Entering the 'New Frontier' of Mathematics Assessment: Designing and Trialling the PVAT-O (online)

Angela Rogers
RMIT University
<angela.rogers@rmit.edu.au>

As we move into the 21st century, educationalists are exploring the myriad of possibilities associated with Computer Based Assessment (CBA). At first glance this mode of assessment seems to provide many exciting opportunities in the mathematics domain, yet one must question the validity of CBA and whether our school systems, students and teachers are ready to harness this form of assessment. The most obvious advantages of CBA are the speed and accuracy of accessing results and the opportunities for innovative item development. This paper will aim to highlight how several factors can obstruct the validity and reliability of this assessment mode, particularly at an item level. These threats to validity must be carefully considered by test designers to ensure CBA is used effectively in primary school mathematics classrooms.

Throughout my seven years teaching in Victorian Catholic Primary Schools, I was constantly dismayed by the level of superficial whole number place value understanding displayed by Year 3-6 students. Similar difficulties were found in research focussing on student's place value understanding beyond two-digits (Major, 2011; Thomas, 2004). The problems students exhibited in Year 3-6 seemed to be related to a lack of quality assessment instruments available to assist teachers to accurately assess this "critical area of mathematics" (Major, 2011, p. 82).

Without access to quality place value assessments, Year 3-6 teachers were being impeded in their attempts to improve place value teaching and learning. As such, my research has centred around developing a comprehensive whole number Place Value Assessment Tool (PVAT) for Year 3-6 students (Rogers, 2012).

Throughout the course of this research, another reality of classroom teaching and assessment has become apparent- the time factor (Ketterlin-Geller, 2009). During the PVAT paper and pen (P&P) trials, teachers expressed their pleasure at the insights the test provided, yet they were concerned about the time taken to correct and collate the test data. This led to the investigation of the possibility of creating an online version of the PVAT.

This paper will describe the process of designing and trialling the PVAT-O (online), report on its comparability with the P&P version of the PVAT and discuss the implications of using this test in a current school setting.

Literature

Whole number place value is a very difficult concept for students to grasp (Rogers, 2012). The complexities associated with the acquisition of two-digit place value knowledge have been well documented (Baroody, 1990; Ellemor-Collins & Wright, 2009). Yet research by Thomas (2004) and Major (2011) suggest the difficulties students encounter comprehending and applying the recursive multiplicative structure of the number system beyond two-digits, are just as widespread.

Assessment, by its very nature is designed to measure a particular attribute (Morgan, 2000). The primary purpose of this 'measurement' is the opportunity it provides for teachers to gain a better understanding of their students' level of this attribute and thus

In V. Steinle, L. Ball & C. Bardini (Eds.), Mathematics education: Yesterday, today and tomorrow (Proceedings of the 36th annual conference of the Mathematics Education Research Group of Australasia). Melbourne, VIC: MERGA.

hopefully improve the effectiveness of their teaching (Morgan, 2000). A comprehensive place value assessment which addressed the seven aspects of place value (*count, make/represent, name/record, rename, compare/order, calculate and estimate)* (Rogers, 2012) was required to provide teachers with the detailed information they need to conduct targeted teaching at each student's point of need (Vale, Weaven, Davies, & Hooley, 2010).

Traditionally mathematics assessment has been delivered via paper and pencil (P&P) means (Griffin, McGaw, & Care, 2012). However, as we move further into the 21st century, computer based assessment (CBA) provides exciting opportunities for the advancement of the mathematics evaluative process. Just as computers provide many avenues for teachers to trial new ways of teaching, CBA provides opportunities for test developers to explore a multitude of possibilities. This, coupled with the recognition that "doing mathematics with the assistance of a computer is now part of mathematical literacy" (Stacey, 2012, p. 11), has led many, including large scale tests such as PISA and NAPLAN, to move towards investigating the potential of CBA (Tout & Spithill, 2012, December).

CBA can be utilized in several ways in a mathematics assessment context. These include facilitating the design of assessments which better address existing constructs (Csapo, Ainley, Bennett, Latour, & Law, 2012), those which address totally new constructs (Stacey & William, 2013) and those which deliver traditional assessment in a more efficient and effective manner (Bridgeman, 2009). Within each of these categories there are also different features of the CBA platform which can be utilized, including fixed and adaptive testing or Internet and other delivery systems (Stacey & William, 2013). While each avenue has its own challenges and validity issues, all provide opportunities for mathematics assessment which have previously been impossible.

Much research associated with CBA has addressed the comparison of a traditional P&P based test with its CBA equivalent (Bennett et al., 2008; Poggio, Glasnapp, Yang, & Poggio, 2004; Thomson & Weiss, 2009; Wang, Jiao, Young, Brooks, & Olson, 2007). Wang et al. (2007) conducted a meta-analysis of 44 mathematics based assessments which compared P&P and CBA versions of the same test, and reported that overall the mode of administration did not have a statistically significant effect on the tests. This supported the work of Poggio et al. (2004) who reported that "there existed no meaningful statistical differences" (p. 30) between the two modes in their research. However, Poggio et al. (2005) did discover, at an item level, there were more substantiative differences.

Item level functioning differences were also explored by Bennett et al. (2009). Their study used two randomly parallel groups of students and found the CBA to be significantly more difficult statistically than the P&P test. The results from this study led Csapo et al. (2012) to warn that until further studies with alternate research designs, such as that used by Bennett et al. (2008), are conducted, the view that P&P is comparable to CBA should, at best, be "viewed as preliminary" (p. 184).

The reasons for the differences in student performance at an item level between the P&P and CBA mode are varied and can be difficult to pinpoint. Csapo et al. (2012) suggest that factors such as the quality of graphics available on a CBA platform have been found to affect the way a student interacts with items. Computer graphics are considered to provide "richer stimulus material" (Csapo et al., 2012, p. 153), which affects a s tudents' engagement (Stacey, 2012) and potentially their proficiency to answer the item. Thus the importance of researchers closely analysing differences in item functioning is critical to ensuring accurate comparisons are made between the two modes.

Not only does CBA provide opportunities for traditional P&P tests to be converted into more efficient computer based forms, it has many applications in the design and presentation of innovative new ways of assessing mathematics. As Stacey (2012) points out, CBA items can be "more interactive, authentic and engaging" (p. 11). The use of item formats such as 'drop and drag', 'radio buttons' and the possibility of using "dynamic stimuli" (Csapo et al., 2012, p. 149) like audio, video or animation allow greater scope for targeting aspects of mathematical constructs that have never been tested before. Yet, as Csapo et al. (2012) suggest, CBA also brings forth many challenging validity issues.

The work of Lowrie and Diezmann (2009) although focusing on P&P assessment, questions whether assessment items which include graphical representations are measuring students' ability to decode and interpret the graphics rather than assessing their content knowledge. These questions become particularly pertinent in the CBA domain, where the use of dynamic stimuli such as images, video or audio input, could potentially change the skill or content that is intended to be assessed (assuming coping with this stimulus is not the intended outcome of the assessment). This phenomenon was noted in the PISA 2006 computer based assessment of science (CBAS) trial, where differences in item scores were not a result of the mode of delivery but of a feature that was associated with the delivery mode (Csapo et al., 2012). This poses significant challenges for CBA test developers.

While in P&P mode, administration errors such as printing or missing pages can cause validity issues, research by Bridgeman, Lennon and Jackenthal (2003) reported that variations such as screen size, screen resolution and display rate can all influence the way students experience CBA. Thompson and Weiss (2009) explain how Internet-based assessments have validity issues associated with the bandwidth, the browser used to access the sites and the general capabilities of the school computer facilities. Clearly these external factors are not within the control of most test developers and thus are difficult to address, making the pursuit of a valid and reliable CBA even more challenging.

Methodology

Background to research

The first phase of this research involved the construction of a Hypothetical Learning Trajectory (HLT) (Simon, 1995) addressing the seven aspects of place value (Rogers, 2012). Assessment items were designed to target a range of difficulties within each of these aspects and these formed the basis of the PVAT P&P test. This test was piloted at two primary schools (A and B) in metropolitan Melbourne. The test was a 45 minute P&P test suitable for Year 3-6 students and included a total of 78 short answer items which teachers considered time consuming to score. The theoretical paradigm underpinning this research is informed by Cobb's (1996) emergent perspective that acknowledges the social and psychological elements at play in the construction of meaning.

PVAT-O trial

The trial for the PVAT-O was conducted at a Catholic Primary school (School C) in metropolitan Melbourne. The school had approximately 253 Year 3-6 students across nine classes, which were all involved in the trial. The trial was undertaken using a counterbalanced measures design (Shuttleworth, 2009). This research design required half the students in each class (randomly selected) to complete the PVAT-O, and then exactly one week later complete the PVAT. Concurrently the other half of each class completed

the PVAT followed by the PVAT-O one week later. The counterbalanced research design was used to minimise factors such as learning effects and order of treatment, adversely influencing the results of the trial (Shuttleworth, 2009).

Both test forms were identical in their mathematical content. The selected items had previously been validated using Rasch analysis in their P&P form (during School A and B trial) and covered a range of item difficulties and aspects of place value (Rogers, 2012).

The items and images used in each mode were "identical". However, as the items were originally written in the P&P form, some needed to be slightly altered for the CBA platform. Figure 1 shows how "Question 46" in the PVAT-O required students to click the dots to colour them, while this item in the PVAT (Figure 2) required traditional colouring skills. This item was designed to address the "count" (Rogers, 2012) aspect of place value.

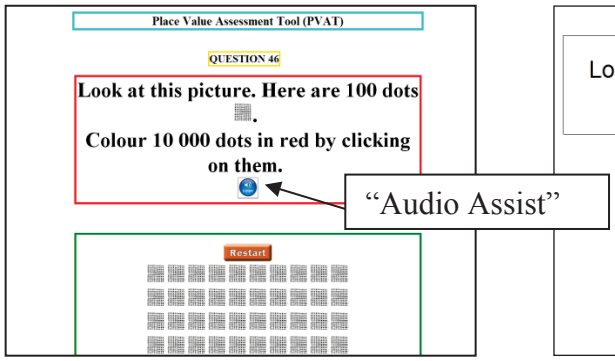

Figure 1. PVAT-O (Count Item).

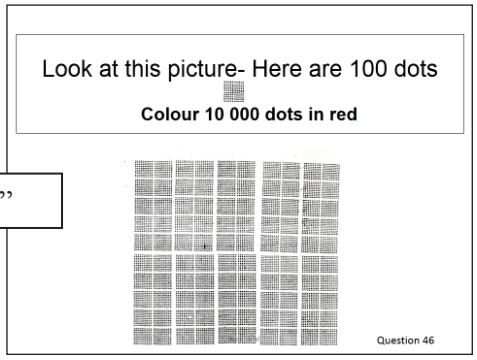

Figure 2. PVAT (Count Item).

Another major difference between the two modes was the inclusion of an "audio assist" button on each PVAT-O item (Figure 1). Each student was provided with earphones during the PVAT-O test and could choose to click on the "audio assist" button to hear the text being read. In order to gain an indication of the frequency this feature was used throughout the trial, students were asked to record the number of times they employed this assistance. This feature allowed PVAT-O students to choose to read the item themselves or listen to the text being read. In the PVAT, students could only read the item themselves.

The time taken for students to complete both the PVAT and the PVAT-O was recorded in order to determine the duration of each mode. After the students had completed both tests they were asked to complete a short survey indicating the version of the test they preferred and reasons associated with their selection. All students completing both versions of the test were supervised by the researcher.

The data from the paper version were coded and scored by the researcher using the same criteria as the online version. Each PVAT-O item response was double checked by the researcher to confirm the online database was consistently scoring the student responses, ensuring valid data for the analysis.

Importantly while 253 students were involved in the research, due to several unavoidable circumstances including student absence, technological issues and challenges in matching the birthdate students entered on the PVAT-O with that on the PVAT, there was a total of 227 students (M=45%, F=55%) who were identified as completing both forms of the test. The analysis was restricted to these students.

Rasch Analysis

A Rasch analysis (Adams & Khoo, 1996) was conducted on the data collected from the trial to address the question of whether the PVAT and the PVAT-O could be considered substantively different in their difficulty. Rasch is a probabilistic model that measures item difficulty and student achievement on the same logit scale (Siemon, Breed, Dole, Izard, & Virgona, 2006). Three Rasch analyses (Run A, B and C) were used to determine the mean of the item difficulties from each mode. The first analysis (Run A) looked at student responses to the PVAT items. The results from Run A created an anchor file composed of the items which were considered to be valid and reliable in this mode of administration.

The next analysis (Run B) looked at student responses to the PVAT-O items. The items which were considered valid and reliable from this analysis were then used in Run C. The final analysis (Run C) combined the student responses to each item on both the PVAT and the PVAT-O. The *anchor file* from Run A was used in this analysis as it allowed the item difficulty estimates of the *surviving* PVAT items to be fixed, so that the PVAT-O items could then be calibrated against them (Izard, 2005). The purpose of Run C was to investigate if items difficulties varied by administration mode when anchoring was used.

The item difficulties for each mode were collated and the mean for the PVAT and PVAT-O was calculated from Run C. These results were compared using effect size measures as this would provide a simple way of quantifying the difference between the item difficulties and student achievement on both tests. Run C of the Rasch analysis also allowed for the mean ability of the students who completed both tests to be compared.

Results

Table 1 summarises the findings of the Rasch analysis (Run C) which looked at the comparison of item difficulties in the PVAT and PVAT-O.

Table 1
Effect Size Estimates for Items by PVAT Administration Mode (Anchored Run)

	PVAT items (N=46)	PVAT-O items (N=62)
Mean	0.34	0.37
Mean Difference	0.03	
Standard Dev.	2.16	1.86
Pooled Std. Dev.	1.99	
Effect Size (Std error)	0.02 (0.19)	
Descriptor	Very Small (Izard, 2004, March)	

The effect size measure calculated for the comparison of the PVAT and the PVAT-O was calculated to be 0.02. This is described to be a "very small (0.00-0.14)" (p. 8) magnitude of effect size (Izard, 2004, March). This suggests that in this study there was not a substantiative difference between the two modes of administration.

Table 2 summarises the findings from the Rasch analysis (Run C) which looked at the difference in estimates of student achievement between the PVAT and PVAT-O.

Table 2
Effect Size Estimates for Students by PVAT Test Administration Mode

	PVAT (N=227)	PVAT-O (N=227)
Mean	0.61	0.63
Mean Difference	0.02	
Standard Dev.	0.23	0.19
Pooled Std. Dev.	0.21	
Effect Size (Std error)	0.09 (0.09)	
Descriptor	Very small (Izard, 2004, March)	

The effect size measure calculated for the comparison of the student ability of those completing the PVAT and PVAT-O was calculated to be 0.09. This is described to be a "very small" (p. 8) magnitude of effect size (Izard, 2004, March). This suggests there are no substantiative differences between the students' achievement in each mode.

Discussion

It must be noted that for this type of analysis, the trial involved a relatively small sample, both in the number of students and the number of items. This limits the scope of conclusions that can be made from the research, particularly at an item level. The trial should be considered merely a population of items and students, not a representative sample. However, with this in mind it is interesting to note that consistent with other research in this area (Poggio et al., 2004; Thomson & Weiss, 2009), at an item level there are several cases where substantiative differences emerge across the two modes.

The differences noted at item level highlights the importance of investigating the features of items that may be influencing the way students are approaching them, particularly in the CBA mode. The use of graphics and other features such as 'drop and drag' are factors which provide great opportunities for innovative item development but also should be considered to be possible threats to the validity of the item.

The difficulty with CBA is accurately pinpointing the factors which are affecting the way students approach items. For example, the data collected in this study suggests that 42% of students reported using the audio assist button on an average of 3.79 items. This may or may not have affected the way these students answered such items.

Another important aspect associated with CBA is the affective side of this mode of delivery. Csapo et al. (2012) notes that the level of proficiency and the general familiarity students have with computers can affect their level of interest and approach to CBA. The student surveys in this study suggest that 55% of students preferred completing the PVAT-O test, stating reasons such as "it's easier to see the graphics", "I like using computers more" and "you can listen to the question if you get stuck". While those who preferred the paper version cited reasons like "it takes longer on the computer", "you can do more working out when you have it on paper in front of you" and "the computer is frustrating". It should be noted that the PVAT-O took students on average 37 minutes while the PVAT took an average of 32 minutes. This discrepancy was mostly due to the speed of the school's internet capabilities. Some computers seemed to take a great deal longer than others to move through the PVAT-O, no doubt frustrating the students working on them.

It became apparent that from a systems perspective, the capability of the school's computer and technological infrastructure is of immense importance when implementing a CBA platform. In order to successfully facilitate CBA within a school, there are many

considerations that need to be made. These include addressing logistical issues such as sourcing enough computers and locating appropriate spaces to administer the test, and practical considerations such as making available assistance to deal efficiently with technological issues which commonly arise in this mode of testing. All of these variables influence the success and validity of the CBA testing process. Clearly this testing mode requires commitment from the school, teachers and students to be a success.

Conclusion

As our society continues to advance technologically, mathematics educators and test designers alike must carefully consider the future of traditional paper and pen assessments in light of the ever evolving computer based assessment platform. The advantages claimed for CBA include immediate results and the opportunities for innovative and creative item development. However, there are also many issues surrounding the validity and reliability of CBA, particularly at an item level. The results of this research paper suggest that, like previous research in this area (Poggio et al., 2004; Wang et al., 2007), the PVAT-O and PVAT appear to be very similar in overall scores of item difficulty. However, at an item level there are several items which appear to display substantively different difficulty thresholds suggesting there may be features of the CBA platform which alter the construct being measured. This is an area of the PVAT-O research which is currently being investigated further. Furthermore, the capabilities and resources of schools that choose to embrace this mode of assessment are all variables that are difficult to control and pose significant threats to validity and success of CBA. It seems Lowrie's (2012, November) suggestion to "hasten slowly" when entering the *new frontier* of CBA is sound advice.

References

Adams, R. J., & Khoo, S. T. (1996). QUEST-Version 2.1 The interactive Test Analysis System.

Baroody, A. (1990). How and when should place-value concepts and skills be taught? *Journal for Research in Mathematics Education, 21*(4), 281-286.

Bennett, R. E., Braswell, J., Oranhe, 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 NEAP. *Journal of Technology, Learning and Assessment, 6*(9). Available: http://escholarship.bc.edu/jtla/vol6/9/

Bridgeman, B. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large scale testing* (pp. 39-44). Luxembourg: Office for Official Publications of the European Communities.

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

Cobb, P., & Yackel, E. (1996). Constructivist, emergent and sociocultural perspectives in the context of developmental research. *Educational Psychologist, 31*(3), 175-190.

Csapo, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143-231). London: Springer.

Ellemor-Collins, D., & Wright, R. (2009). *Developing conceptual place value: Instructional design for intensive intervention.* Paper presented at the (Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia), Palmerston, NZ.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. London: Springer.

Izard, J. F. (2004, March). *Best practice in assessment for learning*. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on Redefining the Roles of Educational Assessment, Nadi, Fiji.

Izard, J. F. (2005). *Trial Testing and Item Analysis in Test Construction: Module 7*. Paris: International Institute for Educational Planning (UNESCO).

Ketterlin-Geller, L. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research and Evaluation, 14*(16), 1-11.

Lowrie, T. (2012, November). *Assessment in digital environments: Hasten slowly*. Presentation given at the Pearson Global Research Conference, Fremantle. Poewerpoint retrieved from http://www.pearson.com.au/marketing/corporate/pearson_global/presentations/10.50am-Thematic_6-Prof_Tom_Lowrie-Assessment_in_Digital_Environments-Hasten_Slowly.pdf.

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

Major, K. (2011). *Place Value: Get it. Got it. Good enough?* (Unpublished Master's thesis), University of Auckland, Auckland.

Morgan, C. (2000). Better assessment in mathematics education? A social perspective. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning.* Westport, CT: Ablex Publishing.

Poggio, J., Glasnapp, D., Yang, X., & Poggio, A. (2004). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning and Assessment, 3*(6), 30-38.

Rogers, A. (2012). Steps in developing a quality whole number place value assessment for Years 3-6: Unmasking the "experts". In J. Dindyal, L. P. Cheng & S. F. Ng (Eds.), *Mathematics education: Expanding horizons* (Proceedings of the 35th annual conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 650-657) Singapore: MERGA, Inc.

Shuttleworth, M. (2009). Counterbalanced measures design. Retrieved 24/12/12, from http://explorable.com/counterbalanced-measures-design.html

Siemon, D., Breed, M., Dole, S., Izard, J., & Virgona, J. (2006). *Scaffolding Numeracy in the Middle Years-Project Findings, Material and Resources. Final Report*. RMIT University. Melbourne. Retrieved from www.eduweb.vic.gov.au/edulibrary/public/teachlearn/student/snmy.ppt

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114-145.

Stacey, K. (2012). *The international assessment of mathematical literacy: PISA 2012 framework and items*. Paper presented at the 12th International Congress on Mathematics Education, COEX, Soel, Korea, July 8-15. Seoul, Korea. Retrieved from http://www.icme12.org/upload/submission/2001_f.pdf

Stacey, K., & William, D. (2013). Technology and Assessment in Mathematics. In M. Clements, A. Bishop, C. Keitel, J. Kilpatrick & F. Leung (Eds.), *Third International Handbook of Mathematics Education*. Netherlands: Springer.

Thomas, N. (2004). The development of structure in the number system. In M. Johnsen Hoines & A. Berit Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*: Vol. 4, pp 305-312. Bergen, Norway: Bergen University College Press.

Thomson, N., & Weiss, D. (2009). Computer and adaptive testing in educational assessment. In S. F & J. Bjornsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large scale testing.* (pp. 127-133). Luxembourg: Office for Official Publications of the European Communities.

Tout, D., & Spithill, J. (2012, December). *From paper to screen: Computer-based assessment of mathematics-Lessons from PISA.* Presentation given at the Mathematical Association of Victoria Conference. Melbourne.

Vale, C., Weaven, M., Davies, A., & Hooley, N. (2010). *Student centered approaches: Teacher's learning and practice.* Paper presented at the MERGA 33- Shaping the future of Mathematics Education, Fremantle, WA. http://www.merga.net.au/documents/MERGA33_ValeEtAl.pdf

Wang, S., Jiao, H., Young, M., Brooks, T., & Olson, J. (2007). A meta analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219-238.