CASE STUDY

Development and analysis of a Numbas diagnostic tool for use in a mathematics refresher program

Donald Shearman, Mathematics Education Support Hub, Western Sydney University, Sydney, Australia. Email: d.shearman@westernsydney.edu.au
Shatha Aziz, School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia. Email: s.aziz@westernsydney.edu.au
Jim Pettigrew, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia. Email: j.pettigrew@unsw.edu.au

Abstract

We describe the development and analysis of an online diagnostic tool implemented in the Numbas e-learning system and used in an Australian university mathematics refresher program. Following the rapid transition to online delivery of the refresher program caused by COVID-19, the diagnostic instruments and methods used within the pre-pandemic, in-person, version of the program were lost. In 2022, we undertook to revive them in a way that would honour their original diagnostic purpose but offer a more sophisticated approach utilising the Numbas diagnostic exam type. Improvement of the tool after its initial deployment has involved the use of Rasch-based item analysis and recursive refinement of the knowledge map underlying the items.

Keywords: diagnostic tool, mathematics refresher program, knowledge map, item analysis, Numbas.

1. Introduction

For more than 15 years prior to the COVID-19 pandemic, the Mathematics Education Support Hub (MESH) at Western Sydney University (WSU) offered a series of mathematics and statistics refresher workshops for incoming - primarily engineering - students. The workshops were designed to prepare students for the mathematical requirements of their studies, and included content covering topics in basic algebra, trigonometry, statistics, and calculus. Five workshops in these areas (two for algebra) were taught face-to-face over a 3-week period. Students taking the algebra and trigonometry workshops were given pre- and post-tests to measure their skill improvement. The pre-test was also used to place students into equal-ability groups, and so served as a crude diagnostic classifier. All of the lesson and exercise material for each workshop was contained in a printed booklet provided to students at the start of the workshop (this material was also available online during the year for out-of-class learning). These provided a highly structured curriculum for each workshop.

Due to the pandemic, it was necessary to make the workshops available for online study at very short notice. To facilitate a quick transition, we used existing external videos and web resources to present the content of each workshop, along with in-house skill development problems built using the Numbas e-learning system. While the pre- and post-tests were still available online, the pre-test's diagnostic function was largely lost. This limited the capacity of MESH educators to classify students according to skill level, and thus inhibited their work in supporting students' preparation for university study.

2. Plan to develop a diagnostic tool

After three years of online operation, anecdotal evidence indicates that many students find the amount of content covered in the Maths Start workshops overwhelming. They are unable to determine for themselves which selection of lessons and exercises to study in order to improve specific skill deficiencies in topics required for their study. In response, we have developed an online diagnostic

tool using Numbas. The tool is designed to identify for each attemptee a selection of topics where mastery criteria have not been met, and hence where further skill development should be focussed.

We have built and deployed diagnostic tools for four of the five Maths Start workshops. We have chosen to refer to our implemented instrument as a diagnostic 'tool' rather than 'test', as the former term better captures the essential enabling function of the instrument. This paper focuses on the first of these to be built, for the Algebra 1 workshop, in 2023.

3. Diagnostic tool design

The definition of a 'diagnostic tool' adopted for this study is an instrument that enables students to identify their areas of skill mastery/non-mastery within a given knowledge domain (Rylands and Shearman, 2022). When considering the design parameters of a suitable diagnostic tool, we were guided by two broad factors: 1. the mechanism by which the tool would identify a students' areas of mastery/non-mastery; and 2. the amount of time it would take a typical student to complete a full attempt (a measure of the diagnostic efficiency of the tool).

A form of computer adaptive test was chosen to meet these criteria. This type of test is designed to update a latent measure of the attemptee's ability while they are answering questions. If they answer a question correctly, they are only presented with more difficult questions in the hierarchically-ordered set of questions to which it belongs (or taken to another set if the most difficult question in the current set has been mastered). Conversely, if they answer a question incorrectly, they are stepped through easier questions in the hierarchy until mastery is met or the questions in the set are exhausted. Attempt times for these tests are usually shorter than for traditional tests because attemptees only complete a subset of the full question set. In tailoring the questions presented to the individual attemptee according to difficulty, computer adaptive tests aim to find the attemptee's threshold skill level and give a more accurate assessment of their knowledge (Meijer and Nering, 1999).

There are many subvariants of computer adaptive test designs; the one chosen for this study relies on an underlying 'knowledge map' consisting of hierarchically-arranged concepts and associated topics. Throughout all of the development and analysis described in this case study, we have assumed that knowledge of a concept is demonstrated by way of mastery of a specific skill (or skills). In practical terms, the implemented design is such that:

- if a question is answered correctly, it is assumed that any question about a concept on which that question relies would also be answered correctly, so these questions are marked as correct;
- if a question is answered incorrectly, it is assumed that any question about a concept which relies on the concept being assessed would also be incorrect, so these questions are marked incorrect.

In developing the diagnostic tool for this study, the first stage was the construction of the abovementioned knowledge map, which acted as a formal network of the necessary connections between the concepts and associated topics being assessed.

4. The knowledge map

The knowledge map used in this study was developed by reviewing the sections and subsections of the Algebra 1 workshop and making links between those exhibiting clear and logical conceptual connections. The result, a directed acyclic graph (DAG), was reviewed by other members of the MESH team and opinions on the included topics (nodes) and their hierarchical connections (directed edges) were discussed and resolved. A section of the final map for the Algebra 1 content is shown in figure 1.

This was created using the graph visualisation application Gephi. Note that only the numbered items are topics; the unnumbered items are learning objectives, included in the map for convenience (as they show the grouping of topics by learning objective).

Once the construction of the knowledge map was complete, the next task was to assign a question to each of its nodes.

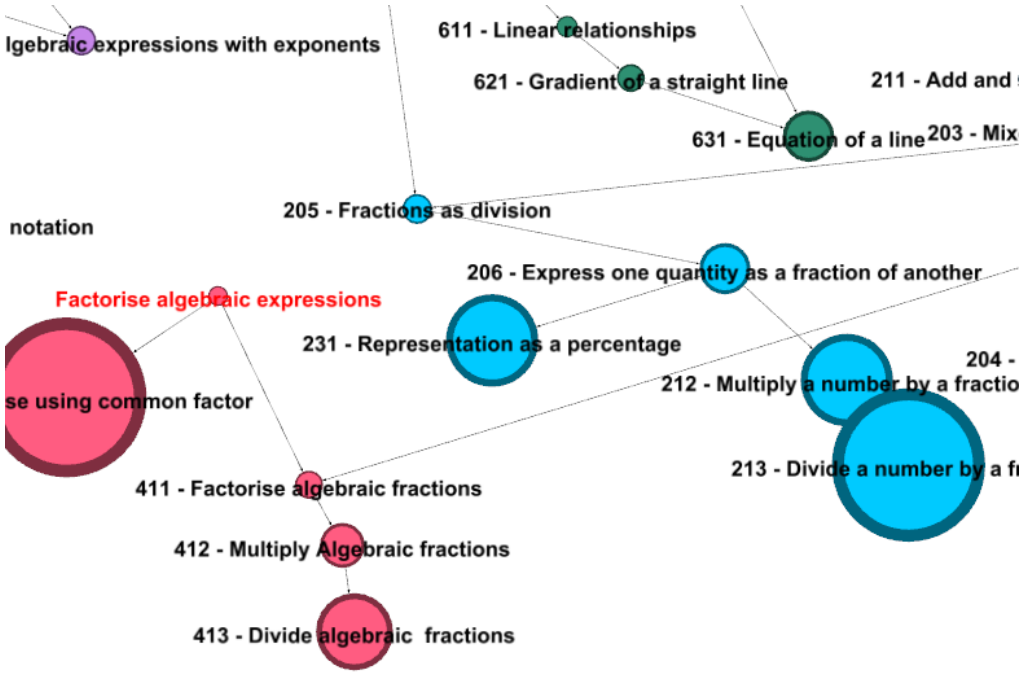

Figure 1. A section of the knowledge map for the Algebra 1 workshop.

5. Implementing the diagnostic tool using Numbas

The system used to implement the Algebra 1 diagnostic tool is the Numbas diagnostic exam (more precisely, the diagnostic mode of the generic Numbas exam object). At the heart of this system is an algorithm, known as DIAGNOSYS, that acts on questions whose underlying concepts are grouped within 'learning objectives'. Individual concepts within learning objective groups are classed as 'topics'. For example, in the Algebra 1 diagnostic tool, we have defined 'Factorise algebraic expressions' as a learning objective and 'Factorise algebraic fractions' and 'Multiply algebraic fractions' as topics within this learning objective (see figure 1). The system allows each topic to be assigned a question, hierarchically linked to other topics, and grouped within a specific learning objective.

DIAGNOSYS conforms to the general principles of computer adaptive test design: the 'next' question presented to the attemptee (except the first) depends on their responses to previous questions; no questions are presented whose underlying concepts are assumed by the algorithm to have been mastered or unable to be mastered. The effect of the hierarchical arrangement of questions in a Numbas diagnostic exam is that:

- an incorrect answer to a question causes the system to mark all harder questions on the same hierarchical path as incorrect;
- a correct answer to a question causes the system to mark all easier questions on the same hierarchical path as correct.

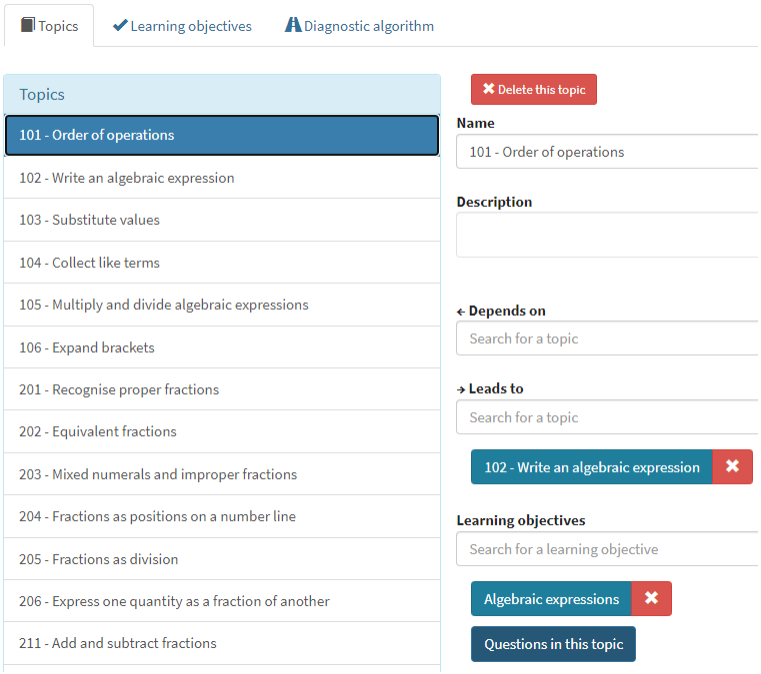

Figure 2. The Numbas authoring interface for topic linking.

In practice, after almost 12 months of use, this has reduced attempt times for the Algebra 1 diagnostic tool. We believe this has led to higher completion rates, and, ultimately, enhanced MESH's ability to support incoming students unsure of how to tackle their skill refreshment.

The knowledge map was written into DIAGNOSYS by creating topics grouped within learning areas and the necessary connections between these topics. Topics were paired by using either the 'leads to' or 'depends on' directed connections (see figure 2). This task was complex as extra care was needed to ensure faithful translation of the knowledge map and accurate reconstruction of the DAG in Numbas.

To complete the construction of the tool, each topic in Numbas was assigned a question. This process was simplified by the fact that the Algebra 1 curriculum, whose structure has been refined over two decades by educators within MESH, coexists with a set of Numbas questions (used in non-diagnostic quizzes prior to the development of the diagnostic tool). For each topic, a search was conducted for a suitable existing question. If one was found, it was either adopted unchanged or modified. In exceptional cases, new questions were created in the absence of suitable existing candidates.

In the second half of 2022, regular review meetings addressed question suitability and the composition of the knowledge map (whose development was recursive). Other educators within MESH who did not participate in these review meetings trialled early versions of the Algebra 1 tool; their feedback allowed for corrections and refinements. The process of review and refinement was repeated until January 2023, when we deployed the final version of the tool on WSU's learning management system.

6. Deployment

The Algebra 1 diagnostic tool was loaded onto a dedicated Learning Tools Interoperability (LTI) server and made available to all WSU students on the Maths Start site. The tool is still 'live' and as of mid October 2023 has been attempted over 800 times by 599 unique students. The first attempt was on 1 February 2023.

7. Analysis

Following deployment of the diagnostic tool in 2023, an analysis of the response data was conducted to determine whether the questions were functioning according to their design specifications. Particular attention was given to the relationship between the questions and the knowledge map, and the extent to which the logical connections between questions were validated by the patterns in the response data.

The analysis method was to compare two sets of question responses (hereafter, we'll refer to questions interchangeably as items): 'raw' responses for scored student attempts, and 'implied' responses for unattempted items whose scores were implied by DIAGNOSYS and its application of the knowledge map. In each case, correct and incorrect responses were assigned a score of 1 and 0 respectively. The implied scoring works as follows: if an attemptee answers an item correctly (call it item A, for instance) and items B, C and D are set as easier than item A, then the tool assumes the attemptee will also answer items B, C and D correctly - and therefore not present these items to them. Figure 3 gives an example of these relationships for four connected items in the knowledge map.

The knowledge map was built in Gephi such that items (nodes) connected (via edges) to other items but lying below them in the graph were dependent: the knowledge and skills required to correctly answer item A (below) include and extend those required to correctly answer items B, C and D (above). Thus the implied scoring algorithm worked by assigning a score of 1 to all easier items above a correctly-answered item and connected to it in the knowledge map (the 1s 'float'), and assigning a score of 0 to all harder items below an incorrectly-answered item and connected to it in the knowledge map (the 0s 'sink').

The response data was exported from the Numbas LTI server as a JavaScript Object Notation (JSON) file and imported into the R statistical analysis program using the 'jsonlite' package. The knowledge map was exported from Gephi as a Graph Exchange XML Format (GEFX) file for manipulation in R.

The attempts for which each of the 50 items received an implied score of 0 or 1 were collected as the implied-scored response data. There were 412 such attempts, giving a 412 by 50 data frame in R. Their associated raw-scored response data was stored in an identically-dimensioned data frame, populated by 0s and 1s as well as NAs for the missing scores (see figure 4).

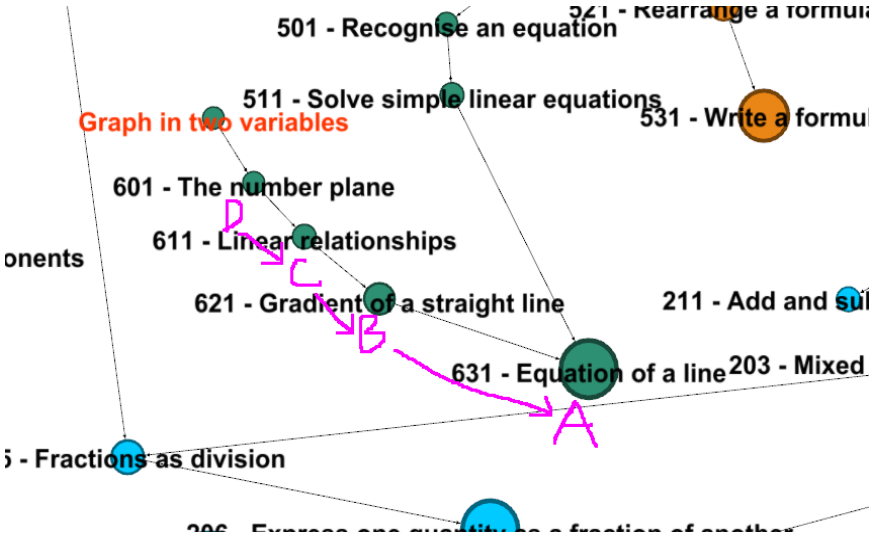

Figure 3. Graphical relationship between items A, B, C and D. Item D is the easiest of the four, C the next easiest, and so on. A is the hardest item in the set.

	Q1_score	Q2_score	Q3_score
1	NA	0	0
2	NA	0	0
3	NA	NA	NA
4	NA	NA	NA
5	NA	1	1
6	NA	NA	NA
7	NA	0	NA
8	NA	0	0
9	NA	1	0

	Q1_implied_score	Q2_implied_score	Q3_implied_score
1	1	0	0
2	1	0	0
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	0	1
8	1	0	0
9	1	1	0

Figure 4. Snippets of the raw- and implied-scored data frames used in R.

Rasch modelling was applied to the raw-scored and implied-scored response data separately to determine, first, item difficulties for the items in each case and then other properties such as item discrimination (point-biserial correlation), and various statistical measures of item and model fit (see Wu, Tam and Jen, 2016). It should be noted that this use of Rasch modelling is unconventional for at least two reasons: 1. the dichotomously-scored data is incomplete in that missing and implied scores represent 'responses' to items that were not presented to students; and 2. randomisation of variables in the tool meant that different students attempted (slightly) modified versions of the same items. The authors make no claim as to a rigorous, scientifically-robust use of this modelling, but note that it has provided an alternative methodology to Classical Test Theory in assigning measures of difficulty to each of the 50 raw-scored and implied-scored items. Furthermore, the ensuing comparison of item rankings by difficulty (raw-scored versus implied-scored) drew attention to a number of potentially dysfunctional items that, upon close inspection, were revealed to be so.

The raw-scored and implied-scored items were ranked (separately) from least to most difficult. For example, the three least difficult raw-scored items were questions 22, 24 and 25, while the three least difficult implied-scored items were questions 7, 1 and 16. The unsigned difference in ranking was calculated for each item, and this yielded a list of 10 items where the difference was greater than or equal to 15. (Most rankings were close, for example 35 of the 50 items had an unsigned difference of 8 or less.) Our analysis proceeded by assuming that these items were anomalous in the sense that they were malfunctioning according to their design, or irregular due to unexpected attemptee behaviour or item interaction.

The final phase of the analysis was to use the results of the Rasch modelling to guide a critical inspection of the anomalous items. We shall present two examples of the findings of this work.

7.1 Question 11 (raw-scored rank: 39; implied-scored rank: 5)

This question is shown in figure 5. It aims to assess skill in dividing one fraction by another. According to the raw-scored responses, it was ranked as the 12th hardest item; but according to the implied-scored responses, it was ranked as the 5th easiest (giving a rank difference of 34). What could explain this difference?

We believe that the answer relates to part (b) of the question, mastery of which would require arithmetic skill beyond the direct division assessed in part (a). As figure 6 illustrates, the attemptees who got this question - appearing as ‘205 - fractions as division’ - wrong would not have been presented with any of the questions below it on the hierarchical path (note that 56 of the 412 attempts were in this

category). This means that they would have received the implied score of 0 for all of these 'harder' questions (recalling that the 0s 'sink'). In particular, an incorrect raw response to Question 11 would result in an implied incorrect response to Question 15 (identified as '213 - Divide a number by a fraction'). Conversely, a correct raw response to Question 15 would result in correct implied responses for all of the 'easier' questions above it in the path (recalling that the 1s 'float').

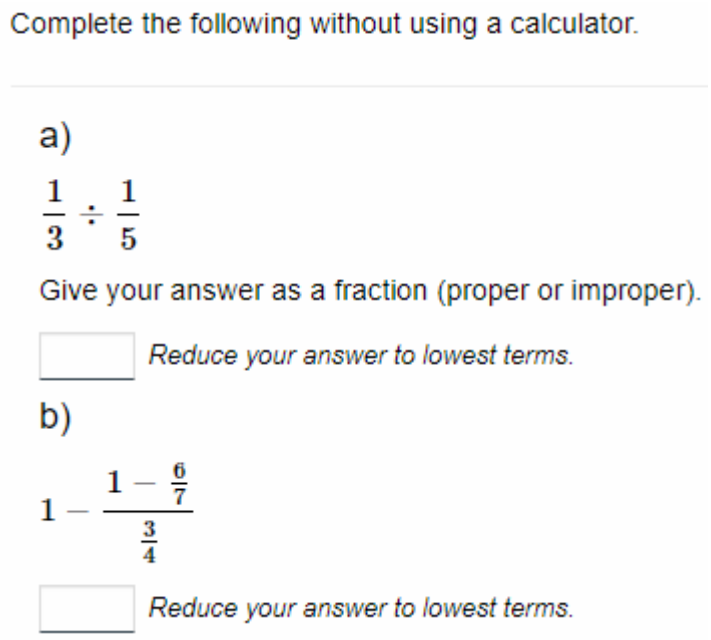

Figure 5. Question 11 of the diagnostic tool (identified in the knowledge map as '205 - Fractions as division').

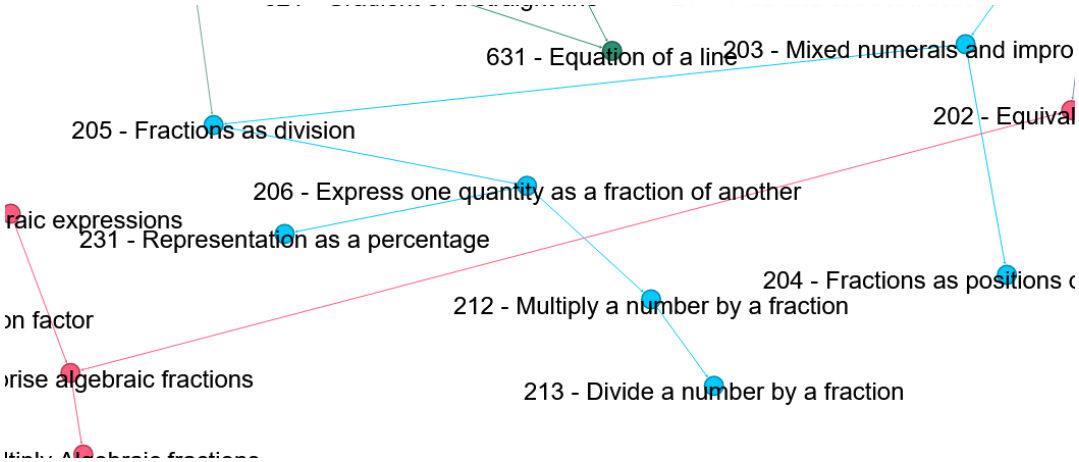

Figure 6. A snippet of the knowledge map, showing the location of Question 11 (identified as '205 - Fractions as division'), and its connection to harder questions below and easier questions above it.

Question 15 is shown in figure 7. Is it reasonable to assume that attemptees who answered this question correctly would be able to answer Question 11(b) correctly? We believe not and so have adjusted Question 11 to better align it with its underlying concept and location in the knowledge map.

If the above analysis is correct, then the reason Question 11 was ranked as significantly easier under implied scoring compared to raw scoring is that many attemptees (167 of 412) were presented with Question 15, answered it correctly and subsequently were not presented with any of the questions higher than it on the path (in particular Question 11 - see figure 6). All of these higher items, assumed easier, would have been given an implied score of 1.

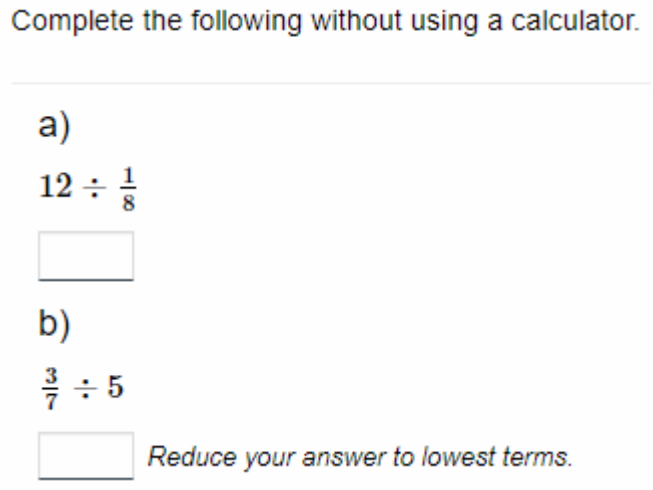

Figure 7. Question 15 of the diagnostic tool (identified in the knowledge map as '213 - Divide a number by a fraction').

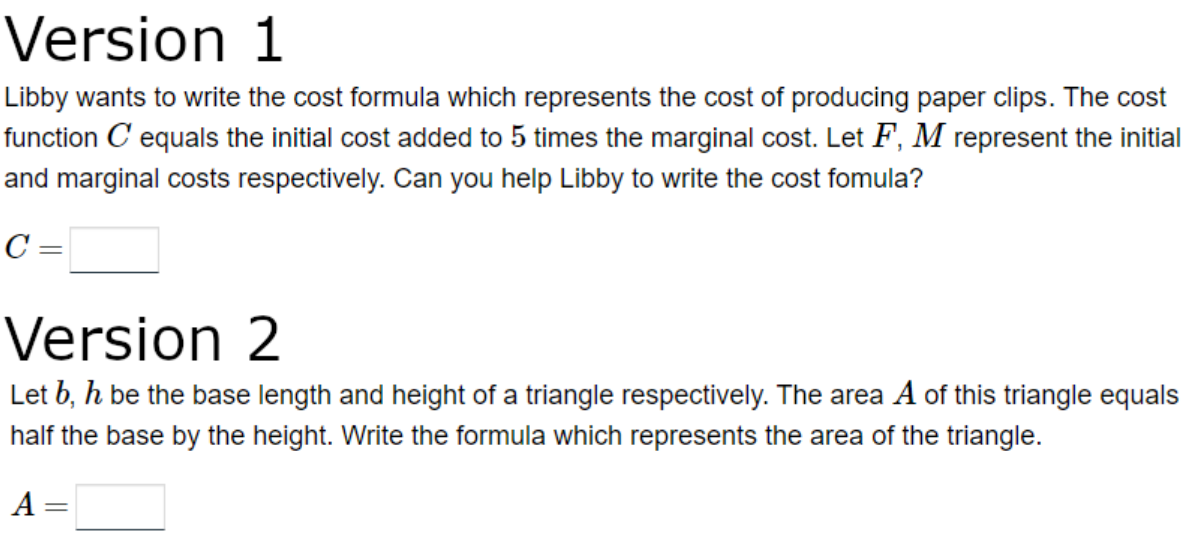

Figure 8. Both versions of Question 46 of the diagnostic tool (identified in the knowledge map as '531 - Write a formula').

7.2 Question 46 (raw-scored rank: 15; implied-scored rank: 42)

This question is shown in figure 8 (both versions; noting that either is presented to attemptees at random). It aims to assess skill in translating a worded description of a formula to its symbolic, mathematically-notated form. According to the raw-scored responses, it was ranked as the 15th easiest item; but according to the implied-scored responses, it was ranked as the 9th hardest (giving a rank difference of 27).

We propose that the reason Question 46 (appearing as '531 - Write a formula' in the knowledge map - see figure 9) was ranked as significantly harder under implied scoring compared to raw scoring is that many attemptees would have been presented with Question 2 (appearing as '102 - Write algebraic expressions', see figures 9 and 10), answered it incorrectly and subsequently be given a mark of 0 for all of the 'harder' questions below it in the hierarchical path (in particular, Question 46). This means that they would have received the implied score of 0 for Question 46 (assumed to be harder). But in our judgement, version 2 of Question 46 is easier then Question 2, and this is reflected in a disproportionately high number of correct raw-scored responses.

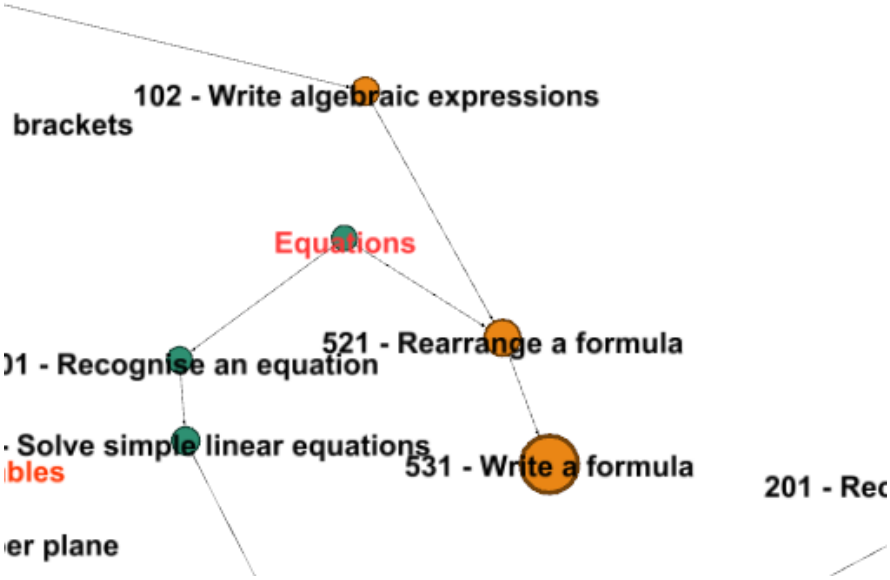

Figure 9. A snippet of the knowledge map showing the location of Question 46 (identified as '531 - Write a formula'), and its connection to easier questions above it.

H is used to represent a standard hourly rate of pay in dollars.

Which one of the following expressions would represent the amount of pay earned for 10 hours work, where 3 of those hours were paid at one and a half times the standard rate?

○ $13H$ ○ $11.5H$ ○ $\frac{10}{1.5}H$ ○ $14.5H$

Figure 10. Question 2 of the diagnostic tool (identified in the knowledge map as '102 - Write algebraic expressions').

8. Conclusion

In this case study we have described the process by which we developed and analysed an online diagnostic tool for deployment within a university mathematics refresher program. The tool's diagnostic power is derived from a knowledge map defining and connecting key concepts in a basic algebra knowledge domain. Use of the tool has enabled attemptees to determine their level in this knowledge domain in a manner that is more efficient and personalised than non-adaptive alternatives (featuring banks of multiple-choice questions, for example). From the educator's perspective, the tool has allowed

for the optimised delivery of support services and learning resources that are tailored to the individual attemptee based on their location in the knowledge domain. An important difference between our tool and other computer-adaptive diagnostic systems is that 'next question' selection is based on the attemptee's trajectory through a knowledge map rather than a real-time estimate of their ability (matched with questions whose difficulty is on the same logit scale in the case where Rasch and related analysis methods are used).

The tool has been built using the Numbas e-learning system, whose essential DIAGNOSYS algorithm operates according to principles of computer adaptive test design. The refresher program's existing curriculum provided an unrefined hierarchical structure for the mathematical concepts and associated skills covered in the program, and this led to the creation of the abovementioned knowledge map defining their necessary connections. The tool was built using a Numbas exam object (diagnostic mode), which was designed carefully to honour the logical connections in the knowledge map. A combination of new and existing Numbas questions was used to populate the exam. Analysis of the item responses for 826 attempts of the tool has revealed significant differences in the difficulty ranking of some raw-scored and implied-scored questions, and this drew attention to anomalies in the design of a subset of questions and their hierarchical relationships in the knowledge map. Addressing these anomalies has involved modifying questions and reconfiguring the knowledge map. In this way, improvement of the tool has been recursive.

Though it has not been developed as an instrument for formative or summative assessment per se, the design of the diagnostic tool would allow for ready applications in these ways. The Numbas system allows for exams to be switched from diagnostic to mastery mode, enabling attemptees to hone their skill with unlimited question attempts. An adapted version of the tool could also be used for delivery of summative assessments. Summative computer adaptive tests are used in many of Microsoft and CISCO's Computer Certification Tests, for example. In broad terms, the idea here is that the test-taker's response patterns allow the system to find their 'point of equilibrium' in settling on the questions – hence concepts – positioned at their level of ability.

An extension of this study would be to compare the existing 'subjectively constructed' knowledge map to one that has been generated using an 'objective' statistical method. The latter would involve administering all of the tool's questions to a sample of students and using one of many (typically naïve Bayesian) approaches to define a network of connections between the concepts underlying the questions. (There are many examples of such approaches in the literature.) We could then compare the 'objective' and 'subjective' knowledge maps and if the differences are minimal adopt the latter as valid for the constrained diagnostic purposes of the tool. Another extension would be to look across the knowledge map and calibrate topics based on the Rasch-determined difficulty of their associated questions. This would add a 'depth' dimension to the knowledge map in introducing a vertical scale to its nodes and edges (some edges being longer than others where the Rasch analysis has determined the difference in the difficulty of the topics they connect is greater, for example).

9. References

Meijer, R. R. and Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement*, *23*(3), pp.187-194. https://doi.org/10.1177/01466219922031310

Rylands L. and Shearman D. (2022). Diagnostic tests: Purposes and two case studies. *MSOR Connections*, 20(3), pp.45-54. https://doi.org/10.21100/msor.v20i3.1281

Wu, M., Tam, H.P. and Jen, T.H. (2016). Educational measurement for applied researchers. *Theory into practice*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5