Computer based mathematics assessment: Is it the panacea?

Angela Rogers
Royal Melbourne Institute of Technology
<angela.rogers@rmit.edu.au>

Test developers are continually exploring the possibilities Computer Based Assessment (CBA) offers the Mathematics domain. This paper describes the trial of the Place Value Assessment Tool (PVAT) and its online equivalent, the PVAT-O. Both tests were administered using a counterbalanced research design to 253 Year 3-6 students across nine classes at a primary school in Melbourne. The findings show while both forms are valid and comparable, the online mode was preferred by teachers. The affordances and constraints of using CBA in the formative assessment process are explored.

Over the past 10 years there has been a rapid uptake of Mathematics *Computer Based Assessments* (CBA) in Australian primary schools. Commercial firms have identified teachers as eager consumers in this market. Companies are acutely aware of the friction points for teachers: the challenges around creating their own formative assessments and time-consuming marking. This has led to the development of several increasingly popular CBA formative assessment "programs". Yet, for these programs be the panacea their advertising suggests, schools must be confident they provide valid formative data teachers can easily interpret and apply.

Currently in Australia, there are very few comprehensive formative whole number place value assessments for Years 3-6 students. To address this, a Rasch analysis-based methodology was used to develop a valid and reliable whole number place value paper-and-pen assessment, called the *Place Value Assessment Tool* (PVAT) (see Rogers, 2014). While the PVAT provided a detailed picture of student knowledge in the construct, the time taken to mark (5-7 minutes per student) was seen as a potential obstacle for teachers. To address this, the researcher investigated if a comparable online version of the test could be created.

Relevant Literature

Place value knowledge has been compared to the framework of a house, such that if a student's knowledge in this area is shaky, his/her understanding of mathematics as a whole is affected (Major, 2011). An understanding of place value has been shown to be closely related to students' sense of number (McIntosh et al., 1992), understanding of decimals (Moloney & Stacey, 1997), and comprehension of multi-digit operations (Fuson, 1990). Underpinning almost every aspect of the mathematics curriculum, it is an integral part of the primary school syllabus. Yet there is considerable evidence to suggest students struggle with whole number place value well into lower secondary school (Thomas, 2004; Wade et al., 2013). Research has shown that place value is often taught superficially, something that can be attributed to the lack of quality formative assessments available in this construct (Major, 2011; Rogers, 2014).

An assessment is essentially a sample of selected tasks intended to allow inferences to be made about a student's level of achievement. The strength of these inferences relies heavily on the quality of the tasks used (Izard, 2002). An assessment which includes a selection of items that are too easy, or too difficult, will not provide teachers with a complete picture of each student's knowledge. Similarly, an assessment that does not comprehensively cover the required content may cause the omitted content to be devalued by teachers (Webb,

2021. In Y. H. Leong, B. Kaur, B. H. Choy, J. B. W. Yeo, & S. L. Chin (Eds.), *Excellence in Mathematics Education: Foundations and Pathways (Proceedings of the 43rd annual conference of the Mathematics Education Research Group of Australasia)*, pp. 329-336. Singapore: MERGA.

2007). In both cases, the inaccurate inferences drawn from these assessments, adversely influence the quality of instruction. Formative assessment is a process that provides teachers with information that can be used to support individual student's future learning (Popham, 2018). It is one of the most effective, empirically proven, processes that teachers can use to improve student performance.

An important consideration when developing assessments is practicality (Masters & Forster, 1996). If an assessment instrument does not justify the time or money required for its administration and marking, it will not be implemented by schools. Doig (2011) noted that some educators (despite appreciating the quality of data they received) avoided using interview-based assessments simply because of their administration time. As a result, many schools consider paper-and-pen tests a more practical assessment option, particularly with older students. Proponents of interview-based assessments disagree, stating clinical interviews provide higher quality assessment information and enhance teacher knowledge of common misconceptions in mathematics (Clements & Ellerton, 1995). While mathematics assessments have traditionally been delivered via paper-and-pen or interview (Griffin et al., 2012), the accessibility of technology has seen test developers investigate the many opportunities provided by CBA (ACARA, 2021).

CBA's major advantage is it delivers traditional assessment in a more efficient and effective manner (Bridgeman, 2009). CBA has the potential to save teachers time marking test papers and means results can be used to guide instruction in a timelier manner (Tomasik et al., 2018). Yet, as Thompson and Weiss (20011) explain, many school's technological capabilities fail the standard required to successfully implement CBA, leading to test administration problems (McGowan, 2019). Thus, while CBA has great potential in schools, further logistical work is required to ensure its success.

Much research associated with CBA has explored the comparison of traditional paper-and-pen based tests with their CBA equivalent (e.g., Wang et al., 2007; Thompson & Weiss, 2011). Wang et al. (2007) conducted a meta-analysis of 44 mathematics-based assessments comparing paper-and-pen and CBA versions of the same test. Overall, they reported that the mode of administration did not have a substantive effect on the students' performance (ES = -0.059). These comparisons aimed to determine whether online and paper versions of the same test could be used interchangeably. This is an important practical consideration, as comparable tests allow schools the flexibility to choose the most appropriate mode for their context. Yet, as Popham (2018) suggests, the decisions around test selection rely heavily on the assessment literacy of teachers and school leaders.

Popham (2018) defines assessment literacy as an "individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (p.13). An assessment literate teacher makes informed choices around the assessments they use, and accurately applies the results to guide their instruction. Research has shown that assessment literacy is not usually a focus of teacher education, meaning most teachers have poor levels (Stiggins, 2006). While providing teachers with assessment literacy professional development has been shown to be effective (Xu & Brown, 2016), without access to this, teachers are left to develop these skills 'on the job'. As CBA is a relatively new mode of assessment, it is realistic to assume that teachers need support to develop their assessment literacy in this mode. As Popham (2008) points out, being provided with assessment data is only the beginning of the process – teachers need the assessment literacy skills to understand a test's construction so they can successfully *interpret* the data.

One proven method of test construction is *Item Response Modelling* (IRM), which has well-established methods for analysis (Wright & Masters, 1982). IRM measures the

relationship between student achievement and item difficulty on the same scale (Wright & Stone, 1979). IRM has been successfully applied to a variety of test modes and used in large-scale assessments through to high-quality classroom-based assessment tools including PAT-M (Australian Council for Educational Research, 2012) and the *Scaffolding Numeracy in the Middle Years* (SNMY) assessment (Siemon et al., 2006). A popular IRM model, devised by Rasch (1960) is used in this research. Rasch analysis is based around the interplay of candidates and items in an assessment. While analysis of assessments traditionally generates a score that summarises the number of items correctly answered by students, Rasch considers the students who correctly answered each item (Izard, 2004). Rasch examines the extent to which the item distinguishes between those who are more and less knowledgeable (Izard et al., 2003). That is, the model assumes that less knowledgeable students have lower probability of answering a difficult item compared with those who are more knowledgeable (Rasch, 1960). Items that are considered not to follow this pattern do not fit the Rasch model and are generally removed from a test. This process verifies that the test content is meaningful and appropriate so that useful inferences can be made about the knowledge of candidates (Izard et al., 2003). Rasch allows different tests to be located on the same scale and allows test designers to determine if they are of comparable difficulty. The next section describes how quantitative Rasch based methods were used to compare the PVAT and PVAT-O, and the qualitative methods used to gather insights from teachers.

Methodology

PVAT-O Creation

Multiple technologies including *HyperText Markup Language* (HTML5), *Javascript*, and *PHP: Hypertext Preprocessor* (PHP) were used to create the PVAT-O assessment. The mathematical content and format of each PVAT-O item was as close as possible to the equivalent PVAT items. However, some items required the inclusion of computer-based features. For example, a ‘drag and drop’ feature was used in items requiring students to place numbers in order from smallest to largest and ‘radio buttons’ were used in multiple choice items.

The Counterbalanced Trial

The online and paper and pen PVAT trial was conducted at School C, a Catholic Primary school in metropolitan Melbourne where approximately 11% of students were from English as an Additional Language or Dialect (EAL/D) families (ACARA, 2020). All Year 3 to 6 students (N = 253) from nine classes took part in the trial (Male= 47%, Female= 53%). The trial took place over a 2-week period in the school library and was supervised by both the researcher and the classroom teacher. The trial was conducted using a counterbalanced measures design (Shuttleworth, 2009). Half of the students in each class (randomly selected) completed the PVAT-O, whilst the other half of the class completed the paper-and-pen PVAT. Exactly one week later, the students completed the alternate version of the test. This research design was used to minimise factors such as learning effects and order of treatment, adversely influencing the results of the trials (Perlini et al., 1998). Only 227 students (Male= 45%, Female= 55%) completed both forms, due to absences and technical issues.

Teacher Surveys

A short survey was given to the nine Year 3 to 6 classroom teachers (Female=100%). The purpose of this survey was to gain an indication of the teacher’s preferred testing mode.

When the survey occurred, the teachers had not yet received their student's results from the PVAT-O database, but it was explained they would receive each student's raw score in a spreadsheet. Due to the small sample size, the survey data was interpreted by the researcher and reported as individual responses (Neuman, 2006).

Rasch Analysis

The paper PVAT tests were scored and coded by the researcher. The PVAT-O was scored by the PVAT-O database and then rechecked by the researcher to ensure consistency and accuracy. In order to determine if the PVAT and PVAT-O could be considered valid tests and comparable in their mean item difficulty and mean student achievement (Kolen & Brennan, 2004), three Rasch analyses were conducted:

- *Run A* was conducted to re-confirm that the paper-and-pen PVAT was a valid and reliable test. The items which fit the model were used to create an anchor file for Run C. This allowed the PVAT and PVAT-O items to be placed on the same scale.
- *Run B* looked at the PVAT-O items in isolation. Rasch analysis was used to determine which PVAT-O items fit the model and determine if it was an internally consistent test.
- *Run C* investigated if the PVAT and PVAT-O could be placed on the same uni-dimensional scale and thus determine if they were comparable in item difficulty and student achievement.

The anchor file from Run A was used to fix the difficulty estimates of the PVAT items that fit the model. This allowed the PVAT-O items to be calibrated against the PVAT items (Izard, 2005). The mean item difficulty and mean student achievement for the PVAT and PVAT-O was then calculated from this run. Effect Size measures were used to quantify the standardised mean difference between the two tests (Izard, 2004). Cohen's (1969) descriptors for the magnitude of Effect Sizes, alongside the assigned ranges for each descriptor as suggested by Izard (2004), were then be used to describe the Effect Sizes in plain language.

Results

Rasch Validation and Comparison

The mean and standard deviation of the PVAT (n = 65) and PVAT-O (n = 59) items which fit the model in Run C were calculated to determine if the PVAT and PVAT-O could be considered comparable tests. The Effect Size measure was calculated to be 0.14, while the difference in student achievement between the tests was 0.01. This is described to be a "very small (0.00 to 0.14)" (Izard, 2004, p. 8) magnitude of Effect Size. This suggests there was not a substantive difference between the mean of item difficulties in the two modes, nor the students' achievement (which is to be expected, given the tests were of similar difficulty).

Teacher Survey

The class teachers (N = 9) at School C completed a brief survey asking them to indicate their preferred mode of administration for the PVAT. Seven teachers preferred the PVAT-O, while two preferred the PVAT. The seven teachers who preferred the PVAT-O stated:

'It will save correcting it' (Teacher #1,#2,#3)

'The results are immediate, I can use them the next day in my teaching' (Teacher #4)

'If the computers all work, online is much better' (Teacher #5)

'I don't have to correct it…and I can use the results tomorrow' (Teacher #6)

'The corrections would save me a lot of time and effort' (Teacher #7)

The two teachers who indicated they preferred the PVAT mode stated:

> 'Correcting them myself gives me a sense of their understanding' (Teacher #8)
>
> 'I'm always concerned students will lose their responses' (Teacher #9)

The small sample of teachers completing this survey limits the inferences that can be made from the data. However, within this group of teachers there was a clear preference for the PVAT-O mode of test administration, largely due to marking time it saved.

Discussion

Formative mathematics CBA continues to be embraced by schools, teachers and test developers. This research highlights several considerations when implementing formative CBA in classrooms: transparency, rigor, flexibility, and assessment literacy.

Transparency

While teachers in this research project were provided access to both the paper and CBA version of the test, this is not always the case. For example, in *Computer Adaptive Tests* (CAT) (Martin & Lazendic, 2018) each child is provided with a different set of items according to their responses. It is impossible for a teacher to view the combination of items individual students encounter, thus they are unable to judge their quality, appropriateness and relevance. Without this transparency, teachers are outsourcing the judgement of student knowledge to test designers. While somewhat appropriate in summative situations, eliminating teacher judgement in the formative assessment process should raise concerns for schools. Teacher #8 at School C echoed this 'transparency' constraint, indicating she was concerned about missing important diagnostic information in the PVAT-O. In response to this, the database was later adjusted to ensure teachers were provided with a summary of student responses to each item. The *Specific Mathematics Assessments that Reveal Thinking* (SMART) tests (University of Melbourne, 2012), are another platform that recognises the importance of allowing teachers to 'see' common student errors in the CBA mode. Doig (2011) reiterates this concern, noting that 'off site marking' does little to assist teachers to develop their knowledge of common student errors and misconceptions. Providing teachers with an overall raw score, rather than access to individual responses, is a major constraint of formative CBA and an issue which needs to be addressed by test designers.

Rigor of the Assessment

Wiliam (2007) states that formative assessment can effectively double the speed of student learning. Yet, as often happens in education, approaches can become diluted when commercial firms become involved. In order for schools and teachers to make informed decisions about the worth of formative CBA programs (particularly those produced commercially), it is critical teachers understand how to evaluate the rigor of a test's construction. The results presented in this paper use Rasch analysis to show both the PVAT and the PVAT-O are valid and reliable tests. For schools, this is essential information as it means the test has been empirically proven and robustly constructed. While the relatively small sample size gathered from only one school limits the scope of conclusions that can be made from this trial, very little difference was detected between the mean difficulties and student achievement of test items. Similarly, the student achievement was found to be comparable. This supports the results of the meta-analysis conducted by Wang et al. (2007), which noted that the mode of administration did not have a substantive effect on student achievement in computer-based and paper-based mathematics assessments. As Popham

(2018) suggests, schools should be encouraged to contact test developers, ask for a test's technical guide, and gather information related to the trialing, reliability and validity so they can make informed decisions about the suitability and rigor of tests.

Flexibility

Providing teachers with access to a comprehensive formative place value assessment that can be administered in two modes is considered to increase the usability and practicality of the PVAT tool. The PVAT-O was designed to support teachers by providing instant feedback on their students' achievement and save them considerable time. As the online and paper PVAT tests were found to be comparable, teachers are now able to choose the mode which works best for them and their students. This flexibility is useful, as not all schools have the technological requirements to successfully implement CBA. As Csapo et al. (2012) note, at a minimum, a school requires the capacity to allow students completing the assessment concurrent access to the Internet while still supporting the Internet requirements of the rest of the school. As Huff and Sireci (2001) correctly note, when this does not occur, the validity of the test is threatened. In the PVAT-O trial it was noted that some computers took a great deal longer than others to move through the PVAT-O. This frustrated and disadvantaged the students working on the 'slow' computers. Teachers #5 and #9 both mentioned their concerns with the fragility of the technology at their school, stating "if the computers all work…"(Teacher #5) and "I'm always concerned students will lose their work" (Teacher #9). Providing teachers with a 'back up' paper version of the test is considered a practical way to alleviate these fears.

Assessment Literacy

Popham (2018) explains that educators who are not assessment literate often make inappropriate decisions about which tests to use. Using formative CBA is a relatively new form of mathematics assessment in schools, so it is critical teachers are helped to understand the affordances and constraints of these tools. Teachers are a critical stakeholder in the formative CBA process. They are required to administer the assessment and their interpretation of the results influences its success (Jones & Truran, 2011). Seven of the nine teachers in this research described how they based their mode preference choice solely on the time it would save. Research by Melleti and Khademi, (2018) showed that for both assessment literate and illiterate teachers, time was their main concern when implementing formative assessment. Yet interestingly, assessment literate teachers considered the time they spent creating and marking assessments a necessary part of the process. Thus, it appears that when teachers do not fully appreciate the advantages of formative assessment, they consider the time spent on it untenable. This reinforces the need to develop teacher's assessment literacy skills around formative assessment, particularly in CBA (Popham, 2018). Without appropriate professional development designed to increase assessment literacy, teachers will continue to focus on selecting assessments based on their perceived ease of administration and marking, rather than the quality of the tool.

Conclusion

The demands on a classroom teacher's time have never been greater. Whilst a major affordance of CBA is the time it saves teachers, one of the major constraints is its lack of transparency. When a computer database marks student responses, a teacher's judgment and involvement in the process is removed. This research suggests in order to retain the fidelity of the formative assessment process, teachers require access to professional development

that aims to grow their assessment literacy skills. Developing these skills will encourage teachers to seek quality empirically proven assessments, and assist them to accurately interpret CBA data.

References

Australian Council for Educational Research. (2012). *Progressive achievement tests in mathematics plus (PATMaths Plus)*. Retrieved from http://www.acer.edu.au/tests/patmaths-plus

Australian Curriculum Assessment and Reporting Authority. (2020). *My school: "School C"*. Retrieved from http://www.myschool.edu.au/

Australian Curriculum Assessment and Reporting Authority. (2021) *NAPLAN Online: Research and development.* https://www.nap.edu.au/online-assessment/research-and-development

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.

Caygill, R., & Eley, L. (2001). *Evidence about the effects of assessment task format on student achievement*. Paper presented at the Annual Conference of the British Educational Research Association, University of Leeds, England. Retrieved from http://www.leeds.ac.uk/educol/documents/00001841.htm

Clements, M., & Ellerton, N. (1995). Assessing the effectiveness of pencil-and-paper tests for school mathematics. In B. Atweh & S. Flavel (Eds.), *MERGA18: Galtha. Proceedings of 18th Annual Conference of Mathematics Education Research Group of Australasia.* (pp. 184-188). Darwin: Mathematics Education Research Group of Australasia.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York, NY: Academic Press.

Csapo, B., Ainley, J., Bennett, R., 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.

Doig, B. (2011). *Reporting large-scale assessment on a single formative-summative scale.* [Unpublished Doctoral dissertation], Deakin University, Melbourne, Victoria, Australia.

Fuson, K. (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit addition, subtraction, and place value. *Cognition and Instruction, 7*(4), 343-403.

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

Huff, K., & Sireci, S. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16-25.

Izard, J. (2002). *Using assessment strategies to inform student learning.* Paper presented at the Annual Conference of the Australian Association for Research in Education, Brisbane, QLD. Retrieved from http://www.aare.edu.au/data/publications/2002/iza02378.pdf.

Izard, J. (2004). *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. (2005). *Trial testing and item analysis in test construction: Module 7*. Paris: International Institute for Educational Planning (UNESCO).

Izard, J., Haines, C., Crouch, R., Houston, S., & Neill, N. (2003). Assessing the impact of the teaching of modelling: Some implications. In S. Lamon, W. Parker, & K. Houston (Eds.), *Mathematical modelling: A way of life: ICTMA 11* (pp. 165-177). Chichester: Horwood Publishing.

Jones, A., & Truran, L. (2011). *Paper and online testing: Establishing and crossing boundaries*. Retrieved from http://www.aare.edu.au/data/publications/2011/aarefinal00704.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York: Springer.

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

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205

Masters, G., & Forster, M. (1996). *Developmental assessment*. Melbourne: Australian Council for Educational Research.

McGowan, M. (2019, May 27). *Naplan's online testing to be reviewed after botched rollout.* The Guardian. https://www.theguardian.com/australia-news/2019/may/27/naplans-online-testing-to-be-reviewed-after-botched-rollout

McIntosh, A., Reys, B., & Reys, R. (1992). A proposed framework for examining basic number sense. *For the Learning of Mathematics, 12*(3), 2-8.

Mellati, M., & Khademi, M. (2018). Exploring Teachers' Assessment Literacy: Impact on Learners' Writing Achievements and Implications for Teacher Development. *Australian Journal of Teacher Education, 43*(6).

Moloney, K., & Stacey, K. (1997). Changes with age in students' conception of decimal notation. *Mathematics Education Research Journal, 9*(1), 25-38.

Neuman, L. W. (2006). *Social Research Methods* (6th ed.). United States of America: Pearson Education Inc.

Perlini, A., Lind, D., & Zumbo, B. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie Canadienne, 39*(4), 299-307.

Popham, W.J (2008). *Transformative Assessment.* Alexandria, VA: ASCD.

Popham, W.J. (2018). *Assessment literacy for educators in a hurry.* Alexandria, VA: ASCD.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark's Paedagogiske Institut.

Rogers, A. (2014). *Investigating whole number place value assessment in Years 3-6: Creating an evidence-based Developmental Progression*. [Unpublished Doctoral thesis]. RMIT University.

Shuttleworth, M. (2009).*Counterbalanced measures design*. Retrieved 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*. Melbourne: RMIT University. Retrieved from www.eduweb.vic.gov.au/edulibrary/public/teachlearn/student/snmy.ppt

Stacey, K., & Wiliam, 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* (pp. 721-752). Netherlands: Springer.

Stiggins, R. J. (2006). Assessment for learning: a key to student motivation and learning. *Phi Delta Kappa Edge, 2*(2), 1–19.

Thomas, N. (2004). The development of structure in the number system. In M. J. Hoines & A. B. 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.

Tomasik, M. J., Berger, S., & Moser, U. (2018, November 20). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, *9*, 2245.

Thompson, Nathan A., & Weiss, David A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, *16*(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

University of Melbourne. (2012). *Specific mathematics assessments that reveal thinking (SMART).* Retrieved from http://www.smartvic.com/smart/samples/select_preset.html

Wade, P., Gervasoni, A., McQuade, C., & Smith, C. (2013). Launching confident numerate learners. *The Australian Mathematics Teacher, 69*(3), 26-32.

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.

Webb, N. (2007). Mathematics content specification in the age of assessment. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 1281- 1292). United States of America: National Council of Teachers of Mathematics.

Wiliam, D. (2007). Content then process: Teacher learning communities in the service of formative assessment. In D. B. Reeves (Ed.), *Ahead of the curve: The power of assessment to transform teaching and learning* (pp. 183–204). Bloomington, IN: Solution Tree.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago, IL: MESA Press.

Xu, Y. & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162.