The role of test-mode effect: Implications for assessment practices and item design

Tom Lowrie, University of Canberra, Australia
Tracy Logan, University of Canberra, Australia

Introduction

It would seem reasonable to suggest that a duplicated static representation would require precisely the same reasoning, irrespective of the platform it is presented on. However, there is burgeoning empirical evidence to suggest that "identical" paper-based and digital-based tests will not obtain the same results (Clariana & Wallace, 2002).

In recent years, computer-based testing (CBT) has grown in acceptance and is likely to become the primary mode for delivering tests in the future. Benefits of digital-based testing include reduced costs and much faster feedback, and the capacity to tailor the test to an individual's needs. In fact, some of the most influential mathematics tests have moved toward an online, digital, mode (e.g., The Program for International Student Assessment [PISA] will have a CBT component in 2016). Elsewhere, the computerization of the National Assessment of Educational Progress (NAEP) is considered inevitable despite the acknowledgement that this will have an effect on the results of at least some subgroups of the population (Thissen & Norton, 2013). Although such testing has advantages over the Pencil-and-Paper Testing (PPT) mode; assessment experts, researchers and practitioners have raised concerns about comparability, transferability, and equity.

Performance across CBT and PPT modes

Although the prominence and attention afforded to CBT has increased in recent years, studies focused on test-mode effect have been researched for more than twenty years (Bugbee, 1996; Clariana & Wallace, 2002; DeAngelis, 2000). With continued advances in technology, it is unsurprising that CBT has become the preferred method of assessment, with inevitable comparisons made to the more traditional PPT (Wang, Jiao, Young, Brooks, & Olson, 2008). Nevertheless, test developers have an obligation to show the equivalence between these modes of delivery (Bugbee, 1996). Many earlier studies were concerned with this equivalence issue because of the ongoing desire to ensure the validity of score interpretations over time (Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008). In mathematics education, this is especially desirable given the longitudinal international data produced by studies such as PISA and Trends in International Mathematics and Science Study (TIMSS).

Research investigating the difference between pencil-and-paper and computer-based tests (i.e., test mode effect) has yielded different results. For example, some studies have reported differences in students' performance (Bennett et al., 2008). By contrast, a comprehensive meta-analysis involving 44 independent experiments revealed no test-mode effect (Wang et al., 2008) on measures such as performance, test duration and students' computer familiarity. Similarly, Threlfall, Pool, Homer, and Swinnerton (2007) found minimal differences on overall test performance. However, there were substantial differences on students' performance across individual items. These findings are especially relevant to mathematics education assessment since there has been an

increased use of graphics and visual displays (Lowrie & Diezmann, 2009), due in part to new technological innovations (Schnotz & Lowe, 2003). As Hegarty, Canham and Fabrikant (2010) demonstrated, salience (i.e., the noticeable element in the display of information) has a large effect on performance in the interpretation of graphic tasks. In fact, they argued that the display design interacts with knowledge to influence the comprehension of visual displays. As mathematics assessment moves toward computer-based (and digital) representation of items, it is critical to determine the influence test mode has on graphic item performance.

A number of studies have concluded that there are minimal differences between overall performance across PPT and CBT modes (Johnson & Green 2006; Threlfall et al., 2007). However, such findings need to be treated with caution. In their study, Johnson and Green (2006) identified a number of differences at an individual task level across mode. These related to the type of question and how it was posed, the magnitude and quantity of numbers in a task and the students' willingness to show working out. They concluded that particular types of questions within defined content areas impacted differently on student performance according to the mode they were presented in. This is particularly problematic given the manner in which data are now reported since fine-grained analysis associated with performance on mathematics content areas is provided in a majority of national and international tests. The present study examines the notion of test equivalence within mathematics content strands across CBT and PPT modes.

The changing nature of assessment

From a psychometric viewpoint, there is strong desire to ensure that CBT formats are equivalent to that of the previous PPT mode. Typically, test designers are mindful of ensuring that the total mean score of one mode is equivalent to the other (Wang et al., 2008). In order to utilize longitudinal data sets, the transfer from PPT to CBT needs to be as seamless as possible. Although it may be the case that items within a test could vary considerably across mode, most testing agencies need to ensure that the entire instrument (a collection of items) is comparable across modes (Wang et al., 2008). Indeed, differences between one version of a test and another would suggest the possibility of different constructs being assessed. MacDonald (2002) identified that these differences cannot readily be determined through statistical methods alone. Noteworthy, when no differences are identified, there is an immediate assumption that the tests are equivalent.

The present investigation examines test mode effect on a specific class of mathematics assessment items; namely, graphics tasks. Graphics tasks are items that contain high concentrations of visual–spatial information, including graphs, maps and diagrams (Lowrie & Diezmann, 2011). Such tasks are becoming more common in mathematics assessments and tests due to advances in technology and the acknowledgement that decoding spatial information is critical for general numeracy and problem-solving development (Lowrie & Logan, 2007).

The following hypothesis was posed for the investigation: *That there would be no difference in student performance on graphics tasks across CBT and PPT modes.*

Method

Participants

This investigation examined 801 Grade 6 (11-12 year olds) Singaporean students' performance on mathematics tasks presented in computer-based (iPad) and pencil-and-paper modes. The analysis involved the six graphics questions from a set of 12 questions sourced from two national tests. Approximately half the students completed the six graphics questions in a pencil-and-paper mode (n=404) with the remainder solving the questions on an iPad (n=397). From visual and salient perspectives, we attempted to ensure that the questions were represented almost identically in both modes.

Procedure and instrument: Mathematics Processing Instrument (MPI)

In the present study we focused on six graphics problems from the Mathematics Processing Instrument (MPI) (see Appendix). The tasks were sourced from the Australian National Assessment Program – Literacy and Numeracy (NAPLAN) and the Singaporean Primary School Leaving Examination (PSLE). The design of the instrument was based on Suwarsono's MPI (described in Lean & Clements, 1981; Lowrie & Kay, 2001) for measuring the visualizer-verbalizer cognitive style and subsequently adapted elsewhere in the literature (e.g., Kozhevnikov, Hegarty, & Mayer, 2002).

The MPI was administered in Singapore schools in March-April 2013. Within each Grade 6 class, half the students solved the six graphic tasks in a pencil-and-paper mode while the other half completed the task in a computer-based (iPad) mode. The mode was assigned randomly. The Instrument consisted of two parts: (a) a mathematics test; and (b) a corresponding questionnaire that encouraged students to describe the strategies and approach they employed to solve the tasks in the test. The mathematics test (Part A) was used to address the hypothesis. Part B (the questionnaire), was used in subsequent analysis to better understand *how* the students solved the respective tasks. Student solutions were classified as either (1) analytic or (2) non-analytic. An analytic solution was defined as an approach where a graphic or image was not used to solve the task. Typically, analytic solutions involved students generating computations or number sentences to solve the task. By contrast, a non-analytic solution involved the use of a graphic or image to solve the task, either mentally or in a concrete form.

Data from Part A of the MPI were classified as correct or incorrect. These data were used to address the study's hypothesis. Solution strategies from Part B were classified as analytic or non-analytic processing. These data were used to determine whether or not students employed different strategies to solve the respective tasks, across the two test-mode effects.

Results and Findings

Hypothesis: That there would be no difference in student performance on graphics tasks across CBT and PPT modes.

An Analysis of Variance (ANOVA) was used to test our hypothesis that performance would be the same irrespective of mode. An ANOVA revealed statistically significant differences between student performance by mode (i.e., pencil-and-paper vs iPad); [$F(1,800)=41.9$, $p<.001$]. Subsequent univariate analyses revealed statistically significant differences (with p values <.008, adjusted for the Bonferroni correction

method) between the two modes on each of the six questions. In each instance, the proportion of success was higher for those students who completed the respective questions in the pencil-and-paper mode. Hence, these graphic questions (which included maps, tables, and diagrams) were more difficult to solve when presented digitally (see Table 1 for means, standard deviations and t-test scores by test-mode effect).

Table 1. Means (and standard deviations) for graphics tasks by test-mode effect

Task	Means (and Standard Deviations)		T-tests (Effect Size)
	PPT	CBT (iPad)	
1	.82 (.38)	.68 (.47)	4.73* (*d*=.33)
2	.88 (.32)	.81(.39)	2.84* (*d*=.20)
3	.81 (.39)	.73 (.44)	2.69* (*d*=.19)
4	.75 (.44)	.59 (.49)	4.69* (*d*=.34)
5	.78 (.41)	.67 (.47)	3.36* (*d*=.25)
6	.59 (.49)	.39 (.49)	4.40* (*d*=.41)

Note: * = probability value p<.008

Given the fact that there were substantial differences in mean scores for each of the six graphics items across the PPT and CBT modes, it is apparent that the students found the tasks more difficult to solve in the CBT mode. Effect sizes for each question ranged from low to moderate. Further analysis was required to establish why performance differences were consistently in favor of students who completed the tasks in a pencil-and-paper form. Since there were dramatic differences in performance across the test modes (as evidenced in Table 1), differences in strategy selection may have been a determining factor in task success.

Subsequent Hypothesis: That there would be no relationship between student solution methods on graphics tasks and test modes.

Cross-tab and Chi-square analysis were undertaken to establish whether there were relationships between strategies employed by the students and test mode. The distribution of strategies used by the students on pencil-and-paper and iPad for the respective graphics tasks are presented (as frequencies) in Table 2. For four of the six graphics tasks, there were statistically significant relationships between the strategies students employed to solve the tasks and the test mode. For two of these items (Questions 1 and 6), students were more likely to use non-analytic strategies to solve the items in the pencil-and-paper mode. By contrast, analytic strategies were more frequently employed to solve Questions 3 and 4 in pencil-and-paper mode. In each of these cases, the opposite is the case in the computer-based (iPad) mode; that is, more analytic strategies (for Questions 1 and 6) and non-analytic strategies (for Questions 3 and 4). Nevertheless, for two of the items (Tasks 2 and 5) there were no relationships between students' strategy use and test mode. For these two tasks, the use of a particular strategy was not influenced by the mode of presentation.

Conclusions and implications

This investigation has revealed distinct differences in student performance on graphics-based mathematics tasks when presented across PPT and CBT modes. The cohort of students who completed the tasks in the PPT mode had significantly higher mean scores across all six of the graphics items. These findings complement those of Threlfall et al., (2007) who suggested that the mode of testing effects students' performance on

subgroups within a whole test. In our study, subsequent analysis failed to highlight distinct patterns of difference between students' strategy use across the two test modes. Differences in performance were certainly evident and for four of the six items there was a significant relationship between strategy use and test mode. Nevertheless, these patterns were erratic. Consequently, these performance differences were influenced by other factors.

Table 2. Cross Tabs and Chi-Square Analysis for Strategy Use by Test-Mode Effect

Task	Mode	Approach		Chi-Square
		Analytic	Non-Analytic	
1	PPT	74	311	$\chi^2(1) = 3.45, p < .05^*$
	CBT	105	257	
2	PPT	63	331	$\chi^2(1) = 1.42\ p > .05$
	CBT	74	311	
3	PPT	364	34	$\chi^2(1) = 3.45, p < .05^*$
	CBT	332	48	
4	PPT	187	207	$\chi^2(1) = 7.24, p < .01^*$
	CBT	146	239	
5	PPT	383	11	$\chi^2(1) = .82\ p > .05$
	CBT	379	7	
6	PPT	39	343	$\chi^2(1) = 12.22, p < .001^*$
	CBT	72	303	

Note: Not all students provided a response to Part B of the MPI, therefore, totals may not equal the total number of participants.

Some studies have found that the cognitive resources required to interact with computer-based tools may impede or interfere with the natural ability to interpret or decode the problem. The attention focus of moving from the screen to either process information mentally or represent it elsewhere (e.g., on working out paper) may interfere with sense making. That is, the cognitive demands of transferring information from one device to another may lead to increases in cognitive load. For example, Chandler and Sweller (1996) demonstrated that the act of using a computer could interfere with learning, since the problem solver was distracted by features that would otherwise be redundant in pencil-and-paper form. Their study highlighted the fact that processing information in different "spaces" could be distracting especially if the demands of the task were already complex.

In the current study, the participants preferred to use non-analytic strategies for most of their processing and such processing was much more effective in the pencil-and-paper environment. Perhaps when presented with the tasks in CBT mode, the participants found it more challenging to draw diagrams and encode information—such that using the iPads resulted in higher cognitive demands. The results of the study have wide reaching implications for test developers, assessment experts and classroom practitioners.

References

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. *The Journal of Technology, Learning and Assessment*, *6*(9). Retrieved March 4 2014 from http://www.jtla.org.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*(3), 282-290.

Chandler, P. & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*(2), 151-170.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602.

DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health*, *29*(3), 161-164.

Hegarty, M., Canham, M.S., & Fabrikant, S.I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 37-53.

Johnson, M. & Green, S. (2006). On-Line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, and Assessment*, *4*(5). Retrieved September 25 from http://www.jtla.org

Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (2002). Revising the visualizer-verbalizer dimension: Evidence for two types of visualizers. *Cognition and Instruction, 20*(1), 47–77.

Lean, G. & Clements, M. A. (1981). Spatial ability, visual imagery, and mathematical performance. *Educational Studies in Mathematics, 12*(3), 267–299.

Lowrie, T. & Diezmann, C.M. (2009). National numeracy tests: A graphic tells a thousand words. *Australian Journal of Education*, *53*(2), 141-158.

Lowrie, T. & Diezmann, C. M. (2011). Solving graphics tasks: Gender differences in middle-school students. *Learning and Instruction, 21*(1), 109-125.

Lowrie, T. & Kay, R. (2001). Relationship between visual and nonvisual solution methods and difficulty in elementary mathematics. *Journal of Educational Research, 94*(4), 248-255.

Lowrie, T. & Logan, T. (2007). Using spatial skills to interpret maps: Problem solving in realistic contexts. *Australian Primary Mathematics Classroom, 12*(4), 14-19.

MacDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education*, *39*, 299-312.

Schnotz, W. & Lowe, R. (2003). External and internal representations in multimedia learning. *Learning and Instruction*, *13*, 117-123.

Thissen, D. & Norton, S. (2013). *What might change in psychometric approaches to statewide testing mean for NAEP?* Report commissioned by the NAEP Validity Studies (NVS) Panel: University of North Carolina, Chapel Hill.

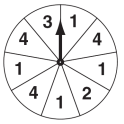
Threlfall, J., Pool, P., Homer, M. & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*, 335-348.

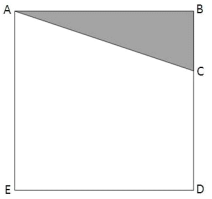
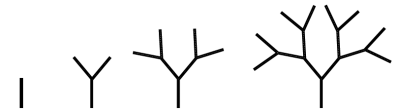
Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and pencil-and-paper testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*(1), 5-24.

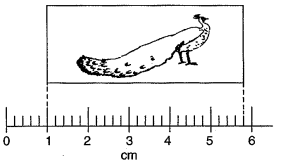
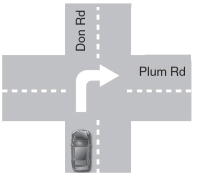
Tom Lowrie
University of Canberra, Australia
thomas.lowrie@canberra.edu.au

Tracy Logan
University of Canberra, Australia
tracy.logan@canberra.edu.au

Appendix. *The six graphics tasks.*

<table>
<tr><td>Task 1</td><td>The spinner is used in a board game.
 

Sanjay spins the arrow.
On which number is the arrow most likely to stop?</td></tr>
<tr><td>Task 2</td><td>In the figure below, ABDE is a square. The length of AB is 3 times the length of BC.
What fraction of ABDE is shaded?
(1) $\frac{1}{3}$
(2) $\frac{1}{4}$
(3) $\frac{1}{6}$
(4) $\frac{1}{8}$
 
</td></tr>
<tr><td>Task 3</td><td>The table shows the rate of charges for each overdue book borrowed from a library.
<table><tr><td>For the first 7 days</td><td>20¢ per day</td></tr><tr><td>After 7 days</td><td>50¢ per day</td></tr></table>Peter borrowed a book from the library. The book was overdue when he returned it.
He paid a total of $3.90 for the overdue book. How many days was it overdue?</td></tr>
<tr><td>Task 4</td><td>Lucy made 4 tree designs using sticks.
There is a pattern in the way the trees grow.
 

Lucy continues the pattern in the same way.
How many sticks will Tree 5 have?</td></tr>
<tr><td>Task 5</td><td>What is the length of the sticker as shown in the figure below?
 
</td></tr>
<tr><td>Task 6</td><td>A car is travelling north-east along Don Road.
The car is about to turn right into Plum Road.
 

In which direction will the car be travelling
after it turns right?</td></tr>
</table>