

Examining mode effects for an adapted Chinese critical thinking assessment

Lin Gu , Guangming Ling , Ou Lydia Liu , Zhitong Yang , Guirong Li , Elena Kardanova & Prashant Loyalka

To cite this article: Lin Gu , Guangming Ling , Ou Lydia Liu , Zhitong Yang , Guirong Li , Elena Kardanova & Prashant Loyalka (2020): Examining mode effects for an adapted Chinese critical thinking assessment, Assessment & Evaluation in Higher Education, DOI: [10.1080/02602938.2020.1836121](https://doi.org/10.1080/02602938.2020.1836121)

To link to this article: <https://doi.org/10.1080/02602938.2020.1836121>



Published online: 28 Oct 2020.



Submit your article to this journal [↗](#)



Article views: 144




View related articles [↗](#)



View Crossmark data [↗](#)



Examining mode effects for an adapted Chinese critical thinking assessment

Lin Gu^a , Guangming Ling^a, Ou Lydia Liu^a, Zhitong Yang^a, Guirong Li^b, Elena Kardanova^c and Prashant Loyalka^d

^aResearch & Development Area, Educational Testing Service, Princeton, NJ, USA; ^bSchool of Education, Henan University, People's Republic of China; ^cInstitute of Education, National Research University Higher School of Economics, Moscow, Russia; ^dGraduate School of Education, Stanford University, Stanford, CA, USA

ABSTRACT

We examine the effects of computer-based versus paper-based assessment of critical thinking skills, adapted from English (in the U.S.) to Chinese. Using data collected based on a random assignment between the two modes in multiple Chinese colleges, we investigate mode effects from multiple perspectives: mean scores, measurement precision, item functioning (i.e. item difficulty and discrimination), response behavior (i.e. test completion and item omission), and user perceptions. Our findings shed light on assessment and item properties that could be the sources of mode effects. At the test level, we find that the computer-based test is more difficult and more speeded than the paper-based test. We speculate that these differences are attributable to the test's structure, its high demands on reading, and test-taking flexibility afforded under the paper testing mode. Item-level evaluation allows us to identify item characteristics that are prone to mode effects, including targeted cognitive skill, response type, and the amount of adaptation between modes. Implications for test design are discussed, and actionable design suggestions are offered with the goal of minimizing mode effect.

KEYWORDS

Mode effect; critical thinking; student learning outcomes; adaption

Introduction

In response to a range of trends, challenges and paradigm shifts in higher education, there is a growing need for internationally comparable data on college student learning outcomes (SLOs; Tremblay, Lalancette, and Roseveare 2012). The need to provide evidence of SLOs at the global level has given rise to an increased interest in the use of standardized assessments in an international context. One notable example is the Assessment of Higher Education Learning Outcomes (AHELO) feasibility study sponsored by the Organization for Economic Co-operation and Development (OECD; Coates and Richardson 2012; Tremblay, Lalancette, and Roseveare 2012; Richardson and Coates 2014), where generic skills, including critical thinking skills, were identified as important part of learning outcomes. A major objective of AHELO has been to assess whether it is possible to develop international measures of SLOs.

Efforts to develop international measures of SLOs have coincided with a transition from paper-based testing to computer-based testing (Buerger and Goldhammer 2016); it has become

increasingly common to use computer-based instead of paper-based test in international assessment programs such as the Programme for International Student Assessment (PISA) and the Programme for International Assessment of Adult Competencies (PIAAC). Compared to paper-based tests, computer-based tests promise a variety of advantages that can help overcome challenges often encountered in international assessments. For example, computer-based tests permit flexibility in test scheduling and location, allowing the delivery of tests to large numbers of test-takers who are typically geographically distant. Enhanced standardization of the test administration also contributes to the comparability among assessment results collected at different geographic locations and/or in different cultural contexts, a concern of paramount importance in international assessments (Richardson and Coates 2014). Test-taking behavioral data (e.g. time spent on task, process sequence information, use of stimulus elements) made available with computer-based tests can assist in understanding test-takers' response processes, which is an integral part of test validation according to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] 2014; hereafter referred to as the *Standards*). Such behavioral data have been used to help develop comparable scores across countries (OECD 2012). Finally, with the COVID-19 impact, it seems likely that more online remote testing is going to be adopted, which presents the needs to examine the comparability of computer-based tests in relation to the traditional paper-based tests or in-person testing.

Although the benefits of computer-based tests are widely recognized, the variability of technology and physical resources as well as user experience and familiarity with technology across countries, makes it challenging to use computer technology in international assessments. In the U.S., even though it is common for students to have access to computer technology, through virtual schools and online learning (Prisacari and Danielson 2017), opportunity for access at school or home may not be equally guaranteed for all students (e.g. Gray, Thomas, and Lewis 2010). This level of access cannot be assumed in other parts of the world. For example, although the AHELO initiative used computer-based tests as the default delivery mode, students in Kuwait who had less exposure to (and were therefore perhaps less proficient in or comfortable with) computer use were provided with the option of paper-based tests (Tremblay, Lalancette, and Roseveare 2012). In the case of PISA and PIAAC, both assessment modes are still in use (OECD 2012, 2013). The variations among countries and regions in terms of the adoption of computer-based tests suggests that both computer and paper-based tests are likely to coexist for an extended duration of time, which presents the necessity of continuous research and monitoring of possible mode effects and ensure the comparability.

In sum, variability in technology capacity, accessibility and usage across countries and regions necessitates the use of both paper-based and computer-based test modes in large-scale international assessments. In such cases, as suggested by the *Standards* (AERA, APA, & NCME 2014) and the International Test Commission Guidelines for Translating and Adapting Tests (International Test Commission 2017), it is imperative to gather and document evidence regarding the impact of mode effects on score interchangeability.

According to Wang et al. (2007), mode effects refer to performance on an assessment due to differences in assessment delivery method and setting rather than the targeted constructs or abilities. Studies on mode effects have investigated the comparability of various types of test information. When mode effects are investigated at the test level, attention has often focused on mean scores and score variability. For example, Bennett et al. (2008) found that computer-based testing results in lower mean scores and more score variability than paper-based tests. Poggio et al. (2005), on the other hand, found equivalence between computer-based and paper-based tests in terms of mean scores. Similarly, Brunfaut, Harding, and Batty (2018) found small to no mode effect at the test level on a writing test suite. Other evaluations include Kim and Huynh (2008), who examined the test characteristic curve (TCC) and test information function (TIF) within item response theory, and found closely overlapping TCC and TIF between computer-

based and paper-based tests. Test-level mode effects have also been investigated in terms of score reliability. For example, comparable reliability estimates have been found by Arce-Ferrer and Guzman (2009) and in Gallagher et al. (2002), indicating equivalent measurement precision across modes. Kroehne, Gnabbs and Goldhammer (2019) also suggested strategies that were designated to promote motivation in the context of unstandardized online testing (unproctored or remote-proctored computer-based tests), acknowledging the mode effects that have been widely studied.

On the other hand, the examination of mode effects can also be conducted at the item level. Item-level evaluation can show how test-takers interact with item features and, thus, provide insights into sources of mode effects by establishing associations between differential item performance and item characteristics (Pommerich 2004). Commonly used methodologies include comparing item statistics (e.g. difficulty and item discrimination) of the same item delivered in different modes, as well as conducting differential item functioning analysis to determine if the same item functions differently across modes at the same ability level. For example, mode effects have been associated with the amount of reading required for item completion. Studies have found that compared to items with little reading demand, mode effects are larger for passage-based items requiring some form of navigation (e.g. scrolling, paging) because the content cannot be viewed in full on a single computer screen (Kim and Huynh 2008; Schwarz, Rich, and Podrabsky 2003). In these studies, it was suspected that navigation complicates the response process by imposing additional cognitive demands on the test-takers.

In addition to examining test and test item performance differences, understanding mode effects between computer-based and paper-based tests has also been approached through the investigation of response behaviors. Past research has compared completion rates at the test level (Gallagher et al. 2002; Goldberg and Pedulla 2002; Pommerich 2004), as well as missing responses at the item level (OECD 2012, 2016). Goldberg and Pedulla (2002), for example, found that a paper test is less speedy than a computer test, highlighting the importance of evaluating the effects of time constraints on performance across modes. Items requiring the use of a scrollbar were found to have higher levels of missing data than paper-based tests (OECD 2012).

Both Kolen and Brennan (1995) and Pommerich (2004) argued that, as mode effects tend to be specific to a given test and computer interface, there is a need to conduct comparability studies for any tests offered in multiple modes. In this study we investigate the differential effects of computer-based and paper-based tests on performance and response behavior on a SLO assessment used in an international context. More specifically, we examine the *HElghten*® Critical Thinking assessment, one of the modules in the *HElghten* Outcomes Assessment Suite developed by Educational Testing Service to measure learning outcomes that are essential for college academic success (Liu, Frankel, and Roohr 2014; Liu et al. 2016). Through research initiatives with institutions across the globe, the *HElghten* Critical Thinking assessment has been translated and adapted from the original source language (English) to other languages (e.g. Chinese, Russian, Korean, and Spanish) to be used in diverse international contexts. Variability in technology capacity, accessibility and usage across countries and regions makes it necessary to offer the translated and adapted assessments in both paper-based and computer-based modes in some countries. Therefore, it is necessary to examine the comparability between paper-based and computer-based tests in these countries where both modes were used.

In addition, there appears to be little empirical research in the literature about testing mode effects for standardized assessments in Chinese, though many mode effects studies have been conducted on English-delivered assessments in the literature. It needs to be noted that even though some tests in China are delivered using a computer (e.g. CET-4 and 6), most of the assessments in the K-12 (e.g. the Senior High School Entrance Examination) and college level (e.g. the College Entrance Examination or Gaokao and Post Graduate Admission Test) are delivered using paper-based tests. Thus, more empirical research is deemed necessary as the digitalization trend continues to evolve in China.

Table 1. Sample sizes across institutions, grades, and modes.

		N	Computer	Paper	%
Institution	1	101	50	51	17.9
	2	74	37	37	13.1
	3	390	191	199	69.0
Grade	Freshmen	299	151	148	52.9
	Juniors	266	127	139	47.1
Total		565	278	287	100

Using data collected from a randomized between-group experiment study in China, we examine mode effects in terms of mean test performance, measurement precision, item functioning, response behaviors and test-taker perceptions. Specifically, the following research question was the focus: Whether there exists any difference between computer-based and paper-based tests on the *HElghten* Critical Thinking Assessment in terms of (a) mean test score, (b) score reliability, (c) item difficulty and discrimination, (d) test-level completion rates and item-level missing values, and (e) test-taker perceptions on test difficulty and testing time.

Method

Study sample

A total of 565 freshmen and juniors, majoring in electric engineering (EE) and computer science (CS), were recruited from three Chinese universities. They were randomly assigned within grade-specific classrooms to either a computer-based ($N = 278$) or paper-based ($N = 287$) version of the *HElghten* CT assessment (see Table 1). There were more than twice as many males as females, as gender imbalance is common in electrical engineering and computer science majors. To evaluate comparability of the two groups, we checked and confirmed there was no statistical dependence between testing mode and each of the background variables collected, including gender, area of origin (rural or not), father's education (high school or below), mother's education (high school or below), high school type (elite or not) and socioeconomic status.

Previous research suggests that test-taking motivation in a low or no-stakes outcomes assessment could be problematic (e.g. Liu, Rios, and Borden 2015). However, Chinese students appear to make effortful performances even with no stakes. For example, Gneezy et al. (2017) reported that Chinese students had top performance on a no-stakes assessment in a research context and their performance did not differ between those who were and were not offered a financial incentive. It seems reasonable to assume that students in this study, regardless of the testing mode assigned, were highly motivated to perform well on the assessment.

The *HElghten* CT assessment

One operational form of the *HElghten* CT assessment, translated and adapted from English to Chinese, in both computer-based and paper-based tests, was used in this study (Liu et al. 2018). This form consisted of 26 items, each with a score of one if correct, and 0 otherwise, which leads to the total raw score scale between 0 and 26. According to our review of literature, both item characteristics and computer interface display could potentially contribute to mode effects between paper-based and computer-based tests.

Item characteristics

There are two types of items: set items and freestanding items. Set items are a set of items that are all based on a common stimulus (e.g. a reading material). There were four item sets. Three of them were each based on a common reading material. The reading materials each had around

Table 2. Summary of HEIghten CT item characteristics in computer-based test.

Item #	Response Type		Item Set	Scrolling	Highlighting
1	Inline choice	Single-selection	Item Set 1	Yes	No
2	Inline choice	Single-selection			Yes
3	Inline choice	Single-selection			No
4	Inline choice	All that apply	Item Set 2	Yes	Yes
5	Radio button	Single-selection			Yes
6	Radio button	All that apply			Yes
7	Radio button	Single-selection	Freestanding	No	No
8	Radio button	Single-selection	Freestanding	No	No
9	Radio button	Single-selection	Item Set 2	Yes	Yes
10	Radio button	Single-selection			No
11	Radio button	Single-selection			Yes
12	Drop-down	Single-selection	Item Set 3	No	Yes
13	Radio button	Single-selection			No
14	Inline choice	Single-selection			No
15	Radio button	Single-selection	Item Set 4	Yes	No
16	Radio button	Single-selection			No
17	Radio button	Single-selection			No
18	Radio button	Single-selection	Item Set 4	Yes	No
19	Radio button	Double-selection			No
20	Radio button	Single-selection			No
21	Radio button	Single-selection	Freestanding	No	No
22	Radio button	Single-selection			No
23	Inline choice	Single-selection			No
24	Radio button	Double-selection	Freestanding	No	No
25	Radio button	Single-selection			No
26	Radio button	Single-selection			No

1100 Chinese characters (about 625 English words), making scrolling necessary to read all the content. The other set had the stimulus visible in its entirety on the screen. Besides the set items, there were four freestanding items; none of these items required scrolling to read the stimulus.

Three response types were used under the computer-based test condition. To indicate an answer choice, test-takers were asked (a) to make an inline choice, that is, to click directly on the relevant part (sentence) in the reading material; (b) to use a radio button (a circle that locates in front of a statement/sentence) to select; or (c) to use a drop-down menu to make a selection. There were 6, 19 and one item(s) of each of these item types, respectively. Most items (22) had only one correct answer, two items had two answer choices together, and two items were select-all-that-apply questions. A handful of the items referred to a specific part of the stimulus (e.g. a passage with multiple sentences). In these cases, the referenced part of the stimulus was highlighted to assist test-takers in locating the relevant information.

Table 2 gives a summary of item characteristics, which shows item ordering number within the test, response type, item set status, scrolling requirement and use of highlighting.

Computer interface design

The computer interface for the assessment was divided into panels. At the top of the screen was an informational panel, showing the current question number, total number of questions, and total time remaining in the test section. At the bottom of the screen was a navigation panel. This panel included buttons that allowed test-takers to navigate forward and backward, mark a question for later review, access help information, display the item review screen and exit the test.

The middle section of the screen was used for displaying items and was split into two parts of equal size, with the stimulus appearing in a panel on the left half and an associated item appearing in a panel on the right half. Items were presented on the screen one at a time. A

vertical or horizontal sliding scroll bar would show if within-panel scrolling was needed to view the content. The font size of the text could be adjusted. During the test, scrap paper could be requested for use in answering any of the test items delivered via computer-based test.

Test-takers were allowed to skip items, meaning that they could proceed to the next item without providing an answer to the current one. Changing previously entered answers was also permitted. A review screen was presented after the last test item was answered or skipped. In the review screen, the status of each question on the test was displayed, including completed, incomplete and marked items. From the review screen test-takers were able to select any incomplete items to complete or to revise any answers that were already provided earlier.

Paper adaptation

The paper-based test was adapted from the computer-based test with the same time constraints and item ordering. Test-takers could move freely throughout the test booklet and could request scrap paper. Although efforts were taken to minimize the differences between computer-based and paper-based tests, the adaptation from computer-based to paper-based test nevertheless necessitated a few changes in terms of item presentation and response elicitation method. First, stimulus input was displayed in its entirety, followed by the associated item or items which may or may not have been visible concurrently with the stimulus as they are in computer-based test. Multiple items could be presented on a single printed page. Second, the response elicitation method for the inline choice items needed considerable adaption as it was no longer feasible to ask the test-takers to click directly in the portion of the stimulus to indicate their answers because of the use of a Scantron-like answer sheet with bubbles. In these cases, test-takers were simply given a list of all possible choices they could have made using inline selection in a computer-based test mode and were asked to select their answers in a multiple-choice format. Third, when an item set had multiple items that required the use of highlighting, multiple parts of the stimulus input were highlighted in the printed test booklet, with each being marked with a note indicating the associated item number. This was different from the computer-based test where items were presented one at a time and, therefore, only one highlighted part would appear for an item.

Both paper-based and computer-based tests were administered in an in-person proctored setting, with the only difference being the test delivery mode.

Analysis

We examined mode effects in multiple steps. To examine mode effects on mean scores across mode, an independent sample t test was used to compare the total score between computer-based and paper-based tests. To evaluate measurement precision, reliability estimates both at the individual and institutional levels were compared. We used Cronbach's alpha to estimate the total score reliability. As the *HElghten* assessment suite is also intended for use at the institutional level, we also calculated institutional-level reliability using a split-sample approach (Klein et al. 2007). This procedure involves randomly splitting the students in each school into two samples (Sample A and Sample B), computing mean scores for both samples at each school, and correlating Sample A and Sample B means across all the schools. A Spearman-Brown correction was used to adjust for the use of half-size samples. In our analyses, the mean of 30 random splits was computed to obtain a stable estimate of the expected value of school-level reliability.

Mode effects pertaining to item functioning were examined through the comparison of item difficulty and item discrimination across modes, both at the test and item level. Item difficulty was calculated as the proportion correct (i.e. p -value). Item discrimination was calculated using uncorrected item-total point-biserial (i.e. r_{pbis}). In addition, we also examined correlations of item difficulty and item discrimination within each mode.

Table 3. Descriptive statistics of total score between computer-based test and paper-based test.

Mode	N	M	SD	Skewness	Kurtosis
computer-based test	278	12.18	3.76	-.55 (<i>SE</i> = .15)	-.18 (<i>SE</i> = .29)
paper-based test	287	13.69	3.33	-.52 (<i>SE</i> = .14)	.50 (<i>SE</i> = .29)
Cohen's <i>d</i>		.43			

We also investigated differential response behaviors across modes, both at the test and item levels. At the test level we calculated the proportion of students who completed 75% and 100% of the test. Chi-square (χ^2) testing was used to examine the association between completion proportion and test mode. At the individual item level, we calculated the difference in the proportion of missing responses across modes by item and identified the items with the highest and lowest proportion of missing data in each mode.

Regarding user perceptions, test-takers were asked to report perceived test difficulty and whether they had enough time to finish the test, both of which were reported on a three-point Likert scale in a post-test survey. We calculated the proportions of those who gave different ratings of test difficulty and testing time and used a chi-square test to compare the proportions across modes.

Results

Mean scores

The descriptive statistics of total test score are shown in Table 3 for each mode. The *t* test results showed the presence of a statistically significant mode effect in favor of paper-based tests. Test-takers who took the paper-based test ($M = 13.69$, $SD = 3.33$) significantly outperformed those who took the computer-based test ($M = 12.18$, $SD = 3.76$) with a close-to-medium effect size, $t_{(563)} = 5.07$, $p < .01$, $d = 0.43$. In addition, computer-based test scores were slightly more heterogeneous.

Measurement precision

Table 4 summarizes total score reliability estimates at the institutional and individual levels. At the institutional level, the reliability was .96 for computer-based tests and .84 for paper-based tests. At the individual level, the reliability was .64 for computer-based tests and .60 for paper-based tests. Both types of reliability estimates were higher for the computer test.

Item functioning

Comparisons of item difficulty and item discrimination are shown in Table 5. The average proportion correct was higher for paper-based ($p = .53$) than for computer-based tests ($p = .47$). Most of the items (19 out of 26) appeared to be easier on paper than on computer. Item difficulty also showed a greater variability under the paper mode than the computer mode. In addition, item difficulty estimates were highly correlated between computer-based and paper-based tests ($r = .97$). The average item discrimination did not differ greatly between computer-based ($r_{pbis} = .32$) and paper-based tests ($r_{pbis} = .30$). About half of the items ($n = 14$) appeared to be more discriminating on computer than on paper. The correlation of item discrimination estimates between computer-based and paper-based tests was only .65, much lower than the correlation of item difficulty estimates.

A closer examination at the item level revealed that Item #4 and Item #9 were the two with the largest difference in difficulty ($p_{diff} = .18$) among all items. Both items appeared to be easier on paper than on computer. Item #4 was an inline choice item in the computer-based test,

Table 4. Institutional- and individual-level reliability across modes.

Mode	N	Institutional-Level Reliability	Individual-Level Reliability
computer-based test	278	.96	.64
paper-based test	287	.84	.60

Table 5. Item difficulty and item discrimination across modes.

	Item Difficulty		Item Discrimination	
	computer-based test	paper-based test	computer-based test	paper-based test
Mean	.47	.53	.32	.30
Range	.14 ~ .81	.10 ~ .92	.12 ~ .60	.01 ~ .52
Between Mode Correlation	.97		.65	

which asked test-takers to review a list of facts presented in the stimulus and select all that apply. The whole list was not visible in its entirety on the screen, and scrolling was needed to view all its content. To respond to this item, test-takers needed to locate the list in the stimulus by scrolling and then to scroll up and down to view the content of the list. Being a “select all that apply” response-type item further complicated the response process. If multiple facts were selected, they were not necessarily visible simultaneously on the same screen, potentially making it challenging for item review.

In comparison, the response process of this item on paper was much more straightforward. The entire list fitted on one page of the booklet so that all the selections made by test-takers could be viewed at once. The list was also repeated, showing immediately after the item stem. Thus, there was no need to turn the pages to locate the list presented earlier on. We speculate that, compared to the paper-based test, responding to this item in the computer-based test required more cognitive load, making this item easier on paper than on computer.

The other item, Item #9, asked test-takers to make a single selection using a radio button. To respond, test-takers were asked to demonstrate the targeted cognitive skill: that is, to connect two pieces of information in the stimulus. The two pieces of information could not be viewed simultaneously in the computer-based test as they were spaced far apart while they were placed on the same page in the paper-based test, potentially making it easier for test-takers to identify the connection.

Response behaviors

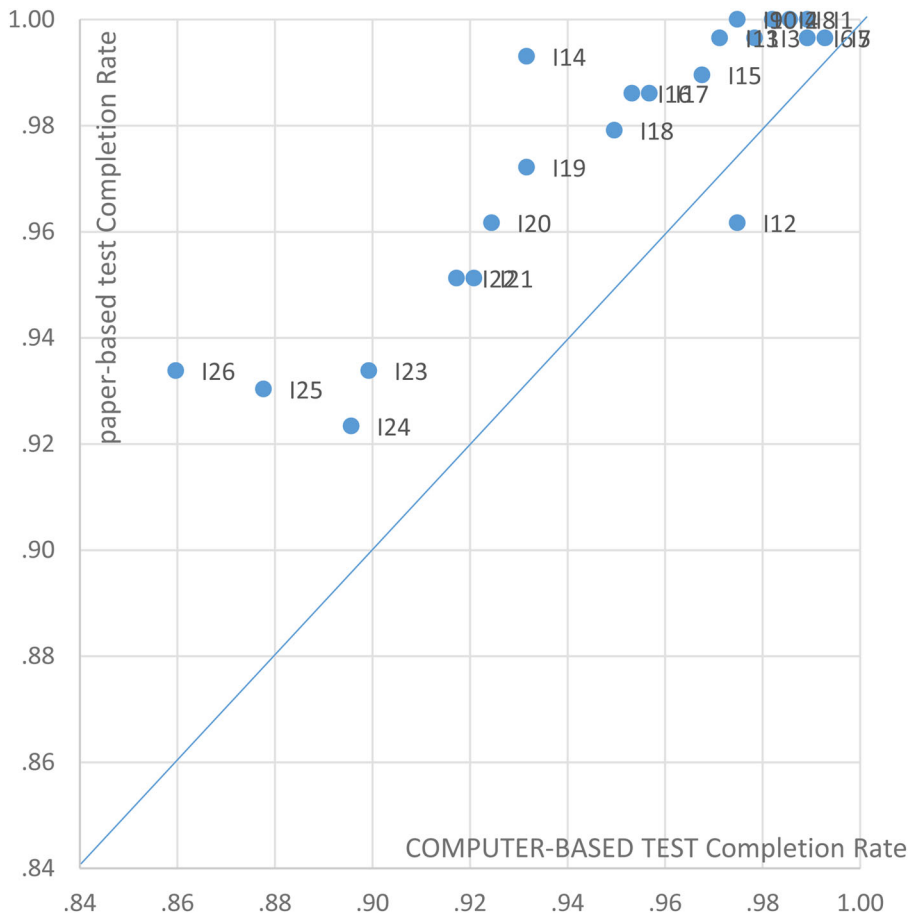
As shown in Table 6, 97% of the paper-based test-takers completed 75% of the test, compared with 92% of the computer-based test-takers. The mode and completion rate association was significant, $\chi^2 = 5.482$, $p < .05$. The proportion of students who completed all items was also significantly dependent on the mode condition (87% versus 80% on paper and computer), $\chi^2 = 5.899$, $p < .05$.

At the item level, paper-based test takers had a completion rate of 98% across items on average, compared to 95% for computer-based test takers. All items, except for one, had a higher proportion of missing data in the computer-based test, consistent with the findings based on test-level completion rate. Figure 1 displays the item-level completion rate in a scatterplot, where all items, except for item 12, were located above the diagonal line (the identical line where the computer-based and paper-based tests have the same item-level completion rate).

As seen from Figure 1, the difference in the completion rate appears to be greater for items located toward the end of the assessment. For example, the first eight items had an average completion rate of 100% for the paper-based test takers and 99% for the computer-based test

Table 6. Test completion and item skipping across modes.

	Computer	Paper	χ^2	p value
75% Test Completion	.92	.97	5.482	.019
100% Test Completion	.80	.87	5.899	.015

**Figure 1.** Scatter plot of item completion rate under computer-based test and paper-based test conditions.

Each dot in this plot represents one item, with the horizontal axis representing the completion rate in the computer-based test mode and the vertical axis representing the completion rate in the paper-based test mode. The diagonal line is the identical line, where it means any items in that line would have equal completion rate between the two modes, a dot above the diagonal line mean that the item's completion rate is higher under paper-based test than that under computer-based test.

takers, with an average difference of 1%. The middle 10 items had average completion rates of 99% for paper-based test and 96% for computer-based test, with an average difference of 3%.

The last eight items observed average completion rates of 94% and 90% for paper-based and computer-based test respectively, with an average difference of 4%. For example, item #26, the last item, had completion rates of 86% and 93% for computer-based and paper-based test, respectively, with a difference of 7%. Item #25 also observed a difference of 5% (93% versus 88% for paper-based and computer-based tests). It seems possible that some computer-based test-takers ran out of time and, therefore, were not able to reach items toward the end of the assessment.

To our surprise, Item #14, which is located in the middle of the test, also had a relatively large difference in completion rate of 6%, 99% for paper-based and 93% for computer-based test. Item #14 was a relatively difficult item, with a p -value of .25 in computer-based test and .24 in

Table 7. User perceptions across modes.

		Computer	Paper	χ^2	p value
Test Difficulty	Too easy	0.04	0.06	0.652	0.722
	At the right level	0.87	0.85		
	Too difficult	0.08	0.09		
Testing Time	More than enough time	0.26	0.16	8.047	0.018
	Enough time	0.54	0.59		
	Not enough time	0.20	0.25		

paper-based test. The computer-based test-takers needed to make an inline choice by clicking on a sentence in a paragraph presented in the stimulus. In the paper-based test, the paragraph was “disassembled” in that all sentences in the paragraph were arranged in a list for the test-takers to select. While it could be the case that the format in the paper-based test for this item may have reduced the cognitive load to some degree compared with the computer-based test format, such an explanation could also be applicable to other items with similar changes in terms of format. However, such differences were not observed in other items that received similar adaptation.

An alternate explanation could be that the differences in Item #14 could have been related to speededness. According to Lord (1980) and Bejar (1985), test items that are omitted (e.g. items in the middle of the test) or not reached (e.g. items toward the end of the test) are both possible results of a timed test, where test takers, even though may have the power and ability to solve them correctly, do not have enough time to answer the items or cannot complete all items. As item #14 was a difficult item, students might be equally likely to skip this item initially in both mode conditions. However, because test takers may need more time to complete the computer-based test, it might have been more likely for the paper-based test-takers to return and answer this item.

Test-taker perceptions

Table 7 displays the proportions of test-takers who rated the test as “too easy”, “at the right level”, and “too difficult”, as well as the proportions of those who reported having “more than enough time”, “enough time” and “not enough time” to complete the test. In general, most students (87% in computer-based and 85% in paper-based test) reported the test had the right difficulty level, and 8-9% of students reported feeling the test was too difficult. A χ^2 test suggested there were no association between perceived test difficulty and the two mode conditions, $\chi^2 = .652, p = .722$.

There was a significant difference across modes regarding users’ perception of test taking time, $\chi^2 = 8.047, p < 0.05$. More computer-based test-takers (26%) thought that there was more than enough testing time than paper-based test-takers (16%), while a lower proportion of computer-based test-takers reported not having enough time (20% versus 25%).

Discussions and implication

In this study we examined the mode effects in assessing critical thinking skills, a key learning outcome in higher education, both domestically and in an international context. The assessment was originally developed as a computer-based test delivered in English and was later translated and adapted into a Chinese version that was delivered under a computer-based and a paper-based mode condition in Mainland China. We took a comprehensive approach to investigate computer versus paper-based mode effects from multiple perspectives, including mean scores, measurement precision, item functioning, response behaviors and user perceptions.

At the test level, we found that paper-based test-takers scored significantly higher on average than computer-based test-takers. We also found that test completion rates were higher for paper-based than for computer-based test. We suspect that the following three factors could have contributed to these observed differences.

One factor could have been the use of item sets in the assessment. Schwarz, Rich, and Podrabsky (2003) argued that items that are locally dependent are more likely to give rise to mode effects. The majority of the *HElghten* CT items were in item sets and, therefore, were likely to be locally dependent within a set. It is presumably easier to navigate through, respond to, and review items associated with the same stimulus input on paper than on computer because of the way this assessment was administered: multiple items could be presented on a single printed page under the paper-based test mode, whereas in the computer-based test items were presented on screen one at a time. This, in turn, could have given the paper-based test test-takers an advantage over those taking the computer test, resulting in better performance and higher test completion. To reduce the likelihood of putting computer-based test test-takers at a disadvantage when taking assessments with mostly set items, we think it could be helpful to provide textual or visual aids to help orient test-takers to the interconnection among items within a set. For example, an overview of the assessment structure could be shown at the beginning of the test, including information such as item ordering, location of the set, number of items associated with each set, and so forth. The computer-based test interface can also be designed in a way that items of the same set can be accessed from a common screen.

The observed differences between modes could have also been attributed to the reading requirement of the assessment. It has been argued that mode effects are associated with the cognitive workload introduced by screen navigation required for passage-based items (Kim and Huynh 2008; Yu 2010). In our case, a substantial amount of reading was required for answering most of the items on the assessment. For these items under the computer-based test condition scrolling became necessary where the reading content could not be viewed in its entirety on one screen, which could have complicated the response process, resulting in increased cognitive load experienced by computer-based test-takers. In comparison, reading the whole of long passages on paper was a considerably more straightforward and less cognitively demanding process.

Related to the scrolling requirement, Pommerich (2004) contended that compared to a computer-based test, it is easier for paper test-takers to locate relevant information given in a passage because the passage occurs in a fixed position on the page. This could also have allowed the paper test-takers to score higher and finish the assessment faster. To minimize mode incomparability for assessments composed of items with long reading passages, we think that test developers could offer tools to assist test-takers in navigating reading. For example, paging, instead of scrolling, can be made available for navigating very long stimulus input. A highlighting tool that lets test-takers highlight each line as they read could also be helpful. In addition, being able to enlarge or maximize the passage display window could be a useful tool.

The third potential cause for the differential performance between modes could have been related to test-taking flexibility. Under the paper-based test mode, test-takers could employ a variety of test-taking strategies, such as highlighting important information in the stimuli, marking a response for later review, marking the response status, crossing out the option considered to be incorrect, skipping items and answering them later, reviewing and revising item responses, and so forth. Past research has found that flexible navigation features contribute to positive user perceptions (Bridgeman, Lennon, and Jackenthal 2003; Arce-Ferrer and Guzman 2009) and better performance (Wise and Plake 1989). The computer interface used in this study was designed with features that allowed test-takers to use many test-taking strategies commonly available in paper testing. Nevertheless, the paper test offered more flexibility than the computer test. For example, in the paper-based test students could move freely among items while in the computer-based test students had to use the navigation arrows or the review screen to reach a

certain item. We therefore suspect that the paper-based test allowed participants to use more test-taking strategies and to use them more effectively, resulting in the observed performance differences between modes. To enable computer test-takers to utilize test-taking strategies commonly available in paper testing, the current computer interface could be enhanced with additional navigational and editing functions. For example, editing tools (e.g. highlight, underline, strikethrough) can be offered to allow computer test-takers to mark up the stimulus input.

Finally, that most Chinese K-12 or college level tests are delivered via paper may suggest that most students may be more used to paper-based tests, and less familiar with computer-based tests. While mode-related familiarity has been found to affect test anxiety and prevent students from demonstrating their abilities or skills targeted by the test (Goldberg and Pedulla 2002), this may have a deeper relation with the mode on which students' daily learning activities are happening. For example, it seems very likely students use paper-and-pencil format more frequently than a computer in a typical Chinese college or K-12 setting.

An investigation was also carried out at the item level to identify item characteristics that could amplify the effects of mode. We found that most of the items appeared to be easier on paper than on computer, which was expected given the direction of the differences in mean scores. In particular, Item #4 and Item #9 showed the greatest differences in item difficulty between modes. A variety of item characteristics were suspected to have made these items more cognitively demanding under the computer-based test mode. In the case of Item #4, performance difference between modes could have been related to the scrolling requirement and the "select all that apply" response type. For Item #9, the targeted cognitive skill, connecting two pieces of information in the stimulus could have made it easier to find the answer on paper than on computer. We also found that Item #14, a relatively difficult item located close to the middle of the test, was more likely to be skipped by computer test-takers. We observed that a paragraph in the computer-based test mode was converted into a list of sentences in the paper-based test mode, which could have made this item appear to be less mentally taxing for paper test-takers, and therefore resulted in a much lower skipping rate.

Past literature has been mainly focused on item display characteristics (e.g. scrolling, graphic display, number of items per page) when exploring potential causes of mode effects. Our findings suggested that factors other than item display, such as the targeted cognitive skill and response type, could also be related to mode effects. Furthermore, our analysis of Item #14 highlighted the need to take a light touch in adaptation between modes. In this case converting the paragraph into a list of sentences was not essential for answering this item and could have made it more prone to mode effects. Further explorations may benefit future practices when creating more comparable versions of the same test items between paper-based and computer-based tests.

Limitations and future directions

A few study limitations must be pointed out. First, our analyses were based on a relatively small sample size in China only, which can be used to inform future practices and research around this topic in a Chinese setting but also limits the extent to which the findings could be generalized to other countries. Small sample size also prohibited us from using latent variable modeling approaches to examine invariances across groups. Buerger and Goldhammer (2016) argued that as the focus of large-scale assessments is often to compare means across populations, it is imperative to examine construct equivalence - that is, whether the test captures the same latent variables in both modes - and to compare means at the latent level by accounting for measurement errors. If sample size permits, future studies should employ latent variable modeling approaches to inform score comparisons across modes.

Second, data on a key test-taker background characteristic, computer use experience or familiarity, were not available in this study, precluding us from controlling for this when comparing performances across modes. Computer use experience and familiarity has been found to have an impact on online test scores (Goldberg and Pedulla 2002; Bennett et al. 2008). Future studies would benefit from collecting information on participants' experiences and familiarity with computer use to determine the extent to which mode effects, if any, could be attributed to this user characteristic. Knowing this can also inform the development of tutorial or training materials to potentially reduce mode differences.

A third study limitation was the lack of external criterion measures. As such, we were not able to examine the differences in the test's relationship to external variables. As argued by Clariana and Wallace (2002), criterion-related validity evidence is critical for determining which mode most accurately reveals the targeted construct. Examining criterion-related evidence should be included in investigations of mode effects in the future.

Taken together, our findings suggest that when international studies are implemented with the aim to produce comparable test scores, one must factor in technological differences in test administration. If multiple test administration methods or testing modes are required, empirical evidence needs to be examined to detect possible impacts of test modes on test-taker performance. Findings from this study also shed light on assessment and item properties that could be sources of mode effects for assessments delivered in Chinese. Our actionable design suggestions will hopefully help minimize mode incomparability in future testing efforts that require the use of multiple delivery modes in Chinese language and other languages.

Acknowledgements

Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dr. Lin Gu is a research scientist at Educational Testing Service (ETS). She has broad training in language pedagogy, language testing, and measurement. Her research focuses on designing and developing AI-powered language learning solutions in digital environments. Lin also has expertise in assessing and improving learning outcomes in the international space.

Dr. Guangming Ling is a managing senior research scientist at the Center for Education and Career Development, Educational Testing Service. Dr. Ling's research focuses on assessment and training of core competencies, student learning outcomes assessments, and factors related to test validity, reliability, and fairness issues. Dr. Ling earned his PhD in Quantitative Psychology from Fordham University.

Dr. Ou Lydia Liu is Principal Research Director in charge of three ETS research centers: Center for Education and Career Development, Center for Language Education and Assessment Research, and Center for K12 Teaching, Learning, and Assessment. Dr. Liu is an internationally recognized expert in assessment and research of critical competencies in higher education. Dr. Liu holds a doctorate in Quantitative Methods and Evaluation from the University of California, Berkeley.

Dr. Zhitong Yang is a project manager in Educational Testing Service. His research interests focus on the assessment of higher-order thinking, problem-solving, and designing innovative computer-based assessments to measure traditionally hard-to-measure cognitive and noncognitive constructs through both process and outcome data.

Dr. Guirong Li is a professor in the School of Education at Henan University. Her research focuses on the policy evaluation, policy experiment and efficient allocation of education resources. She is the leader of the International

Center for Action Research on Education (ICARE) and conducts large-scale empirical research in central China to improve education policy and reduce the education inequality.

Elena Kardanova is a Tenured Professor and a Director of the Center for Psychometrics and Educational Measurement at the Institute of Education at National Research University Higher School of Economics, Moscow. Her research focuses on psychometrics, assessing complex constructs, large scale assessment, measuring individual progress in academic achievements, cross-national comparability of test results. She also is a scientific supervisor of master program “Measurement in Psychology and Education” at HSE.

Dr. Prashant Loyalka is an Associate Professor at the Graduate School of Education and a Senior Fellow at the Freeman Spogli Institute for International Studies at Stanford University. His research focuses on examining/addressing inequalities in the education of children and youth and on understanding/improving the quality of education received by children and youth in multiple countries including China, India, Russia, and the United States. He also conducts large-scale evaluations of educational programs and policies that seek to improve student outcomes.

ORCID

Lin Gu  <http://orcid.org/0000-0001-7283-0749>

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.
- Arce-Ferrer, A. J., and E. M. Guzman. 2009. “Studying the Equivalence of Computer-Delivered and Paper-Based Administrations of the Raven Standard Progressive Matrices Test.” *Educational and Psychological Measurement* 69 (5): 855–867. doi:[10.1177/0013164409332219](https://doi.org/10.1177/0013164409332219).
- Bejar, I. I. 1985. “Test Speededness under Number-Right Scoring: An Analysis of the Test of English as a Foreign Language.” *Ets Research Report Series* 1985 (1): i–57. (doi:[10.1002/j.2330-8516.1985.tb00096.x](https://doi.org/10.1002/j.2330-8516.1985.tb00096.x)).
- Bennett, R. E., Braswell, J. Oranje, A. Sandene, B. Kaplan, B. Yan. F. 2008. “Does It Matter If I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP.” *Journal of Technology, Learning, and Assessment* 6 (9), 1–35.
- Bridgeman, B., M. L. Lennon, and A. Jackenthal. 2003. “Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance.” *Applied Measurement in Education* 16 (3): 191–205. doi:[10.1207/S15324818AME1603_2](https://doi.org/10.1207/S15324818AME1603_2).
- Brunfaut, T., L. Harding, and A. O. Batty. 2018. “Going Online: The Effect of Mode of Delivery on Performances and Perceptions on an English L2 Writing Test Suite.” *Assessing Writing* 36: 3–18. doi:[10.1016/j.asw.2018.02.003](https://doi.org/10.1016/j.asw.2018.02.003).
- Buerger, S., and F. Goldhammer. 2016. “The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes.” *Psychological Test and Assessment Modeling* 58 (4): 597–616.
- Clariana, R., and P. Wallace. 2002. “Paper-Based versus Computer-Based Assessment: Key Factors Associated with the Test Mode Effect.” *British Journal of Educational Technology* 33 (5): 593–602. doi:[10.1111/1467-8535.00294](https://doi.org/10.1111/1467-8535.00294).
- Coates, H., and S. Richardson. 2012. “An International Assessment of Bachelor’s Degree Graduates’ Learning Outcomes.” *Higher Education Management and Policy* 23 (3): 1–19. doi:[10.1787/hemp-23-5k9h5xkx575c](https://doi.org/10.1787/hemp-23-5k9h5xkx575c).
- Gallagher, A., R. E. Bennett, C. Cahalan, and D. A. Rock. 2002. “Validity and Fairness in Technology-Based Assessment: Detecting Construct-Irrelevant Variance in an Open-Ended, Computerized Mathematics Task.” *Educational Assessment* 8 (1): 27–41. doi:[10.1207/S15326977EA0801_02](https://doi.org/10.1207/S15326977EA0801_02).
- Gneezy, U., J. A. List, J. A. Livingston, S. Sadoff, X. Qin, and Y. Xu. 2017. *Measuring Success in Education: The Role of Effort on The Test Itself* (No. w24004). Cambridge, MA: National Bureau of Economic Research. <http://www.nber.org/papers/w24004.pdf>
- Goldberg, A. L., and J. J. Pedulla. 2002. “Performance Differences according to Test Mode and Computer Familiarity on a Practice Graduate Record Exam.” *Educational and Psychological Measurement* 62 (6): 1053–1067. doi:[10.1177/0013164402238092](https://doi.org/10.1177/0013164402238092).
- Gray, L., N. Thomas, and L. Lewis. 2010. *Teachers’ Use of Educational Technology in U. S., Public Schools: 2009* (NCES 2010-040). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education.
- International Test Commission. 2017. The ITC Guidelines for Translating and Adapting Tests (Second edition). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

- Kim, D.-H., and H. Huynh. 2008. "Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test." *Educational and Psychological Measurement* 68 (4): 554–570. doi:[10.1177/0013164407310132](https://doi.org/10.1177/0013164407310132).
- Klein, S., R. Benjamin, R. Shavelson, and R. Bolus. 2007. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31 (5): 415–439. doi:[10.1177/0193841X07303318](https://doi.org/10.1177/0193841X07303318).
- Kolen, M. J., and R. L. Brennan. 1995. *Test Equating: Methods and Practices*. New York: Springer-Verlag.
- Kroehne, U., T. Gnamb, and F. Goldhammer. 2019. "Disentangling Setting and Mode Effects for Online Competence Assessment." In *Education as a Lifelong Process*. 2nd ed., edited by H.-P. Blossfeld and H.-G. Roßbach, 171–193. Wiesbaden, Germany: Springer VS.
- Liu, O. L., L. Mao, L. Frankel, and J. Xu. 2016. "Assessing Critical Thinking in Higher Education: The HEIghten™ Approach and Preliminary Evidence." *Assessment & Evaluation in Higher Education* 41: 677–694.
- Liu, O. L., J. A. Rios, and V. Borden. 2015. "The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment." *Educational Assessment* 20 (2): 79–94. doi:[10.1080/10627197.2015.1028618](https://doi.org/10.1080/10627197.2015.1028618).
- Liu, O. L., A. Shaw, L. Gu, G. Li, S. Hu, N. Yu, L. Ma, et al. (2018). "Assessing College Critical Thinking: Preliminary Results from the Chinese HEIghten® Critical Thinking Assessment." *Higher Education Research & Development* 37 (5): 999–1014.
- Liu, O. L., L. Frankel, and K. C. Roohr. 2014. *Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment (ETS RR-14-10)*. Princeton, NJ: Educational Testing Service. doi:[10.1002/ets2.12009](https://doi.org/10.1002/ets2.12009).
- Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Organization for Economic Co-operation and Development. 2012. *PISA 2012 Technical Report*. Paris, France: OECD. https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf.
- Organization for Economic Co-operation and Development. 2013. *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris, France: OECD. <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>.
- Organization for Economic Co-operation and Development. 2016. *The PISA 2015 Field Trial Mode-Effect Study*. Paris, France: OECD Publishing. www.oecd.org/pisa/data/PISA-2015-Vol1-Annex-A6-PISA-2015-Field-Trial-Mode-Effect-Analysis.pdf.
- Poggio, J., D. R. Glasnapp, X. Yang, and A. J. Poggio. 2005. "A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large-Scale State Assessment Program." *The Journal of Technology, Learning and Assessment* 3 (6):1–32.
- 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):1–45.
- Prisacari, A. P., and J. Danielson. 2017. "Rethinking Testing Mode: Should I Offer my Next Chemistry Test on Paper or Computer?" *Computers & Education* 106: 1–12.
- Richardson, S., and H. Coates. 2014. "Essential Foundations for Establishing Equivalence in Cross-National Higher Education Assessment." *Higher Education* 68 (6): 825–836. doi:[10.1007/s10734-014-9746-9](https://doi.org/10.1007/s10734-014-9746-9).
- Schwarz, R., C. Rich, and R. Podrabsky. 2003. "A DIF Analysis of Item-Level Mode Effects for Computerized and Paper-and-Pencil Tests." Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL, April 22–24.
- Tremblay, K., D. Lalancette, and D. Roseveare. 2012. *Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study Report: Design and Implementation*. Vol. 1. Paris, France: Organization for Economic Co-operation and Development.
- Wang, S., H. Jiao, M. J. Young, T. Brooks, and J. Olson. 2007. "A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests." *Educational and Psychological Measurement* 67 (2): 219–238.
- Wise, S. L., and B. S. Plake. 1989. "Research on the Effects of Administering Tests via Computers." *Educational Measurement: Issues and Practice* 8 (3): 5–10. doi:[10.1111/j.1745-3992.1989.tb00324.x](https://doi.org/10.1111/j.1745-3992.1989.tb00324.x).
- Yu, G. 2010. "Effects of Presentation Mode and Computer Familiarity on Summarization of Extended Texts." *Language Assessment Quarterly* 7 (2): 119–136. doi:[10.1080/15434300903452355](https://doi.org/10.1080/15434300903452355).