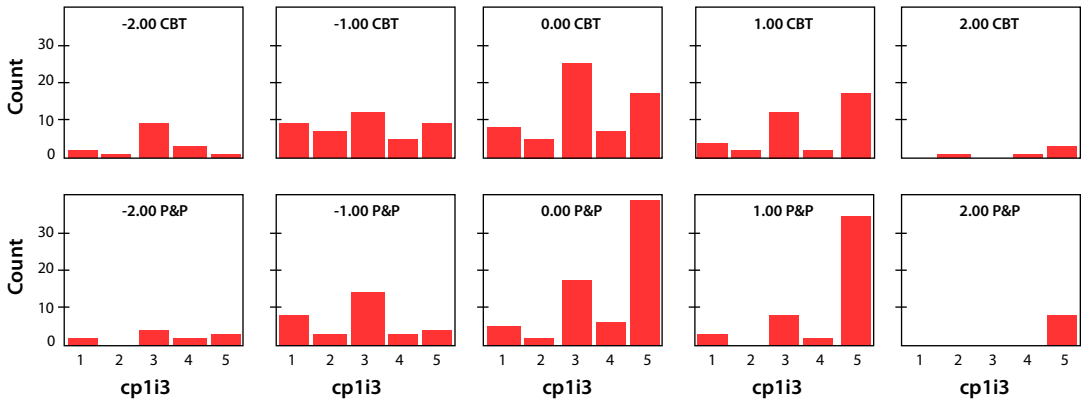



Figure 5. Illustrations of differential item category functioning

# Conclusions

Based on the analyses reported, results make very clear that there existed no meaningful or statistical significant differences in the composite test scores attained by the same students on a computerized fixed form assessment and an equated form of that assessment when taken in a traditional paper and pencil format. Reasonable study controls support the generalizability of these findings, but findings are limited to middle level/grade assessment of mathematics in a general education population. While we observed some items (9 of 204) that did behave differently based on mode and that the difference tended to find these particular items being more difficult in the CBT mode, close review and inspection of these items has not (as yet) identified factors accounting for the differences. Inspection of these items by the investigators suggests that some attention be given to cognitively complex questions that require scrolling in order to see all parts of items' stimuli. We were not able to confirm this finding but it merits further study.

For the few item level differences observed, they had no impact on total or part test scores. This finding supports previous research that suggests that scores obtained from CBTs will be equivalent to those obtained from traditional P&P tests if the CBT is constructed in such a way that it reflects the P&P version on the computer screen. While the needs of this testing program have been established psychometrically, it is important to consider other points of view (Parshall et al., 2002). Issues affecting examinees must be evaluated as well, such as prior experience with computers, proficiency, and examinee comfort, as these factors may act as mediators or moderators in performance across modes. As the CBT in this study was a literal transfer of P&P items to computer, examinee experience with computers was not likely to be an issue for most examinees. A thorough and detailed tutorial for taking the CBT was provided for examinees to familiarize them with taking tests via computerized delivery. Examinee reactions to this mode of administration are summarized in another paper in this volume (see Glasnapp, Poggio, Poggio, & Yang, 2005).

Important to recognize from this investigation is that based on findings for the implementation evaluated, no special provisions appear necessary to offer simultaneously both P&P and CBT forms of assessments. To date there has been justifiable attention toward determining if separately readied and developed assessments coupled with the need for separate cut scores and equating would be required if both CBTs and P&Ps are used in a state's system. Policy makers have expressed concern regarding the costs and management burdens that would be required if operating a "dual" program, i.e., providing both CBTs and P&P to validly measure student learning. Based on these results and analyses to date in the Kansas

environment and context for testing, this investigation advises that dual programs do not appear needed and additional psychometric manipulation and tweaking appears unnecessary. Again, this finding is limited to middle level mathematics, with content coverage and test item structures typical to that in the state studied, relying on a CBT application as the one now used in Kansas. For now, CBT testing appears to provide a credible and comparable option to the P&P testing modality.

# References

Bergstrom, B. A. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*, 191–205.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Dunlap, W. P, Cortina, J. M, Vaslow, J. B, & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 2*, 170–177.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*(2), 133–147.

Glasnapp, D. R., Poggio, J. P., Poggio, A. J., & Yang, X. (2005). Student Attitudes and Perceptions Regarding Computerized Testing and the Relationship to Performance in Large Scale Assessment Programs. *Journal of Technology, Learning, and Assessment*. (in press.)

Hetter, R., Segall, D. O., & Bloxom, B. M. (1997). Evaluating item calibration medium in computerized adaptive testing. In W.A. Sands, B.K. Waters, and J.R. McBride (Eds.) Computerized Adaptive Testing: From Inquiry to Operation. Washington D.C.: American Psychological Association.

Huff, K.L. & Sireci, S.G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25.

Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement, 40*(1), 1–15.

Mazzeo, J. & Harvey, A. L. (1988). *The equivalence of scores from conventional and automated educational and psychological tests: A review of literature* (College Board Report No. 88-8). Princeton, NJ: Educational Testing Service.

Mead, A. D., & Drasgow, F. (1993). Effects of administration medium: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical Considerations in Computer-Based Testing. New York: Springer-Verlag.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6).

Puhan, G. & Boughton, K.A. (2004). Evaluating the Comparability of Paper and Pencil Versus Computerized Versions of a Large-Scale Certification Test. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Spray, J. A., Ackerman, T.A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement*, *26*, 261–271.

Taylor, C., Jamieson, J., Eignor, D. R., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (ETS Research Report No. 98-08). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L, & Howard, W. (1993). Detection of differential item functioning using the parameters of item response models. In. P. W. Holland and H. Wainer, (Eds.). *Differential item functioning*, (pp 67–114). Hillsdale, NJ: Lawrence Earlbaum Associates.

Wang, T. & Kolen, M.J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, *38*, 19–49.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education,* 235–241.

Wise, S.L., & Plake, B.S. (1990). Computer-based testing in higher education. *Measurement and evaluation in counseling and development, 23*(1), 3–10.

# Author Biographies

John Poggio is a Professor in the Department of Educational Psychology and Research and Co-Director of the Center for Educational Testing and Evaluation at the University of Kansas. His research interests are in psychometrics, applied measurement, and educational assessment policy.

Douglas R. Glasnapp is a Research Professor and Co-Director of the Center for Educational Testing and Evaluation at the University of Kansas. Research interests are in applied statistical analysis, and measurement and assessment issues.

Xiangdong Yang is a research associate, Center for Educational Testing and Evaluation, University of Kansas. His research interests are cognitive item design, psychometric models, item response theory, computerized adaptive testing.

Andrew J. Poggio is a doctoral candidate in Educational Measurement and Statistics at the University of Iowa. His research interests are in applications of statistical and measurement methodologies, including computer-based testing, and the use of achievement information for making decisions about students and educational programs.

J·T·L·A

The Journal of Technology, Learning, and Assessment



www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College