LOGAN, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal*, *27*(4), 423-441. https://doi.org/10.1007/s13394-015-0143-1

Math Ed Res J (2015) 27:423–441
DOI 10.1007/s13394-015-0143-1

ORIGINAL ARTICLE

The influence of test mode and visuospatial ability on mathematics assessment performance

Tracy Logan[1]

Received: 17 November 2014 / Revised: 17 February 2015 / Accepted: 16 March 2015 /
Published online: 29 April 2015

Abstract Mathematics assessment and testing are increasingly situated within digital environments with international tests moving to computer-based testing in the near future. This paper reports on a secondary data analysis which explored the influence the mode of assessment—computer-based (CBT) and pencil-and-paper based (PPT)—and visuospatial ability had on students' mathematics test performance. Data from 804 grade 6 Singaporean students were analysed using the knowledge discovery in data design. The results revealed statistically significant differences between performance on CBT and PPT test modes across content areas concerning whole number algebraic patterns and data and chance. However, there were no performance differences for content areas related to spatial arrangements geometric measurement or other number. There were also statistically significant differences in performance between those students who possess higher levels of visuospatial ability compared to those with lower levels across all six content areas. Implications include careful consideration for the comparability of CBT and PPT testing and the need for increased attention to the role of visuospatial reasoning in student's mathematics reasoning.

Keywords Secondary data analysis · Computer-based testing · Visuospatial ability · Mathematics assessment

Introduction

This investigation is framed within two salient topics within mathematics education assessment, namely the following: mode of assessment delivery and the influence of visuospatial reasoning on student performance. Indeed, the influence of test mode effect

✉ Tracy Logan
Tracy.Logan@canberra.edu.au

[1] Faculty of Education, Science, Technology and Mathematics, University of Canberra, Canberra, ACT, Australia

on student performance has come to the fore again recently as The Organisation for Economic Co-operation and Development (OECD) prepares to administer the Programme for International Student Assessment (PISA) as a fully computer-based test (CBT) for the first time, the long-term implications for which are relatively unknown. Despite many individual states within the USA undertaking CBT for their students, much of the research in this area has focused on test performance overall and the students' familiarity and competency within the computer environment.

For over two decades, attention has been afforded to student performance across different modes of test delivery, that is, computer-based testing (CBT) and pencil-and-paper testing (PPT) (Bugbee 1996; Clariana and Wallace 2002; DeAngelis 2000; Hardré et al. 2007). With continued advances in technology, CBT has become the preferred method of assessment, with inevitable comparisons made to the more traditional PPT (Wang et al. 2008). However, 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 et al. 2008). In mathematics education, this is especially desirable given the longitudinal international data produced by studies such as PISA and the move into CBT.

In tandem, there is a worldwide push to increase the number of young people engaging in science, technology, engineering and mathematics (STEM)-based subjects and careers (Marginson et al. 2013). Advanced knowledge of and skill with mathematics is commonly accepted as a prerequisite for most of these careers. Uttal and Cohen (2012) found that visuospatial ability is a strong predictor of students who will enter the STEM fields. Visuospatial reasoning—being able to understand spatial relationships—is considered to be an important aspect of mathematics performance in the sense that students with high visuospatial ability tend to do better on mathematics test than those students with lower visuospatial ability (Kyttälä 2008). These findings may be due to shared processes associated with the fact that mathematics tasks are represented in various spatial formats (Cheng and Mix 2014). These visuospatial demands are potentially different across content areas since assessment items have varying degrees of spatial information presented within a given task—with geometry and measurement content particularly reliant on spatially based representations.

This study was undertaken as a secondary data analysis and as such, the paper considers both topics in parallel due to the nature of the data set utilised.

Mode of assessment delivery

Mathematics tests and other forms of comparative data are increasingly moving from PPT to CBT (Wang et al. 2008). Reasons for this transition include the cost of producing materials, the capacity to provide results almost instantaneously and the flexibility for tailored test design. From a psychometric viewpoint, there is strong desire to ensure that CBT formats are equivalent to that of the previous PPT formats. In order to utilise 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.

A number of studies have concluded that there were minimal differences between overall performance across PPT and CBT modes; however, differences were identified at an item level. In their studies, Johnson and Green (2006) and Threfall, Homer and Swinnerton (2007) 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' ability and willingness to show working out. Indeed, Threfall et al. (2007) identified that in some cases, the cognitive load of the students increased when tasks were presented in CBT form. Sweller's (1994) cognitive load theory suggests that for some students, tasks are less cognitively demanding when all the elements of that task can be addressed sequentially as opposed to concurrently. The study suggested that even if this mode effect applies to only some students, it is still of concern for those attempting to assess mathematics knowledge fairly. Threfall et al. (2007) also concluded that particular types of questions within defined content areas impacted differently on student performance according to the mode in which they were presented. This is particularly problematic given the manner in which data are now reported. 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.

Visuospatial ability and mathematics

Within this paper, visuospatial ability is considered as the ability to visually perceive objects and understand the spatial relationships among those objects in both two- and three-dimensional space (Halpern and Collaer 2005; Tversky 2004). Elements of mathematics are inherently visuospatial in nature (Skemp 1986), so it is unsurprising that studies have found visuospatial ability to be a significant predictor of mathematics ability (Battista 1990; Casey et al. 1997; Reuhkala 2001; Rohde and Thompson 2007; Zhang et al. 2014) and strongly predicts students who will go into STEM fields as adults (Uttal and Cohen 2012). There is an even stronger case developed for the link between visuospatial ability and certain content areas within mathematics such as geometry (Clements 2004; Clements and Battista 1992), with Pittalis and Christou (2010) finding that three separate measures of spatial ability were influential in predicting performance on various 3-dimensional (3D), geometry tasks. More specifically, Kirby and Boulter (1999) found that for transformational geometry, "in which students learn to identify and illustrate movement of shapes in two and three dimensions" (p. 285), student's posttest geometry score was significantly predicted by their pretest spatial ability score, and stable individual differences were evident with those students who were "ahead at the outset had made by far the greatest progress by the end" of the treatment (p. 291).

Other studies have considered how aspects of visuospatial ability connect with various content areas and structural elements of representation within mathematics.

Tolar, Lederberg and Fletcher (2009) found moderate links between 3D visuospatial ability and algebra achievement in undergraduate college students. However, their study considered algebra "as symbol manipulation" only (p. 260), and did not consider other algebra topics such as pattern recognition, modelling and word-problem solving. They also suggested that for this type of algebra, computational fluency had a stronger influence on performance. Fuchs et al (2010) found that in younger children (approximately 5 years old), visuospatial ability did not predict procedural calculation or word-problem development. However, Lowrie and Diezmann (2007) identified that visuospatial ability accounted for over 20 % of the variance in performance across four categories of graphic mathematics tasks. These categories of graphics tasks were not aligned to specific content areas; however, all tasks required the decoding of graphic information which was essential in order to solve the task. Hence, there is evidence that an individual's visuospatial ability is closely linked with their overall mathematics ability and in particular with their geometry ability. What is less well known is whether visuospatial ability relates to more specific content areas within mathematics for students in the middle years of school. This is an important consideration as assessment frameworks are often based on content areas and hence performance data may not be reflective of content area knowledge, but different cognitive abilities.

This investigation utilised secondary data analysis techniques to analyse a pre-existing data set. "Secondary analysis uses data for purposes other than that for which they were originally collected" (Devine 2003, p. 287). As such, the nature in which the data were collected and recorded meant that multivariate approaches to data analysis were not feasible and separate examinations of mode of test delivery and visuospatial ability were deemed to be appropriate. Hence, the purpose of this paper was to identify if the mode of delivery and the level of students' visuospatial ability were influential in mathematics assessment task performance across content areas.

Research questions and hypotheses

Given that the two areas of mode of assessment and visuospatial ability are investigated in parallel, two research questions were developed:

- Does the mode of assessment (iPad vs pencil-and-paper) impact on student performance on tasks from different mathematics content areas?
- To what extent does a student's level of visuospatial ability influence their performance on mathematics tasks from different content areas?

In light of the research literature on the two areas, the following two hypotheses were generated:

- There will be no difference in performance on the iPad and pencil-and-paper modes across the content areas.
- Those students with high visuospatial ability will perform better than the moderate or low level students on the content areas usually associated with high spatial demands, such as geometry and measurement.

Methods and materials

Participants

The participants from the larger study were 807 grade 6 Singaporean students (aged 11–12 years) from 8 Singapore schools (six government and two government-aided). There were 392 boys and 415 girls in this sample. For this secondary analysis, the participants were a sub-sample of the larger cohort. For the Mathematics instrument, in order to have a sample where an even number of students undertook the items in both CBT and PPT, the number of cases analysed was reduced to 788. For the visuospatial instrument, a sample of 804 complete cases was analysed. With regard to sampling, students were randomly assigned to complete the mathematics instrument via CBT or PPT. All students completed the visuospatial instrument on pencil-and-paper. It is not clear from the original data set if the participants were familiar with the use of iPads or not.

Design

A secondary data analysis was undertaken using the knowledge discovery in data (KDD) design (Fayyad et al. 1996). The KDD design aims to extract useable knowledge from a collection of data and can be seen as "the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" (Fayyad et al. 1996, pp. 40-41). While traditionally KDD stems from fields of research such as artificial intelligence, machine learning and statistics, the process is interdisciplinary in nature and is applicable in many research contexts. Within the field of education, KDD has emerged as an interdisciplinary research area, with national data repositories (such as the National Centre for Educational Statistics in the USA) providing opportunities to work with data that was previously inaccessible. Indeed, data mining within education has gained momentum since 2008 after the establishment of a dedicated research conference and journal for educational data mining. Baker (2010) provides an overview of this field and highlights possible techniques and uses.

The KDD process is both interactive and generative and involves a series of sequential steps and corresponding decision making processes. Figure 1 illustrates the KDD process. According to Fayyad et al. (1996), there are five steps in the KDD process. (a) The *Selection* step involves selecting data from the larger database to create a target data set. The target data set is based on the goals of the project and the relevant prior knowledge of the data, i.e. focusing on a subset or a sample of data. (b) *Preprocessing* involves reducing the target data set to the useful features which represent the goals of the project, essentially cleaning and organising the data. Preprocessing requires the researcher to look at the data in a manner which allows them to make decisions about the exact nature of analysis. (c) *Transformation* of the data requires data reduction procedures through identifying a suitable analysis technique based on the goals of the project and the type of data being utilised. Data can be transformed through any analysis technique, with the aim to classify, cluster and summarise the data. (d) *Data mining* is seen as searching for and "determining patterns from observed data" (Fayyad et al. 1996, p. 43). This step can often involve a form of visual representation of the extracted patterns. (e) The *Interpretation*/*Evaluation* step

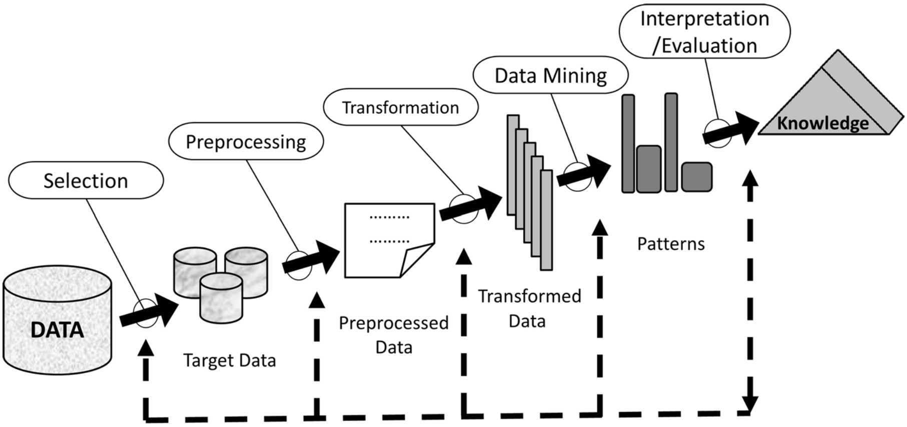


Fig. 1 An overview of the KDD process (Adapted from Fayyad et al. 1996, p. 41)

consists of interpreting any patterns and themes identified in the data mining step in relation to the project goals and evaluating their usefulness and potential interest to others. This paper followed these steps.

Selection of the data

Data for this secondary analysis was sourced from a larger research project, whose aim was to investigate the processing strategies (visual or non-visual) students' employed to solve mathematics tasks presented in digital and non-digital forms. The data for this larger project consisted of three sets of data, namely the following: (1) performance results on a 24-item test—where the test items were sourced from both Singaporean and Australian national tests and presented in both iPad and pencil-and-paper modes; (2) students' self-reported strategy use on those test items—where students were presented with a processing instrument to indicate the types of strategies they used when solving the test items; and (3) results from the cognitive test, the Paper Folding Test (Ekstrom et al. 1976)—completion of up to 20 items which provided a measure of students visuospatial ability. The target data for this paper was the performance results on the 24-item test in both iPad and pencil-and-paper form and the results from the Paper Folding Test.

Instruments

This study will focus on two of the instruments used in data collection: the 24-item mathematics test and the Paper Folding test. The mathematics test was developed by the Chief Investigator of the larger research project using national testing items from the mathematics components of the Singaporean grade 6 Primary School Leaving Examination (PSLE) and the Australian grades 5 and 7 National Assessment Program: Literacy and Numeracy (NAPLAN). Students were scored as correct or incorrect on these items with their results revealing that the minimum total score was 3 and the maximum score was 24.

For the Paper Folding instrument, students were required to visualise the folding action of a square sheet of paper. A hole is then punched in one part of the fold and students are to identify how the punched sheet would appear when fully reopened. An example of an item in this test is shown in Figure 2. The Paper Folding Test consisted of a set of 20 items, with participants given6 min to answer the set. According to the test protocol (Ekstrom et al. 1976), a correct item is given a score of 1 mark and the total score is calculated as follows: number of items marked correctly minus a fraction of the number marked incorrectly. Following from the scoring used by Mayer and Massa (2003), the total score for these participants was calculated as the number correct minus one-fifth the number incorrect in 6 min. The minimum score was −4, and the maximum score was 20 for these participants.

Preprocessing and transformation of the data

These steps require the data to be organised for analysis. Since the 24 items from the mathematics test were sourced from two separate national assessments and therefore developed under different curriculums, the more universal 2011 Trends in International Mathematics and Science Study (TIMSS) Mathematics Assessment Framework (Mullis et al. 2009) was used to classify items into content domains. Also, as the items were originally designed for grades 5, 6 and 7 students, the grade 8 TIMSS content areas were utilised, as the grade 4 TIMSS framework would not cover the types of content presented in the higher grades. The 24 items were independently classified under the content domains by five mathematics education researchers, with an interrater reliability analysis using the intraclass correlation (ICC) statistic (two-way mixed ANOVA) performed to determine consistency among raters. The interrater reliability for the researchers was found to be ICC=0.78 (p<0.0.001), 95 % confidence interval (CI) (0.656, 0.885), which is deemed to be a substantial level of agreement on item classification (Shrout and Fleiss 1979).

Data collation

For the paper folding test, the raw scores of the students were classified as low, high and moderate (Kozhevnikov et al. 1999). This data was transformed as follows: (i) low, bottom 25 % of the distribution; (ii) high, top 25 % of the distribution and (iii) moderate, the middle 50 % of the distribution. The purpose for the classification was to address the second research question and determine if a certain level of visuospatial ability influenced performance on certain content areas.

For the mathematics test, the items were collated under the 2011 TIMSS content domains and sub-domains (see Table 1). The items loaded heavily on the Number domain, and in particular, the Whole Number sub-domain. In order to have a more even spread of items across the content areas, items were collated into new categories based on the TIMSS sub-domains. For the Number domain, items assessing Whole Number remained in this category, while items under the Fractions and Decimals and Ratio,

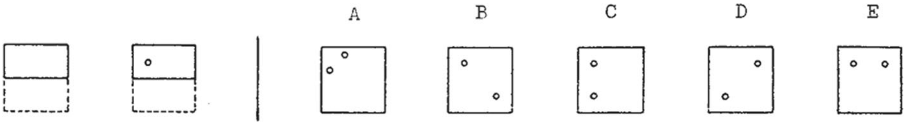


Fig. 2 Paper Folding Test (Reproduced with license and permission of Educational Testing Service, New Jersey, USA)

Table 1 2011 grade 8 TIMSS mathematics assessment framework content domains and new variables

Content Domain	Sub-domain	No. of items	New variables (n=No. of items)
Number	Whole Number	6	Whole Number (n=6)
	Fractions and Decimals	2	Other Number (n=5)
	Integers	0	
	Ratio, Proportion, and Percent	3	
Algebra	Patterns	4	Algebra Patterns (n=4)
	Algebraic Expressions	0	N/A
	Equations/Formulas and Functions	0	
Geometry	Geometric Measurement	5	Geometric Measurement (n=5)
	Geometric Shapes	1	
	Location and Movement	1	Spatial Arrangements (n=2)
Data and Chance	Data Organisation and Representation	1	Data and Chance (n=2)
	Data Interpretation	0	
	Chance	1	

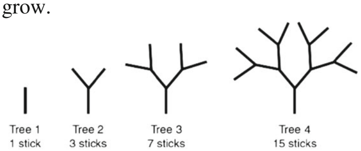
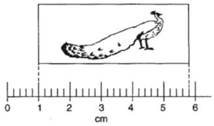
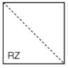
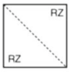
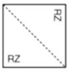
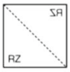
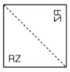
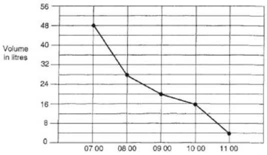
Proportion, and Percent sub-domains were collated under the new category Other Number. Within the Geometry domain, items assessing Geometric Measurement remained in this category and the items assessing Geometric Shape and Location and Movement were collated into a new category classified as Spatial Arrangements. Although there were only two items in the Spatial Arrangement category, it was classified on its own since the Geometric Measurement items all required a unit of measurement to be used while the Spatial Arrangement items related to the spatial relationships among objects, such as orientation and transformations. Within the Algebra domain, the two sub-domains of Algebraic Expressions and Equations/Formulas and Functions were not represented by any of the items present in the 24-item test and so only Algebra Patterns was used as a category (this was unsurprising given the grade levels the original items were aimed at). The Data and Chance category was collated given the small number of items present in all three sub-domains. Table 1 provides the content domains and sub-domains within the grade 8 TIMSS framework, the number of items which were classified under each domain, and the new categories created for this analysis. An example item from each of the six new categories is provided in Table 2. The creation of these new categories was based on the content of the items themselves in order to create a set of data that would be suitable for analysis for the purposes of this paper.

Data mining and interpretation/evaluation

These two steps consist of applying data analysis techniques and interpreting the results. Scores for these variables were generated by using percentages correct as means due to there being an unequal number of items in each category.

Since there were an unequal number of items contained in each content area variable, homogeneity of variance statistics were run in order to identify whether parametric tests could be performed. The results revealed that for the first independent variable of mode (iPad vs pencil-and-paper), the Levene's test for equality was

Table 2 Example items from the six content areas used in this study

Category	Example of performance test item
Whole Number	The chairs in a hall were arranged in rows. Each row had the same number of chairs. Weiming sat on one of the chairs. There were 5 chairs to his right and 5 chairs to his left. There were 7 rows of chairs in front of him and 7 rows of chairs behind him. How many chairs were there in the hall? © 2009 Singapore Examinations and Assessment Board
Other Number	Ben has 2 identical pizzas. He cuts one pizza equally into 4 large slices. He then cuts the other pizza equally into 8 small slices. A large slice weighs 32 grams more than a small slice. What is the mass of one whole pizza? © 2010 Australian Curriculum, Assessment and Reporting Authority
Algebra Patterns	Lucy made 4 tree designs using sticks. There is a pattern in the way the trees grow.  Tree 1 1 stick Tree 2 3 sticks Tree 3 7 sticks Tree 4 15 sticks Lucy continues the pattern in the same way. How many sticks will Tree 5 have? © 2010 Australian Curriculum, Assessment and Reporting Authority
Geometry Measurement	What is the length of the sticker as shown in the figure below?  © 2009 Singapore Examinations and Assessment Board
Spatial Arrangements	Ron paints these letters on a piece of paper.  While the paint is still wet, he folds the paper along the dotted line. When Ron unfolds the paper, what will it look like?     © 2010 Australian Curriculum, Assessment and Reporting Authority
Data and Chance	A tank was filled with 48ℓ of water at 07 00. Water flowed out the tank from 07 00 to 11 00. The line graph shows the amount of water in the tank from 07 00 to 11 00.  During which one-hour period was the decrease in the volume of water the greatest? © 2009 Singapore Examinations and Assessment Board

significant for three of the six content variables, and for the second independent variable of visuospatial ability (low, moderate or high), the Levene's test was significant for all six content variables. As such, homogeneity of variance cannot be assumed and non-parametric tests must be undertaken (Field 2013).

In order to answer the research questions, two non-parametric tests were conducted. Firstly, to determine if the mode of presentation was influential on performance, the Mann-Whitney U test was undertaken (n=788). The six dependent variables were percentage correct scores for each of the content areas identified in Table 1, namely the following: Whole Number, Other Number, Algebra Patterns, Geometric Measurement, Spatial Arrangements, and Data and Chance. The mode was classified as whether the tasks were completed on the iPad or in pencil-and-paper form. Secondly, to determine if visuospatial ability was influential on performance, a Kruskal-Wallis H test was conducted (N=804) on the six content area dependent variables with visuospatial ability as the independent variable. The visuospatial ability variable was low, moderate or high ability. Interpretation and evaluation will be considered as the discussion section.

Data mining results

The purpose of the study was to identify if the mode of presentation and the level of students' visuospatial ability were influential in mathematics task performance across content areas.

Mode: iPad vs pencil-and-paper

The Mann-Whitney U test revealed that there were significant differences in performance across the mode of presentation for three of the six content areas: Whole Number, Algebra Patterns, and Data and Chance.

Performance on Whole Number content area differed significantly between those who answered on pencil-and-paper (mean rank=430.5) and those who answered on iPad (mean rank=358.5), U=91793, z=4.55, p=0.000, $E.S$=0.16. Similarly, performance on the Algebra Patterns content area differed significantly with those who answered on pencil-and-paper (mean rank=426.0) outperforming those who answered on iPad (mean rank=363.0), U=90,045, z=4.04, p=0.000, E.S=0.14. The same pattern emerged for the Data and Chance content area with those answering on pencil-and-paper (mean rank=418.4) performing significantly better than the iPad cohort (mean rank=370.6), U=87,019, z=3.82, p=0.000, $E.S$=0.14. Although the effect sizes were relatively small, the overarching result indicated that when students answered tasks from these three content areas on paper, they outperformed those who answered them on iPad. Figure 3 helps to illustrate the trend of those answering on pencil-and-paper performing better than those answering on iPad.

Visuospatial ability

To determine if level of visuospatial ability was influential in performance, a Kruskal-Wallis H test was undertaken and revealed that there were significant differences in performance across the all six content areas for level of visuospatial ability:

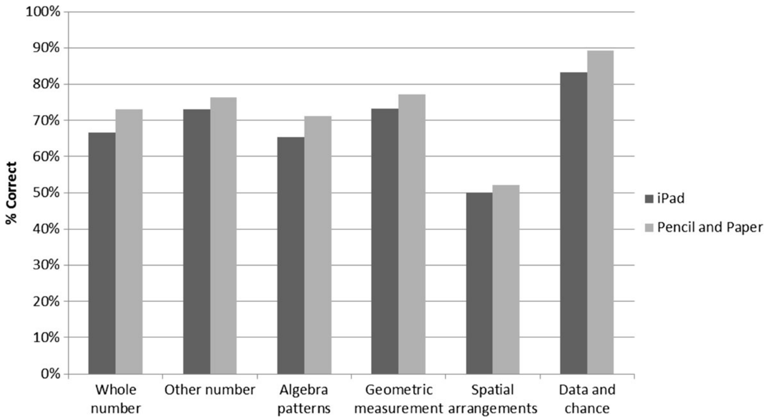


Fig. 3 Performance difference on the iPad and pencil and paper across the six content areas

Whole Number, $H(2)=42.56$, $p=0.000$; Other Number, $H(2)=33.27$, $p=0.000$; Algebra Patterns, $H(2)=71.73$, $p=0.000$; Geometric Measurement, $H(2)=54.12$, $p=0.000$; Spatial Arrangements, $H(2)=87.47$, $p=0.000$; and Data and Chance, $H(2)=37.24$, $p=0.000$. Table 3 provides the subsequent pairwise comparisons for each content area.

Across the six content areas, those students with high visuospatial ability outperformed those students with moderate and low ability, and in turn, those with moderate ability outperformed those with low ability. The largest effect sizes can be found across the low–high ability ranges, although they represent only small to moderate differences. Figure 4 highlights the percentage correct differences between the high level students, the moderate level students and the low level students across each content area.

Table 3 Pairwise comparisons for visuospatial ability level indicating mean rank difference, adjusted *p* values and effect sizes for the six content areas

Pairwise comparison	Whole Number	Other Number	Algebra Patterns	Geometric Measurement	Spatial Arrangements	Data and Chance
Low–moderate	–58.25	–51.99	–73.27	–65.88	–89.48	–67.71
	p=0.009	*p*=0.020	*p*=0.000*	*p*=0.002*	*p*=0.000*	*p*=0.000*
	E.S=–0.10	E.S=–0.09	E.S=–0.13	E.S=–0.12	E.S=–0.17	E.S=–0.15
Low–high	–145.40	–126.18	–186.10	–161.47	–200.93	–106.22
	p=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.000*
	E.S=–0.23	E.S=–0.20	E.S=–0.29	E.S=–0.26	E.S=–0.33	E.S=–0.21
Moderate–high	–87.15	–74.19	–112.83	–95.59	–111.44	–38.51
	p=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.000*	*p*=0.037
	E.S=–0.16	E.S=–0.14	E.S=–0.21	E.S=–0.17	E.S=–0.21	E.S=–0.09

* p levels significant using Bonferroni adjustment where $p \leq 0.008$

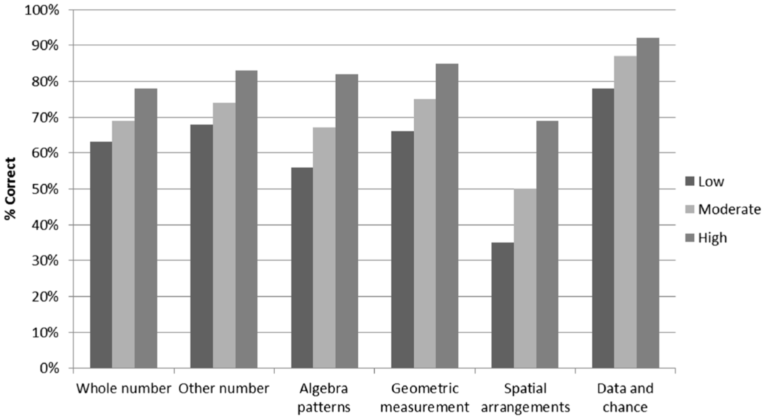


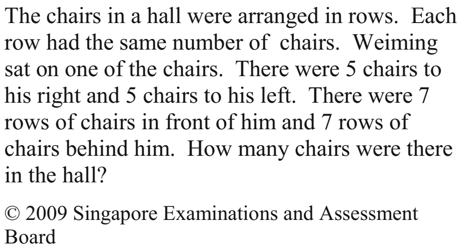
Fig. 4 Percentage correct differences for the low, moderate and high visuospatial ability levels across the content areas

Interpretation and evaluation

The results of this secondary analysis were somewhat surprising given the initial hypotheses. The first hypothesis predicted that there would be no difference in performance across the two modes given the recent research conducted on the topic. However, the statistical analysis revealed that for three content areas, Whole Number, Algebra Patterns and Data and Chance, those students who answered on pencil-and-paper outperformed their counterparts who answered on the iPad. Based on the data alone, it is difficult to provide any clear reasons for why such differences would occur. Threfall et al. (2007) found similar results in their study, where overall test performance across the two modes of delivery was relatively stable, but individual item differences in favour of either PPT or CBT were found. In order to interpret their results, they discussed the characteristics of these items and why such characteristics may have influenced performance. For this paper, a similar type of interpretation was undertaken, where the characteristics of the three content areas may provide suggestions for why there were performance advantages in favour of PPT.

In order to understand the characteristics of the three content areas, the definitions from TIMSS assessment framework were considered. The first content area was Whole Number, for which, the TIMSS framework (Mullis et al. 2009, p. 31) suggested that students should be able to “solve problems by computing, estimating, or approximating with whole numbers”. Four of the six items in this category asked students to complete multi-step problems with whole numbers. These multi-step problems required students to select and apply appropriate strategies for the four operations and coordinate multiple pieces of information. For example, in The Chairs task (Fig. 5), students must monitor not only the chairs and rows and which operations to use, but also remember to include Weiming’s chair and row in the calculation. Another example is The Flowers task (Fig. 6), where students need to coordinate the information in the first part (7 flowers in each of 8 vases, with 3 left over) with the requirements of the second part (9 in each vase). The affordance of being able to show working out in the PPT allowed students to coordinate multiple pieces of information, perform calculations and monitor their thinking.

Fig. 5 The Chairs task



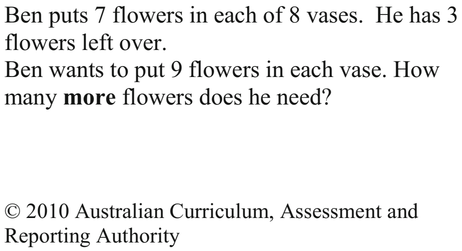
The chairs in a hall were arranged in rows. Each row had the same number of chairs. Weiming sat on one of the chairs. There were 5 chairs to his right and 5 chairs to his left. There were 7 rows of chairs in front of him and 7 rows of chairs behind him. How many chairs were there in the hall?

© 2009 Singapore Examinations and Assessment Board

For the Algebra Patterns category, students should be able to "extend well-defined numeric, algebraic, and geometric patterns or sequences using numbers, words, symbols, or diagrams" (Mullis et al. 2009, p. 33). For two of the four items, algebraic knowledge and strategies would have been advantageous. Both The Trees task (Fig. 7) and The Paper Dolls task (Fig. 8) required students to identify the increasing pattern, where an algebraic rule could be formulated if you had the knowledge or another form of representation could be developed. For the remaining two items, the patterning required either multiplication or addition operations. It is certainly the case that the Singaporean students', whose data was utilised in this study, would be trained in specific heuristics techniques such as drawing the model method (Ng and Lee 2009) to answer such questions and would require the use of pencil-and-paper working out. Again, the affordance of using PPT where students could systematically provide step-by-step representations of their notations could contribute to the performance differences.

For the final content category, Data and chance, the two items represented one from the Data Organisation and Representation sub-domain, The Tank task (Fig. 9) and one from the Chance sub-domain, The Spinner task (Fig. 10). For Data Organisation and Representation, Mullis et al. (2009, p. 37) suggested that students "read scales and data from tables, pictographs, bar graphs, pie charts, and line graphs". For Chance, they suggested students should be able to "judge the chance of an outcome as certain, more likely, equally likely, less likely, or impossible" (Mullis et al. 2009, p. 38). Unlike the previous categories, the two items within this category do not share many content features. One common element is the use of a graphic within the task, where the graphic must be decoded in order to answer the problem. However, there is little evidence to suggest that this was a contributing factor to the performance difference.

Fig. 6 The Flowers task

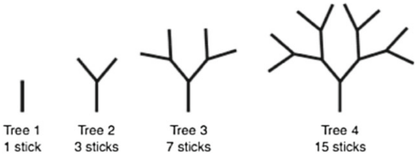


Ben puts 7 flowers in each of 8 vases. He has 3 flowers left over.
Ben wants to put 9 flowers in each vase. How many **more** flowers does he need?

© 2010 Australian Curriculum, Assessment and Reporting Authority

Fig. 7 The Trees task

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?

Although the interpretation of the differences in test mode is somewhat speculative, there is evidence from the items themselves to suggest that assessment questions which require systematic working out and the coordination of multiple pieces of information maybe placing extraneous cognitive load on students when answered in a computer-based environment. The increased cognitive load of students' performing such tasks mentally when answering on the iPad could be a contributing factor for the higher performance in PPT.

The second hypothesis predicted that students with high visuospatial ability would perform better on content areas usually associated with spatial structures, such as geometry and measurement. This was indeed the case; however, the differences were also found across all content areas. There was also a consistent trend where the students with high visuospatial ability outperformed those with moderate visuospatial ability, who in turn, outperformed those students with low visuospatial ability, for all six content areas. This finding highlights the challenges faced by students who struggle with visuospatial activities. Hence, there is a requirement to undertake further research in order to better understand the notion that visuospatial training may improve mathematics performance.

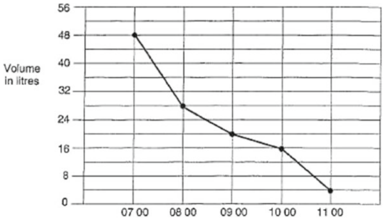
The lower scores across the board for the two tasks associated with Spatial arrangement were unanticipated. It could be argued that the measure of visuospatial ability (see Fig. 2) was directly related to the types of tasks in this content area (for example Fig. 11), as the mental action of folding and unfolding was required for both. However, even those students who were classified as having high visuospatial ability did not perform as well as on this content area as the other content areas. It could be the case

Fig. 8 The paper dolls task

Lili spent 4 days making paper dolls for her friends. Each day she managed to make 2 paper dolls more than the day before. She made a total of 24 paper dolls.
How many paper dolls did she make on the last day?

Fig. 9 The tank task

A tank was filled with 48ℓ of water at 07 00. Water flowed out the tank from 07 00 to 11 00. The line graph shows the amount of water in the tank from 07 00 to 11 00.



During which one-hour period was the decrease in the volume of water the greatest?

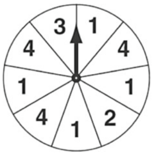
that the measure of visuospatial ability was actually measuring a different cognitive ability or the fact that the two tasks in the Spatial Arrangement content

Knowledge discovery

In drawing interpretations from this study, it is should be noted that the participants' mathematics knowledge could be considered advanced. Most international studies have consistently found that Singaporean students are among the highest performing in the world and that only a very small proportion of this country's students do not reach international benchmarks (Mullis et al. 2012). As a consequence, such studies would anticipate findings with less cohort variance than would be the case in other countries. Nevertheless, this study highlighted differences in students' understanding of particular mathematics content (a) across test mode and (b) in relation to their level of visuospatial ability.

Fig. 10 The spinner task

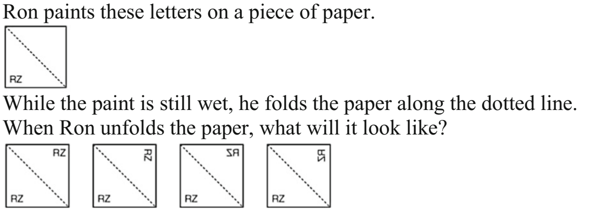
The spinner is used in a board game.



Sanjay spins the arrow.

On which number is the arrow **most** likely to stop?

© 2010 Australian Curriculum, Assessment and Reporting Authority


Ron paints these letters on a piece of paper.

While the paint is still wet, he folds the paper along the dotted line.
When Ron unfolds the paper, what will it look like?

© 2010 Australian Curriculum, Assessment and Reporting Authority

Fig. 11 An example of a task in the Spatial Arrangement content area

Differences across test mode were especially evident when high levels of self-monitoring and systematic planning were required. Students' mean scores were higher when the mathematics tasks were presented in PPT form. The three content categories that produced statistically significant differences across test mode included Whole Number, Algebra patterns and Data and Chance. Within the Whole number and Algebra patterns content areas, it was necessary for the participants to represent information in a relatively structured and sequential manner and the solution approaches often required the recording and monitoring of information. If these solutions were performed mentally, the problem solver was required to perform mentally demanding operations, which in line with Sweller's (1994) theory, increased the cognitive load of the student. The difference in the Data and Chance content area across test mode was hard to interpret within the confines of this study. Hence, there are opportunities for further research to discover the types of processing strategies that students' utilise when answering assessment questions in both CBT and PPT and whether the CBT tests are fairly testing the assigned knowledge and skills.

With respect to the participants' visuospatial ability, the three content areas that revealed the largest effect sizes between participants' who were categorised as high and low visuospatial ability were Spatial arrangements, Algebra patterns and Geometric measurement. It was unsurprising to find differences between the Spatial arrangement and Geometric measurement areas given the formulation of hypothesis two. There were, however, surprising differences across other content areas that are not necessarily considered spatial in nature. In particular, the larger effect size on the Algebra pattern area was unexpected; however, the nature of these tasks required less formal algebra and more pre-algebraic thinking. In some primary school mathematics curriculums, formal symbolisation is not required but rather students are able to recognise and analyse patterns and develop generalisations regarding those patterns. These processes and skills require higher levels of visuospatial thinking than analytic or symbolic thinking. Consequently, those students with higher levels of visuospatial thinking would be better equipped to perform these tasks.

Implications

Two implications arise from the study. Firstly, with the demand for international and national assessments to be administered in CBT form, longitudinal analysis of student performance will need to be carefully considered if attempting to compare results of CBT to PPT, if in fact, it is undertaken at all. Despite research attempting to prove test equivalency, for certain content areas, differences were evident and may not provide an adequate comparison between the two modes. Given this study was undertaken within a high performing nation, international assessments such as PISA will need to research and investigate more fully the effect of presenting certain items within CBT mode and the long-term impact this may have on accurate measurements of students' mathematics knowledge.

Secondly, there needs to be increased attention devoted to improving students' visuospatial reasoning ability in the Singaporean primary classrooms, since it is evident that across all content areas, those students with higher levels perform better. With more and more information being presented visually and spatially, often within digital environments, such reasoning ability is especially critical. Such findings could be influential in recommending students with high visuospatial ability undertake STEM related subjects or move into the STEM related fields. Given that Singapore's main resource are its people, being strategic about how best to utilise that resource would be advantageous. Further research needs to be undertaken in different cultural settings to better understand if the results highlighted here are universal.

Limitations of the study

There were two limitation associated with this study. The first concerns the uneven number of items in each content area category. It is acknowledged that the classification of items into content areas produced unequal number of tasks, and that for two of the content areas, there was limited variance. The second limitation is associated with the notion that research involving secondary data analysis is likely to include some methodological limitations due to the fact that the design and data structure are predetermined by others.

Acknowledgments The author wishes to thank Professor Tom Lowrie for generously allowing me to re-analyse his data and for his comments and insights into the paper.

References

Baker, R. S. J. D. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 112–118). Oxford, UK: Elsevier.

Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education, 21*, 47–60.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, *6*(9), Retrieved September 25 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.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33*, 669–680.

Cheng, Y.-L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development, 15*(1), 2–11. doi:10.1080/15248372.2012.725186.

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.

Clements, D. H. (2004). Geometric and spatial thinking in early childhood education. In D. H. Clements, J. Sarama, A.-M. DiBiase, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: standards for early childhood mathematics education* (pp. pp. 267–pp. 297). Mahwah: Erlbaum.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 420–464). New York: Macmillan.

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

Devine, P. (2003). Secondary data analysis. In R. L. Miller & J. D. Brewer (Eds.), *The A-Z of social research* (pp. pp. 286–pp. 289). London: SAGE Publications, Ltd. doi:10.4135/9780857020024.n97.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton: Educational Testing Service.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Thousand Oaks, CA: Sage Publications Inc.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., Bryant, J. V., & Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical and general cognitive abilities? *Developmental Psychology, 46*, 1731–1746. doi:10.1037/a0020662.

Halpern, D. F., & Collaer, M. L. (2005). Sex differences in visuospatial ability: more than meets the eye. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 170–212). New York: Cambridge University Press.

Hardré, P. L., Crowson, H. M., Xie, K., & Ly, C. (2007). Testing differential effects of computer-based, web-based and paper-based administration of questionnaire research instruments. *British Journal of Educational Technology, 38*(1), 5–22.

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.

Kirby, J. R., & Boulter, D. R. (1999). Spatial ability and transformational geometry. *European Journal of Psychology of Education, 14*(2), 283–294.

Kozhevnikov, M., Hegarty, M., & Mayer, R. E. (1999). *Students' use of imagery in solving qualitative problems in kinematics*. Washington, DC: U.S Department of Education. (ERIC Document Reproduction Service No. ED433239).

Kyttälä, M. (2008). Visuospatial working memory in adolescents with poor performance in mathematics: Variation depending on reading skills. *Educational Psychology: An International Journal of Experimental Educational Psychology, 28*(3), 273–289.

Lowrie, T., & Diezmann, C. M. (2007). Solving graphics problems: Student performance in the junior grades. *Journal of Educational Research, 100*(6), 369–377.

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.

Marginson, S., Tytler, R., Freeman, B., & Roberts, K. (2013). *STEM: Country comparisons*. Report for the Australian Council of Learned Academies. Melbourne, Vic: Australian Council of Learned Academies. Retrieved 17 Feb 2015 from www.acola.org.au.

Mayer, R. E., & Massa, L. J. (2003). Three facets of visual and verbal learners: cognitive ability, cognitive style, and learning preference. *Journal of Educational Psychology, 95*(4), 833–846.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *Chestnut Hill, MA: TIMSS & PIRLS International Study Center*. International Association for the: Evaluation of Educational Achievement (IEA). *TIMSS 2011 Assessment Frameworks*.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, International Association for the Evaluation of Educational Achievement (IEA).

Ng, S. F., & Lee, K. (2009). The model method: Singapore children's tool for representing and solving algebraic word problems. *Journal for Research in Mathematics Education, 40*(3), 282–313.

Pittalis, M., & Christou, C. (2010). Types of reasoning in 3D geometry thinking and their relation with spatial ability. *Educational Studies in Mathematics, 75*, 191–212. doi:10.1007/s10649-010-9251-8.

Reuhkala, M. (2001). Mathematical skills in ninth-graders: relationship with visuo-spatial abilities and working memory. *Educational Psychology: An International Journal of Experimental Educational Psychology, 21*(4), 387–399. doi:10.1080/01443410120090786.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92. doi:10.1016/j.intell.2006.05.004.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Skemp, R. R. (1986). *The psychology of learning mathematics*. London: Penguin.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312.

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. doi:10.1007/s10649-006-9078-5.

Tolar, T. D., Lederberg, A. R., & Fletcher, J. M. (2009). A structural model of algebra achievement: computational fluency and spatial visualisation as mediators of the effect of working memory on algebra achievement. *Educational Psychology: An International Journal of Experimental Educational Psychology, 29*(2), 239–266. doi:10.1080/01443410802708903.

Tversky, B. (2004). Visuospatial reasoning. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 209–240). New York: Cambridge University Press.

Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking in STEM education: when, why and how. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 147–182). San Diego: Academic.

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

Zhang, X., Koponen, T., Räsänen, P., Aunola, K., Lerkkanen, M.-K., & Nurmi, J.-E. (2014). Linguistic and spatial skills predict early arithmetic development via counting sequence knowledge. *Child Development, 85*(3), 1091–1107. doi:10.1111/cdev.12173.