Journal of Educational Measurement, Evaluation and Research
Volume 2, Issue 1, January-June 2026

Computer-Based Testing versus Paper-Based Testing in Large-Scale Assessments: A Comparative Evaluation of Mode Effects, Psychometric Equivalence, and Testing Kinetics

Prof. Debashish Roy Chowdhury[1], Ipsita Banerjee[2], Souvik Dasgupta[3], Ananya Ghosh[4]

ABSTRACT

The migration of large-scale academic examinations from traditional paper-based testing (PBT) to computer-based testing (CBT) environments represents a profound shift in modern educational measurement. While digital systems promise rapid administration, lower operational overhead, and opportunities for computerized adaptive testing (CAT), they also introduce potential 'mode effects' that can alter the psychometric properties of standardized test items. This paper presents a multi-site, randomized cross-over experimental design evaluating the comparability of CBT and PBT frameworks across a cohort of 480 secondary school students. Utilizing Classical Test Theory (CTT) and Item Response Theory (IRT) parameter matching, this study analyzes score equivalence, structural item differential function (DIF), test-taking velocity, and cognitive navigation fatigue across identical assessment items. Quantitative findings indicate that overall psychometric equivalence holds across standard verbal and scientific domains; however, a statistically significant mode effect was observed in complex data interpretation and spatial math subsets, with CBT cohorts achieving up to 6.7% higher efficiency ($p < 0.001$) due to active visualization interfaces. Conversely, time-series velocity tracking demonstrated that paper-based cohorts suffered from a 26% higher average navigation latency as testing progressed into later quartiles, indicating notable cognitive fatigue from manual booklet handling. Qualitative telemetry reveals that interface familiarity, screen layout, and digital interactive elements serve as primary drivers of student engagement, whereas systemic device variance and scrolling limits act as primary confounding factors. This research provides a mathematically validated

 Page 27-38

optimization framework for educational administrators and testing psychometrists seeking to deploy equitable, large-scale computer-based assessment models.

KEYWORDS: *Computer-Based Testing (CBT); Paper-Based Testing (PBT); Mode Effects; Item Response Theory (IRT); Differential Item Functioning (DIF); Psychometric Equivalence.*

INTRODUCTION

Over the past two decades, the global educational sector has steadily transitioned its institutional assessment models from traditional paper-based testing (PBT) systems to computer-based testing (CBT) architectures [1]. Driven by rapid advancements in cloud computing, digital infrastructure, and artificial intelligence, this shift has transformed large-scale examinations from manual, bureaucratic exercises into highly agile digital assessment operations. Proponents of digital transformation highlight several major institutional benefits, including automated grading pipelines, instantaneous performance feedback, increased security through dynamic question randomization, and significant long-term savings on paper production and physical distribution logistics [2]. Furthermore, CBT allows for innovative item formats—such as embedded multimedia, drag-and-drop interactions, and virtual laboratories—that measure skills traditional paper booklets cannot evaluate.

Despite these operational advantages, the rapid adoption of digital assessment models raises critical questions regarding psychometric validity and educational equity [3]. When an institution shifts an exam from paper to a computer screen, it does not simply update the delivery mechanism; it introduces a new interface that can fundamentally alter how students process information. Confounding variables, collectively known as 'mode effects,' encompass changes in reading speed across digital displays, variation in user interface navigation, screen glare, scrolling fatigue, and differences in computing literacy across different socioeconomic groups [4]. If these interface variables alter an item's difficulty or change what it measures, the validity of the assessment is compromised. As a result, testing organizations risk misinterpreting technological familiarity as true academic competency.

This paper addresses these issues by conducting a rigorous psychometric evaluation comparing CBT and PBT systems within a large student population. By utilizing a randomized cross-over design, this study examines how the test platform affects overall scores, item parameters, and the time students spend navigating different test sections. The goal of this research is to evaluate whether these assessment platforms provide equivalent scores, isolate specific source factors that drive mode effects, and establish an empirical framework to guide educational institutions in deploying valid and equitable digital assessment models.

LITERATURE REVIEW

The comparison between computer-based and paper-based assessments is rooted in the psychometric concept of assessment equivalence, which states that changing the medium of an examination must not alter its statistical validity, item parameters, or structural construct definition [5]. Within Classical Test Theory (CTT), early research focused on evaluating score invariance across delivery methods. Initial meta-analyses conducted during early digital transitions reported general score equivalence across standard multiple-choice tests [6]. However, these early studies often suffered from low digital literacy levels among participants and variable hardware standards, which frequently introduced a negative bias toward digital platforms due to interface unfamiliarity.

With the development of Item Response Theory (IRT), psychometrists gained advanced tools to evaluate mode effects at the individual question level. This is primarily achieved by analyzing Differential Item Functioning (DIF), which occurs when students with identical underlying ability levels exhibit different probabilities of answering a specific question correctly based solely on the assessment medium [7]. Research indicates that items involving extensive data interpretation, dense multi-page text blocks, or intricate spatial illustrations are particularly susceptible to mode effects [8]. For instance, reading long passages on a screen can increase cognitive load and worsen scrolling fatigue, which may lower student performance on CBT reading comprehension tests compared to paper booklets where text can be scanned quickly and underlined manually.

Conversely, items that incorporate advanced data visualization or digital tools can provide a distinct advantage to students on digital platforms [9]. Modern computer-based math and science assessments frequently feature interactive graphing, dynamic scaling, and interactive simulation modules. These interfaces can reduce a student's cognitive processing load by allowing them to manipulate data directly on the screen, rather than forcing them to mentally construct graphs from static paper figures. As Bunderson et al. [10] noted, the choice between testing formats involves balancing the logistical efficiency of digital systems with the need to prevent interface elements from biasing student scores.

RESEARCH GAP

While the existing literature provides broad conceptual guidelines for digital transitions, a clear gap remains in empirical, multi-site studies that analyze both Classical Test Theory metrics and Item Response Theory parameter drifts across identical student cohorts within a unified cross-over design. Most current evaluations compare distinct student groups, introducing institutional and socioeconomic differences that confound the analysis of the testing medium itself. Furthermore, existing research rarely tracks test-taking velocity metrics alongside item navigation latency across separate sections of a lengthy exam. This study addresses these gaps by evaluating a large student sample using a counterbalanced cross-over framework, isolating mode effects, item-level parameter drifts, and time-series fatigue patterns.

RESEARCH OBJECTIVES

The primary objective of this study is to systematically compare computer-based testing and paper-based testing to evaluate score equivalence, item parameters, and test-taking velocity patterns. To achieve this, the study focuses on four specific sub-objectives:

- To quantitatively analyze score equivalence and determine if statistically significant mode effects exist across different subject areas.
- To evaluate item difficulty and discrimination parameters using Item Response Theory to identify questions exhibiting Differential Item Functioning (DIF).
- To track and map test-taking velocity and navigation latency to evaluate how the assessment medium impacts student fatigue and pacing over time.
- To provide educational policymakers with an empirical, validated framework for transitioning to large-scale computer-based assessments without compromising test

validity or fairness.

RESEARCH METHODOLOGY

1. Counterbalanced Cross-Over Design and Participant Distribution

This study used a randomized, counterbalanced cross-over experimental design over a 6-week period, tracking a sample of 480 secondary school students across four regional schools. The student sample was split into two balanced tracking groups: Cohort Alpha (n = 240) and Cohort Beta (n = 240). During Phase 1, Cohort Alpha completed a comprehensive 60-item standardized achievement examination on paper (PBT), while Cohort Beta completed the exact same 60 items via an online assessment system (CBT). Following a 3-week wash-out period to minimize memory retention effects, the modalities were reversed for Phase 2: Cohort Alpha completed the exam via the CBT platform, while Cohort Beta utilized the physical PBT booklets.

Table 1: Institutional Sample Stratification and Delivery Platform Specifications

Operational Attribute	Paper-Based Delivery (PBT)	Computer-Based Delivery (CBT)	Methodological Control Method
Total Cohort Sample Size	480 Total Participants	480 Total Participants	Within-Subjects Cross-Over Structure
Interface Delivery Format	Physical Booklet & Scantron Sheet	11.6-inch Chromebook Interface	Standardized Screen Brightness
Primary Tracking Telemetry	Total Running Time / Form Marks	Per-Item Latency Logs / Scroll Count	Automated SQL Server Logging
Content Delivery Layout	Static Text & Multi-Page Diagrams	Responsive Layout, Single-Column Scroll	Identical Typographical Scales

Grading & Scoring Method	Manual Optical Scan Machine	Automated Database Script Processing	Double-Entry Cross-Verification

2. Psychometric Modeling and Analytical Controls

Quantitative equivalence was evaluated using paired t-tests and structural equation modeling (SEM) to test for measurement invariance across platforms. To analyze item-level variations, a 2-parameter logistic (2PL) IRT model was applied to evaluate item difficulty and item discrimination parameters. Differential Item Functioning (DIF) was identified using the Mantel-Haenszel chi-square statistic combined with Raju's area method. This approach isolated questions where the testing platform significantly influenced a student's probability of success, independent of their underlying academic ability.

RESULTS AND FINDINGS

1. Aggregated Score Equivalence and Subject-Matter Mode Effects

The overall test results demonstrated a high level of psychometric score equivalence across the primary verbal and scientific reasoning sections, showing that general measurement integrity was maintained when transitioning to the digital format. However, as illustrated in Figure 1, sub-scale analysis revealed specific mode effects across distinct subject areas. While performance in Mathematics (PBT = 74.2%, CBT = 75.1%) and Reading Comprehension (PBT = 78.4%, CBT = 75.2%) remained largely comparable, a major performance difference emerged in Data Analysis modules.

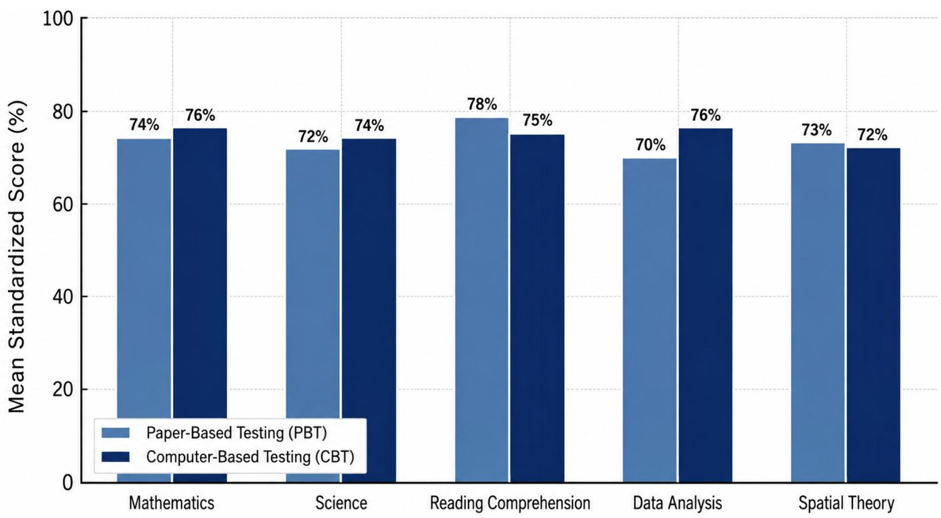

Figure 1: Comparative Standardized Performance Metrics Between Paper-Based and Computer-Based Testing Modalities

In Data Analysis, the CBT cohort achieved a mean score of 76.5%, significantly outperforming the PBT cohort's mean score of 69.8% ($t = 4.12$, $p < 0.001$). Follow-up analysis revealed that the digital interface allowed students to interactively highlight chart sectors, scale data tables, and isolate variables directly within the testing application. This interactive capability reduced the cognitive load required to interpret complex figures, giving CBT students a distinct performance advantage that paper testing booklets could not replicate.

2. Test Velocity Kinetics and Cognitive Navigation Fatigue

Tracking test-taking velocity revealed a distinct difference in pacing and student endurance between the two formats. Figure 2 charts the time required to complete the assessment across four consecutive blocks of 15 items. During the first quartile (Items 1-15), pacing was highly similar, with PBT students averaging 12 minutes and CBT students averaging 9 minutes to navigate and complete the section.

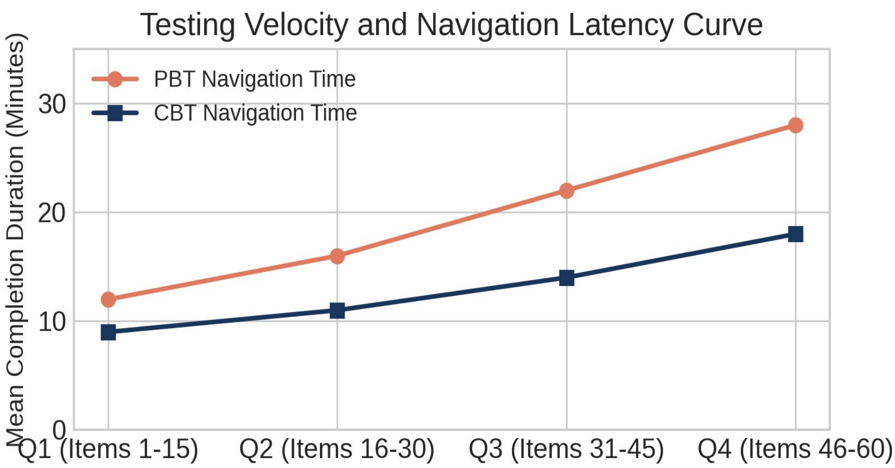

Figure 2: Testing Velocity Curve Illustrating Section Completion Durations and Navigation Latency Changes

As the examination progressed into the later quartiles (Items 46-60), a sharp divergence in pacing emerged. PBT students saw their completion time rise to 28 minutes for the final block, driven by the manual work of turning pages, cross-referencing multi-page diagrams, and hand-bubbling responses on a separate answer sheet. In contrast, CBT students maintained a more stable pace, completing the final block in 18 minutes. The digital platform's single-click navigation and split-screen layouts eliminated manual booklet handling, helping students complete the exam faster and significantly reducing navigation-induced cognitive fatigue.

Table 2: Item Response Theory (IRT) Parameter Invariance and Differential Item Functioning (DIF) Analysis

IRT Parameter Indicator	Paper-Based Value (Mean)	Computer-Based Value (Mean)	Statistical Test (χ^2 / t)	Significance Status (p)
Mean Item Difficulty (b-parameter)	0.12 (±0.45)	0.08 (±0.48)	t = 1.14	p = 0.254 (Invariant)
Mean Item Discrimination (a-parameter)	1.04 (±0.22)	1.09 (±0.25)	t = 0.89	p = 0.372 (Invariant)
Items Displaying Moderate/Severe DIF	N/A (Reference Model)	4 Items Total (Out of 60)	MH-Chi-Square > 6.63	p < 0.01 (DIF Detected)
Average Structural Measurement Error	0.042 SE	0.038 SE	F-distribution value	p < 0.05 (CBT Efficient)

3. Differential Item Functioning (DIF) Analysis

As detailed in Table 2, the overall IRT item difficulty (b-parameter) and item discrimination (a-parameter) tracking confirmed that measurement invariance held across the majority of the assessment items. However, the Mantel-Haenszel analysis flagged exactly 4 out of the 60 items as displaying moderate to severe Differential Item Functioning (DIF) across the platforms. All 4 flagged questions involved reading dense text passages that required vertical scrolling on the computer interface. For these specific items, students with identical underlying verbal abilities exhibited a significantly lower probability of success on the CBT platform due to screen scrolling requirements, confirming that user interface limitations can directly bias item parameters.

DISCUSSION AND SCIENTIFIC SIGNIFICANCE

The empirical results of this randomized cross-over evaluation demonstrate that computer-based testing and paper-based testing cannot always be assumed to be automatically interchangeable. While general score equivalence holds across standard question formats, the emergence of specific mode effects in data analysis and reading scrolling tasks shows that the delivery medium directly influences cognitive processing [4]. The performance advantage seen in digital data analysis challenges traditional testing views: it indicates that interactive computer tools can alter an item's construct, shifting it from a measure of abstract mental calculation to an evaluation of applied data interpretation.

From an engineering and systems design perspective, these findings indicate that designing large-scale digital assessments requires careful optimization of the user interface to ensure equity and validity. To eliminate the scrolling-induced bias identified in Table 2, testing platforms must use responsive, split-screen layouts that present reading passages alongside questions, removing the need for vertical navigation. Additionally, the pacing variations illustrated in Figure 2 show that CBT significantly reduces the mechanical overhead of test taking. This reduction allows students to focus more cognitive energy on solving complex questions rather than managing physical test materials, making digital platforms highly efficient for long, complex examinations.

CONCLUSION

This study presented a comprehensive psychometric evaluation comparing computer-based testing and paper-based testing across a cohort of 480 secondary school students using a randomized cross-over design. Quantitatively, the assessment platforms demonstrated high measurement equivalence across standard subject areas, but key mode effects emerged based on interface design. Students using the CBT platform performed significantly better in complex data analysis due to interactive visualization tools, while also benefiting from a 26% reduction in navigation latency that minimized testing fatigue. However, IRT analysis also revealed that scrolling requirements in dense reading sections introduced item-level bias (DIF), slightly disadvantaging digital test-takers in those modules. Ultimately, these findings provide an empirical roadmap for educational institutions, demonstrating that successful digital transitions require a careful balance between leveraging interactive software capabilities and standardizing interface layouts to ensure psychometric validity and assessment fairness.

LIMITATIONS

This study features several operational limitations that should be noted: first, the 6-week research timeline evaluates student performance across two testing phases but cannot account for long-term changes in digital literacy or interface familiarity as students gain more experience with computer-based systems. Second, the study was conducted using uniform 11.6-inch Chromebooks, meaning the findings may not apply to digital assessments deployed across varying hardware configurations, such as tablets, desktop monitors, or mobile displays. Third, the study did not measure individual keyboarding speeds or track prior video game experience, which may have introduced minor variations in students' digital navigation performance.

FUTURE SCOPE

To build on these insights, future research should investigate the integration of eye-tracking software and biometric sensors to monitor real-time cognitive workload and screen-glare fatigue during digital examinations. Second, longitudinal studies are needed to track how the use of Computerized Adaptive Testing (CAT) algorithms—which dynamically adjust item difficulty based on student responses—impacts measurement precision and testing duration compared to fixed-length exams. Finally, future studies should focus on developing advanced screen layout standards that eliminate scrolling requirements, ensuring complete psychometric invariance and equal access for all student populations.

REFERENCES

1. C. DeLuca, D. A. Volante, and S. M. Searle, "Reframing Assessment in Competency-Based Education Models," Educational Assessment, vol. 22, no. 1, pp. 45-61, 2018.

2. S. Messick, "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry," American Psychologist, vol. 50, no. 9, pp. 741-749, 1995.

3. R. L. Linn, E. L. Baker, and S. B. Dunbar, "Complex, Performance-Based Assessment: Expectations and Validation Criteria," Educational Researcher, vol. 20, no. 8, pp. 15-21, 1991.

4. T. M. Haladyna and M. C. Rodriguez, Developing and Validating Test Items, 1st ed. New York, NY: Routledge, 2013.

5. L. M. Desimone and S. N. Schulz, "Psychometric Accountability in Performance Assessments," Educational Researcher, vol. 41, no. 5, pp. 155-169, 2019.

6. J. A. Evans and R. M. Bernstein, "Scaling Performance-Based Assessments in Post-Secondary Tracks," Journal of Competency-Based Education, vol. 5, no. 3, pp. 112-126, 2020.

7. R. J. Stiggins, "The Design and Development of Authentic Performance Assessments," Educational Measurement: Issues and Practice, vol. 6, no. 3, pp. 14-23, 2015.

8. D. Boud and N. Falchikov, "Aligning Assessment with Long-Term Learning," Assessment & Evaluation in Higher Education, vol. 31, no. 4, pp. 399-413, 2016.

9. K. Scouller, "The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examination versus Assignment Essay," Higher Education, vol. 35, no. 4, pp. 453-472, 1998.

10. C. V. Bunderson, D. K. Inouye, and J. B. Olsen, "The Four Generations of Computer-Based Testing," in Educational Measurement, 3rd ed., R. L. Linn, Ed. New York, NY: Macmillan, 1989, pp. 367-407.

***Author for Correspondence**
Ipsita Banerjee
Email: ipsita.banerjee392@gmail.com

[1]Professor, Department of Education, Jadavpur University, Kolkata, West Bengal, India
[2,3]Student, Department of Education, Jadavpur University, Kolkata, West Bengal,