Learning from Their Mistakes - An Online Approach to Evaluate Teacher Education Students' Numeracy Capability

Thuan Thai
University of Notre Dame Australia
<thuan.thai@nd.edu.au>

Kate Hartup
University of Notre Dame Australia
<kate.hartup1@my.nd.edu.au>

Adelle Colbourn
Western Sydney University
<a.colbourn@westernsydney.edu.au>

Amanda Yeung
University of New South Wales
<amanda.ws.yeung@unsw.edu.au>

Teachers' numeracy capability is essential for student learning in the classroom and important across all subject areas, not only within mathematics. This study investigated the use of online diagnostic tests as a form of assessment for learning, to evaluate and support teacher education students (TES) in developing their numeracy skills. Data were collected using the "Test" feature through the Blackboard Learning Management system at two Australian universities. In this paper, we report on trends amongst TES who showed growth in their numeracy capability through the repeated use of the diagnostic test.

Introduction

As part of the general capabilities outlined by the Australian Curriculum and Reporting Authority (ACARA), all teachers are required to teach numeracy skills across all areas at all year levels (ACARA, n.d.). Since teacher knowledge is an important element that informs preparation and teaching (Shulman, 1987), it is essential for teachers to demonstrate an adequate level of personal numeracy capabilities to successfully teach numeracy across the curriculum. Given that research has shown that teachers' mathematical content knowledge affects their students' performance, it is reasonable to postulate that a link may also exist between teachers' numeracy skills and students' numeracy capabilities and students' numeracy capabilities.

There is currently little research that has investigated TES' numeracy skills in Australia and thus this research aims to address this gap. One particular study that specifically explored TES' numeracy skills in Samoa reported that participants demonstrated persistent misconceptions of basic numeracy skills across various topics, including fractions, decimals, percentages, and geometry (Afamasaga-Fuata'i, Meyer, Falo, & Sufia, 2008). Interestingly, Afamasaga-Fuata'i et al. (2008) also reported that in a follow up test, after two semesters of normal load coursework studies, 34 out of 46 research participants showed an overall improvement. A closer inspection of the areas of improvement showed that TES in this study performed better in less difficult questions in the follow up test but showed little improvement with more challenging questions. A more recent study of TES in New Zealand showed that less than half the cohort demonstrated the mandated level of foundational mathematical content knowledge (Linsell & Anakin, 2012). More specifically, only 41% of TES (n=153) in 2010 and 43% of TES (n=122) in 2011 met the numeracy skills standard in this study. These studies show concerning results about the professional standards of numeracy possessed by TES. Therefore, it is important for initial teacher education providers to have knowledge of their TES' numeracy skills and mechanisms to support their development.

2019. In G. Hine, S. Blackley, & A. Cooke (Eds.). Mathematics Education Research: Impacting Practice *(Proceedings of the 42nd annual conference of the Mathematics Education Research Group of Australasia)* pp. 707-714. Perth: MERGA.

Research Aims and Significance

This research identified and evaluated trends amongst TES from two Australian universities whom showed growth in their numeracy skills through the repeated use of an online diagnostic test. This was achieved by evaluating learning analytics captured through the diagnostic test developed and hosted on Blackboard, the Learning Management System (LMS) at both institutions.

It is anticipated that TES will be able to improve their numeracy skills through participating in the diagnostic test, which encourages self-assessment, self-error identification, and active learning through immediate feedback provided for each question (Blanco, Estela, Ginovart, & Saà, 2009; Metz, 2008). As such, knowledge gained from this research will benefit education program providers that wish to adopt an online approach to support and/or track TES' numeracy capabilities. In the long-term, the provision of a method for improving TES' numeracy skills will benefit schools by having increasingly more numerate teachers educating Australian students.

Theoretical Framework

In 1998, Black and Wiliam conducted a comprehensive review of formative assessment research and discussed the specific significance of the roles of feedback, student goal orientation, self-perception, peer-assessment, self-assessment, teacher choice of assessment task, teacher questioning behaviour, teacher use of tests, and mastery learning systems. Of interest to this study is the element of feedback and skills mastery, which is widely discussed in the literature. For example, while acknowledging that there is evidence to suggest that formative assessments promote student learning in higher education, Yorke (2003) described that the "important determinant of the effectiveness of formative assessment is the quality of feedback received by learners" (p. 482). Feedback and the other factors that Black and Wiliam (1998) outlined can be considered as the framework for Assessment for Learning (AfL). According to Berry and Kennedy (2008), AfL enables students to make the decisions that matter most by allowing them to gain continuous information about their learning, including identifying where they are succeeding and where they should focus efforts for improvements, and determining the strategies they need to improve. This work extends on traditional AfL by taking an online approach, which has been reported to have a positive effect on students' learning and future assessment results (Blanco et al., 2009; Metz, 2008). Studies have also reported that students performed better in assessments when coupled with online diagnostic tests (DeSouza & Fleming, 2003; Fletcher-Flinn & Gravatt, 1995), an effect attributed to more consistent and better quality of instructions provided as well as the opportunity for students to develop mastery of the skills assessed. As such, this research adopts the AfL framework of Black and Wiliam (1998) and extends on it by taking an online approach to develop and evaluate the benefits of online diagnostic tests as an AfL tool to improve TES' numeracy capability.

Methodology

Diagnostic Test

The Literacy and Numeracy Test for Initial Teacher Education (LANTITE) Assessment Framework (ACER, 2017) was used as an external objective measure to inform the style, content, and difficulty of the test items in the Diagnostic Test. Specifically, the LANTITE Assessment Framework's prescribed target proportions for levels of difficulty, and process

and context domains were applied to the Diagnostic Test. There were 270 questions developed, including multiple choice, true/false, and fill-in-the-blank (including short response, matching questions with answers, and numerical calculation questions).

Each question was assigned to one of three test categories, according to their content strand (Number and Algebra [N&A], Measurement and Geometry [M&G], or Statistics and Probability [S&P]). Within these categories, sub-pools were created according to the mathematics topic that the question assessed. A fourth Non-Calculator [NC] test category was also created, with questions covering content from all three content strands. The test consists of 40 randomly selected questions, ten from each of the four categories, with a specified number of questions randomly drawn from each topic. Although it is possible that students might see the same question across different attempts, given the volume of questions in the pool, there is a low chance that this will occur. This meant students received the same spread of questions but were exposed to different questions on each test attempt and the distribution of topics are aligned with the LANTITE Assessment Framework. A key component of the test design is the feedback with worked solutions for every question. This encourages self-assessment and supports AfL.

Data Collection and Analysis

Learning analytics were collected through Blackboard LMS at both institutions. For every attempt, data included the questions displayed, students' responses and the score given for each question. Purposive (criterion) sampling was used for this study in order to determine commonalities amongst students who showed considerable improvements over a number of test attempts. The selected sample satisfied the following conditions: (1) Only genuine attempts were selected (defined as attempts with at least 32 out of 40 questions answered), (2) Students who had three or more genuine attempts, and (3) Improved by at least 10% between first and final attempt. Overall, 35 students satisfied all these conditions.

Data was analysed using GraphPad Prism (version 8.0.1). Students' performance in their first and final attempts were assessed using a Mann-Whitney non-parametric *t*-test to determine if there was statistical significance (Figure 1).

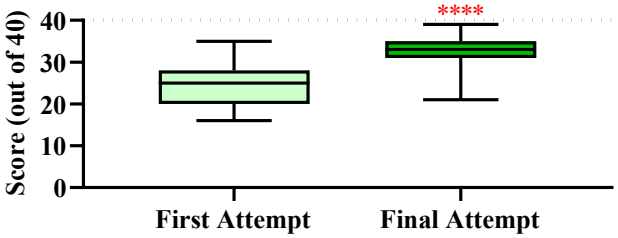

Figure 1. Students' performance in the first and final attempt in the Diagnostic Test. ****$p<0.0001$.

An Ordinary one-way ANOVA with Dunnett's multiple comparisons test was used to determine statistical significance between the mean performance of each attempt with the mean of the first attempt (Table 1). Test categories and topics were assessed using a two-way ANOVA with the Bonferroni's *post-hoc* test, which compares the mean of the first and final attempts within each category or topic (Figure 4 and 5). Results were considered statistically significant where $p<0.05$.

Table 1
Statistical analysis between attempts

Attempt	Mean	SD	N	Mean difference from attempt 1	P value
1	24.51	4.80	35	-	-
2	27.92	4.82	35	3.41	0.0218*
3	29.38	5.05	35	4.87	0.0001*
4	31.11	4.62	28	6.6	<0.0001*
5	30.48	3.68	19	5.97	<0.0001*
6	31.72	3.23	14	7.21	<0.0001*
7	29.19	4.67	11	4.68	0.0358*
8	32.67	1.87	6	8.16	0.0008*
9	32.00	1.83	4	7.49	0.0224*
10	31.75	4.43	4	7.24	0.0311*
11	33.00	1.74	3	8.49	0.0237*
12	31.34	3.06	3	6.83	0.1364
13	28.00	-	1	3.49	-
14	32.00	-	1	7.49	-

Note. * indicates statistical significance.

To frame the analysis of the data and subsequent discussion of findings, the following general questions were investigated: (1) What is the extent of improvements made in the overall test results? (2) What are the most common areas of improvements? (3) What are the areas that require further development?

Findings

Diagnostics Test Performance

Initially, we compared students' performance in their first and final attempt to ensure that the sample captured by the criteria in our purposive sampling was statistically significant. Data from students' performance in the Diagnostic Test showed that the mean for students' first attempt was 24.51±4.80 (mean±SD, out of 40) compared to 32.29±3.99 in the final attempt. Similarly, the median (25 vs. 33), mode (25 vs. 35), minimum (16 vs. 21) and maximum (35 vs. 39) were all higher in the final attempt compare to students' first attempt (Figure 1). Overall, students' performance in the final attempt was significantly higher compared to their first attempt (p<0.0001). Between their first and final attempts, 15 out of 35 students improved by 8 points or more (out of 40). Of these students, ten improved by 25% or more in the test between their first and final attempt. The greatest improvement amongst this cohort was achieved by one student who improved by 42.5%.

In addition to the first and final attempts, test scores were also collected for the other attempts that the students made. Our result shows that the majority of students attempted the Diagnostic Test up to five times (n=21). Eleven students attempted the test between six to ten times and three students attempted the test more than ten times (Figure 2, column). In light of our first research question, we sought to clarify whether the students' final attempt marked their highest performance, and if not, which attempt it was. More than half of the students performed their best in their final attempt. An additional 26% of students achieved their highest result in their penultimate attempt. Our data also shows that students who attempted the test only three times consistently performed their best in their final attempt (Figure 2, cross). When we compared the number of times each student attempted the test with the maximum score they achieved, there appears to be no observable trend. Therefore,

similar maximum results were achieved by students (mean=34, SD=3.23), irrespective of the number of attempts made (Figure 2, line). Further analysis to determine if there are any correlations between the total number of attempts, the attempt that achieved the maximum score, and students' maximum score showed that there are no significant correlations between these variables (data not shown).

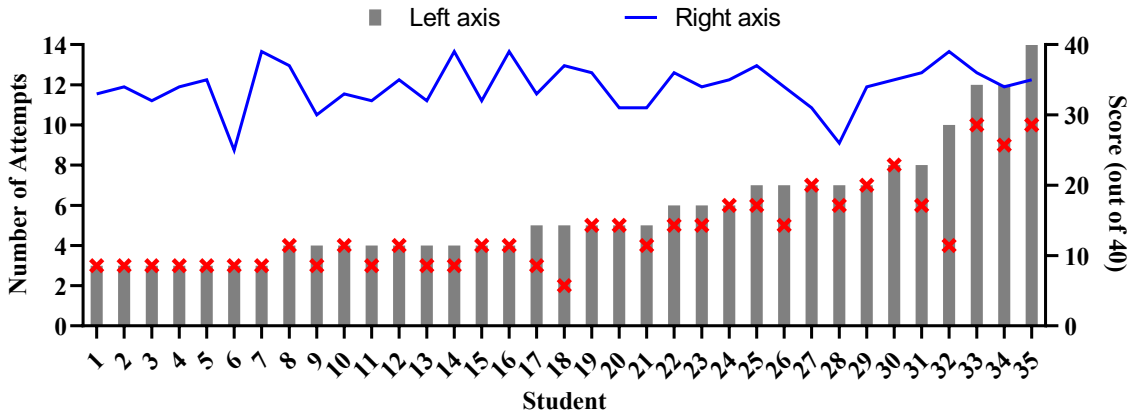

Figure 2. The total number of attempts made (column, left axis), the attempt with the maximum score (cross, left axis), and students' highest score (line, right axis).

To better understand students' learning progression through using the Diagnostic Test as an AfL tool, we evaluated trends between individual attempts. Our data shows that there was progressive improvement with repeated use of the test (Figure 3). The highest rate of improvement occurred within the first four attempts, plateaued by the 8th attempt (mean diff. of 8.15) and reached a peak by the 11th attempt (mean diff. of 8.48). Analysis between the attempts shows that there was a statistically significant improvement in all attempts up to and including the 11th attempt when compared with the first attempt (Table 1). Given that there were limited data points from the 9th attempt (n=4) onwards, we contend that changes past this point should be disregarded.

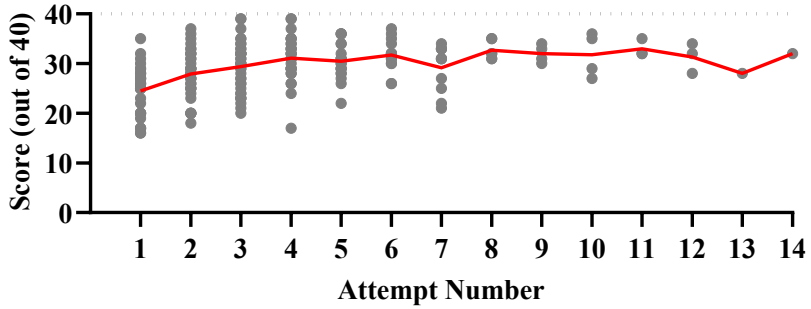

Figure 3. Scores from individual attempts. Line indicates the mean.

Scores in Test Categories

To address our second and third research questions on the areas that improved and areas that need development, we evaluated students' performance in each test category (N&A, M&G, S&P, and NC). Our analysis shows that students' mean in their first attempt was 6.514 (out of 10) in N&A, 5.971 in M&G, 5.571 in S&P, and 6.457 in NC. The mean difference in score between the first and final attempt was between 1.80 and 1.86 for N&A, M&G, and NC, and was 2.271 for S&P. Therefore, whilst S&P was the lowest performing category for students' first attempts, it was also the category with the highest improvement

in students' final attempts. There was no statistical significance between different categories for both first attempt and final attempt. When comparing results between students first and final attempts, we observed a statistically significant improvement in all four categories (p<0.0001 for all categories). We also noted that the spread in the students' final attempt was less in N&A compared to the other three categories (Figure 4). Furthermore, the only category in which any student achieved full marks in their first attempt was NC. In contrast, full marks were achieved in all categories in their final attempt (Figure 4).

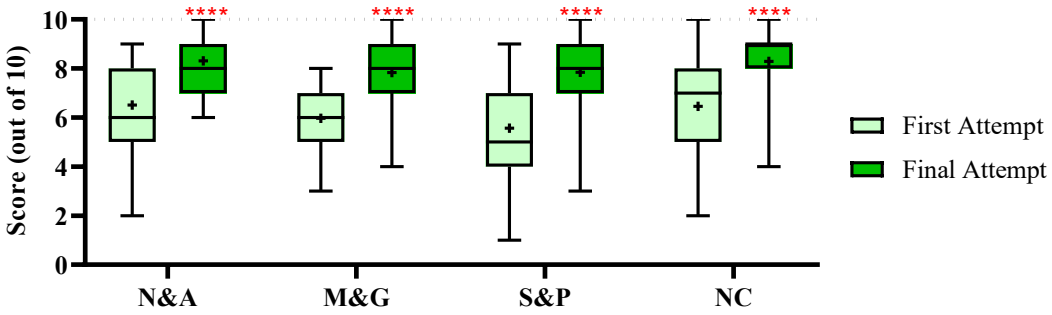

Figure 4. Students' performance in the first and final attempt across the four test categories. + indicates the mean. ****p<0.0001.

Scores in Content Areas

We further explored the students' performance between their first and final attempt by evaluating changes at the content area level. Scores for each topic were tallied and expressed as a percentage of the total number of questions displayed for that topic. Our data shows that improvement was achieved in all content areas assessed (Figure 5). The most statistically significant improvement was in decimals and combinations (p<0.001 for both), followed by probability (p<0.01) and then fractions (p<0.05).

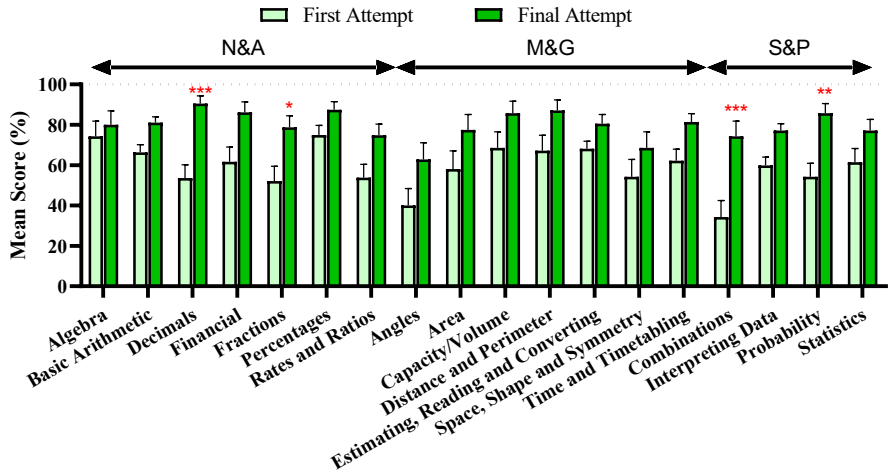

Figure 5. Students' performance in the first and final attempt across content areas. Error bar indicates SEM. *p<0.05, **p<0.01, ***p<0.001.

Discussion and Conclusion

There is a unanimous desire amongst stakeholders for teachers to possess a high level of personal literacy and numeracy, especially since these qualities have been identified to be essential for effective classroom teaching (Allington & Johnston, 2000). Therefore, it is

critical for initial teacher education program providers to have knowledge of their TES' literacy and numeracy capabilities to ensure they meet teaching standards as well as have mechanisms in place to support TES in developing these requisite skills in order to become effective classroom teachers.

In this study, we showed that online diagnostic tests can help track TES' numeracy skills (Figures 1-3). Although there was some fluctuation in the mean between attempts, the overall trend in our result shows that repeated attempts in the Diagnostic Test was associated with improved student performance, which continued to improve even after the 8th attempt (Table 1). Whilst repeated attempts improved students' performance, there was no correlation between the number of attempts and the maximum score students attained. A possible explanation for this is that individual students are improving as they continue to use the Diagnostic Test but at different rates. For example, one student might take three attempts to achieve a personal goal compared to another student who might take ten attempts to achieve the same level. This would also align with our observation that more than three-quarter of students achieved their personal best in their final or penultimate attempt (Figure 2). Therefore, not only is the Diagnostic Test a useful form of AfL, it also allowed students to self-assess the level of support needed. Indeed, our data shows that 80% of students' final attempts achieved a score of 30 or more (out of 40).

Further analysis of the four test categories shows that there was a significant improvement in performance in all three mathematical content strands (N&A, M&G, and S&P) as well as NC (Figure 4). The biggest improvement occurred in S&P, which had the lowest mean in students' first attempts. This result differs to that reported by Afamasaga-Fuata'i et al. (2008), who showed that students were more likely to improve in less difficult questions. Future studies could consider exploring the types of questions (e.g., multiple choice and short answer) and the literacy demands of questions to determine if these factors influence students' performance and progress. In addition, a breakdown of the test into individual content areas showed that whilst there was a trend of improvement in all topics, significant improvement was made in decimals and fractions in the N&A strand, and combination and probability in S&P (Figure 5). There was no significant improvement in any topic in the M&G strand.

A potential limitation of this study is the possibility that students were improving from memorising solutions given in the feedback and/or through the repeated attempts of the test. However, given the volume of the pool of questions, this is unlikely to be the main factor. A possible explanation for the improvement is that students engaged in additional support and used the Diagnostic Test as a benchmark for the numeracy level required. It would also explain the motivation for students to attempt the Diagnostic Test several times. Determining the factors that led to students' numeracy improvement is an area for further investigation.

Overall, this study shows that online diagnostic tests can be used as a sustainable form of AfL to track TES' numeracy skills improvement. The incorporation of detailed feedback in questions promotes self-assessment, and active and independent learning, through repeated attempts of the test.

References

Afamasaga-Fuata'i, K., Meyer, P., Falo, N., & Sufia, P. (2008, 2007). *Future teachers' developing numeracy and mathematical competence as assessed by two diagnostic tests.* Paper presented at the Australian Association for Research in Education, Conference.

Allington, R., & Johnston, P. (2000). What do we know about effective fourth-grade teachers and their classrooms? In C. Roller (Ed.). Learning to Teach Reading: Setting the Research Agenda. Newark, DE: International Reading Association. Retrieved from https://www.albany.edu/cela/reports/allington/allington4thgrade13010.pdf

Australian Council for Educational Research (ACER). (2017). Literacy and numeracy test for initial teacher education students assessment framework. Retrieved from https://teacheredtest.acer.edu.au/files/Literacy-and-Numeracy-Test-for-Initial-Teacher-Education-Students-Assessment-Framework.pdf

Australian Curriculum and Reporting Authority (ACARA). (n.d.). General capabilities. Retrieved from https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/

Blanco, M., Estella, M. R.; Ginovart, M., Saà, J. (2009). Computer Assisted Assessment through Moodle Quizzes for Calculus in an Engineering Undergraduate Course. *Quaderni di Ricerca in Didattica* (Scienze Matematiche), *9*(2), 78-84.

Berry, R., & Kennedy, K. J. (2008). *Assessment for learning*. Hong Kong: Hong Kong Univeristy Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

DeSouza, E., & Fleming, M. (2003). A comparison of in-class and online quizzes on student exam performance. *Journal of Computing in Higher Education, 14*(2), 121-134. Retrieved from https://link.springer.com/article/10.1007/BF02940941

Fletcher-Finn, C., & Gravatt, B. (1995). *The efficacy of computer assisted instruction (CAI): A Meta-analysis.* Retrieved from https://journals.sagepub.com/doi/abs/10.2190/51D4-F6L3-JQHU-9M31

Metz, M. (2008). *A study of high school mathematics teachers' ability to identify and create questions that support students' understanding of mathematics.* (Doctoral Disseration). Retrieved from http://d-scholarship.pitt.edu/9494/1/Metz2.ML.8.2007_final.pdf

Linsell, C., & Anakin, M. (2012). Diagnostic assessment of pre-service teachers' mathematical content knowledge. *Mathematics Teacher Education and Development, 14*(2), 4-27.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education,45(4),* 477-501. https://doi.org/10.1023/A:1023967026413