Using Computerised Adaptive Tests in Assessment for Learning

by

Poh Hua, TAY (tay_poh_hua@seab.gov.sg)
Joyce, LIU (joyce_liu@seab.gov.sg)
Yick Chee, FONG (fong_yick_chee@seab.gov.sg)

Singapore Examinations and Assessment Board, Singapore

Paper presented at

The 44th International Association for Educational Assessment (IAEA)

Annual Conference

Conference Theme: Assessment & Big Data
Sub-Theme: Reporting on progress and achievement using Big Data

Oxford University, United Kingdom
September 9 – 14, 2018.

ABSTRACT

In a self-directed learning environment, students take tests anytime and anywhere to ascertain their mastery of learning. Traditionally, Computerised Adaptive Test (CAT) was used to automatically score and provide an ability estimate for each student. Although these ability estimates are good for summative purposes such as comparing of students or ranking students relative to the cohort, it may not be useful to teachers for formative assessment.

To support teachers in the assessment and reporting of students' progress and achievement in this big data environment, this paper presents an approach, illustrated using primary school Fractions, to transform CAT to provide teachers with both precise results as well as detailed information about students' proficiency in Fractions in the form of profile descriptors. During our prototyping in schools, teachers generally found the reports useful in helping them identify student proficiency level in Fractions as well as customising their interventions to close individual student learning gaps.

Sub-themes: Reporting on progress and achievement using Big Data, Computerised Adaptive Testing, Assessment for Learning

INTRODUCTION

Amongst the recommendations made by the Singapore Primary Education Review and Implementation (PERI) Committee in 2009, Holistic Assessment was highlighted as a key approach to support student learning and development (MOE, 2009). In the years following its implementation, there has been an increasing emphasis on Assessment for Learning (AfL) in schools. Teachers are trained to make use of information gathered from assessments to diagnose student needs, plan follow-up instruction, and give effective feedback to students to move them forward in their learning (Lee, Oh, Ang, & Lee, 2014; Tan, 2017). At the same time, students are also taught the importance of feedback and how to use it to improve the quality of their work.

The definition of AfL has undergone many iterations since its introduction in the UK in the late 1990s through Black and William's seminal work (1998a, 1998b). The original definition by the Assessment Reform Group (the ARG) states that "Assessment for Learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (2002). After many years of research and refinement, there are now many diverse definitions of AfL used in different contexts, but all with a central focus on promoting student learning (Bennett, 2011). Black et al. (2004) provided the following expanded definition:

> "Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of promoting students' learning. It thus differs from assessment designed primarily to serve the purposes of accountability, or of ranking, or of certifying competence. An assessment activity can help learning if it provides information that teachers and their students can use as feedback in assessing themselves and one another and in modifying the teaching and learning activities in which they are engaged. Such assessment becomes "formative

> assessment" when the evidence is actually used to adapt the teaching work to meet learning needs."

The concept of AfL has appeared in a multitude of research studies exploring its potential to facilitate learning and improve student achievement. William and Black (1996) described an assessment cycle involving the elicitation of evidence, which is then appropriately interpreted and acted upon. The cycle repeats naturally as more evidence is generated through the process. In order to maximise the impact of AfL, assessments intended for this purpose need to be carefully conceptualised and implemented, and teachers also need to be equipped with the necessary knowledge and skills to use these assessments effectively (Bennett, 2011).

The use of the Computerised Adaptive Test (CAT) has become increasingly popular over the past decade and extensive research is being conducted in this area. It is well-known that the Graduate Record Examination (GRE) developed and implemented by Educational Testing Service (ETS) is adaptive in nature. More recently, organisations such as Cito (The Netherlands) and the Northwest Evaluation Association (Oregon, USA) have also developed their own suites of CATs to be used in schools nationwide. Other high-stakes tests such as the ACT Test (used for college admissions in the US) and the Korean Medical License Examination (KMLE) are also making preparations for transitioning to the CAT format (ACT, 2016; Seo and Choi, 2018).

CAT presents many advantages, the most well-known being its efficiency and control over measurement precision (Weiss & Vale, 1986). Students sit for a test which is customised (adapted) to their individual abilities, and this eliminates the need for students to answer a large number of questions which may be too easy or too difficult for them. Also, this allows more information to be gathered about a student's level of mastery as compared to traditional linear tests, resulting in a more accurate measurement (Rezaie & Golshan, 2015). CAT lets the student showcase what he knows, not just what he or she does not know, and this is especially useful for diagnostic purposes.

Typically, CAT is used to automatically score and provide an ability estimate for each student. Although these ability estimates are good for summative purposes such as comparing of students or ranking students relative to the cohort, they may not be as useful for AfL purposes. The potential of CAT as an AfL tool is explored in this paper, and it offers the exciting possibility of providing teachers with both precise results as well as detailed information about their students' proficiency.

In the teaching of Mathematics, teachers use a variety of strategies to carry out formative assessment of students' learning (Black et al., 2004; Nortvedt et al., 2016; Anderson and Palm, 2017). Many of these are seen in the classroom where teachers pose questions and provide feedback to students' responses. Often, teachers also make use of the results from school-based assessments (such as class tests and end-of-semester examinations) to elicit evidence of students' mastery in particular topics and plan further actions. Our version of CAT serves as a supplementary resource, with additional benefits (such as saving time on the teacher's part) on top of those mentioned above.

In this paper, we present an approach to make use of CAT to support AfL in schools by helping teachers to gain insight into their students' mastery of a particular topic in Mathematics. Through the use of profile descriptors, we are able to surface students' learning gaps and provide teachers with meaningful information on their students' content mastery (apart from scores alone). Teachers are also guided in their interpretation of the reports to help them to decide on the next steps in instruction and remediation.

APPROACH

In this section, we present a methodology to transform CAT to efficiently and accurately provide school teachers with detailed information about their students' proficiency in any Mathematics Topic in the form of profile descriptors. For illustration purposes, Fractions, which is one of the Mathematics Topic in Singapore Primary Mathematics Curriculum, will be used. This methodology comprises 6 key steps, abbreviated as 'D.E.S.I.G.N.':

1. **D**efine item bank blueprint based on the Learning Outcomes (LOs) stipulated in the Primary Mathematics teaching and learning syllabus;
2. **E**xpedite item development to populate the item bank using an Automatic Item Generation approach;
3. **S**tructure and construct the Topical Continuum by establishing and arranging the LO mastery thresholds in ascending order of demand across grade levels;
4. **I**ntegrate the Computerised Adaptive Testing engine to efficiently and accurately determine whether a student could meet each LO mastery threshold along the Continuum;
5. **G**enerate students' performance by profiles at the individual, class and school levels;
6. **N**otify and provide teachers with a report describing the LOs that each student can and cannot manage to inform customised interventions for individual students or groups of students.

Step 1: Define item bank blueprint based on the Learning Outcomes (LOs) stipulated in the Primary Mathematics teaching and learning syllabus. In Singapore, the topical Mathematics content chart can be found in the Primary Mathematics teaching and learning syllabus document (Ministry of Education, 2012). The content chart on Fractions stipulates the sub-topics in Fractions and their corresponding Learning Outcomes (LOs) from Grade 2 to Grade 6. It guides primary school teachers in their teaching as well as in the assessment of Fractions. The Primary Mathematics curriculum is designed to be spiral in nature. This means that the teaching of concepts at each grade level is built upon students' prior knowledge and skills to allow gradual mastery from one grade level to the next. This systematic approach not only allows students to learn Mathematics topics and skills which are appropriate to their developmental or cognitive stages, but also strengthens students' retention and mastery of the topics and skills acquired.

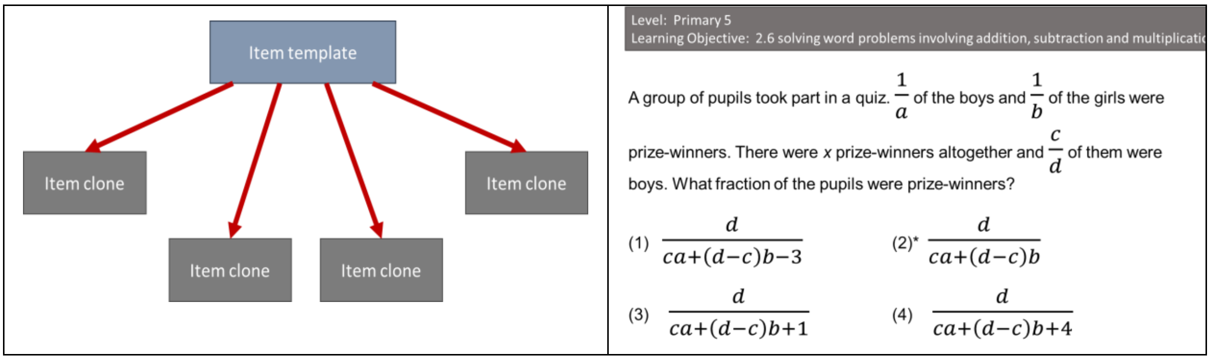
The first step of the methodology is to define the item bank blueprint based on the topical LOs in the teaching and learning syllabus across all levels. This is to ensure that the item bank developed covers the entire content of the particular topic taught in schools. The alignment of the bank blueprint to the syllabus is critical as it will permit assessment of only what students have learnt and provide valid inferences of their topical LO mastery.

Step 2: Expedite item development to populate the item bank using an Automatic Item Generation approach. To adequately support the implementation of CAT, one would need to have a reasonably large bank of items. Depending on the required measurement precision and content coverage, a typical CAT item bank would require 300 to 500 items. In standard practices, items are usually written by subject matter experts who use their knowledge and experience to construct items. This current approach of item writing to develop the entire item bank required by CAT administration is both time consuming and costly.

In order to expedite the item development process, an Automatic Item Generation (AIG) approach can be adopted. Instead of writing one single item at a time, item writers develop item templates (see Figure 1 below) that can be used to clone numerous similar items. For example, a Multiple Choice Question type of item template would contain the following components:

(i) **Template Descriptor**. To provide a description of the LO;
(ii) **Stem**. To parametrise the item using parameters;
(iii) **Constraints**. To apply the constraints to the parameters defined in (ii) so as to control the item demand appropriately;
(iv) **Options**. To provide definitions of the answer key and distractors based on the parameters defined in (ii); and
(v) **Auxiliary information**. To indicate the graphics and labels needed.

Figure 1: Item Template



In the AIG approach, the subject matter experts use their knowledge and experience to decide the number of item templates needed to cover each LO in the item bank blueprint adequately. Depending on the complexity of individual LOs, some may require two or more item templates to represent them well. After developing the item templates, each of them will be used to clone a specified number of items to populate the item bank. The cloning of items can be programmed using statistical software such as SAS, which randomly assigns suitable numbers (subject to the constraints defined by the subject matter experts) to the parameters in the item templates.

Step 3: Structure and construct the Topical Continuum by establishing and arranging the LO mastery thresholds in ascending order of demand across grade levels. After generating the items from each template, one can proceed to calibrate these items based on a sample of students with a good spread of ability for the Fractions topic. This can be achieved by randomly selecting students from various grade levels so that the sample includes students with different levels of mastery in Fractions. Once the items are calibrated, the mastery threshold of each LO template can be determined as follows:

(i) Plot the Test Characteristic Curve of each item template testlet that consists of all the items generated by that template with the vertical axis representing the score obtained and the horizontal axis representing the Fraction Continuum logit scale that has been established during the calibration of the item bank.

(ii) Determine the ability location along the Fraction Continuum scale that corresponds to a typical student correctly responding to 75% of items generated by the template. For example, in Figure 2, suppose a total of four items has been generated from an LO item template, then the mastery threshold for this item template will correspond to the ability location of –0.048 on the logit scale, which is the ability of a typical student attaining a score of 3 out of 4 items.

Figure 2: Establishment of Mastery Threshold for each Item Template

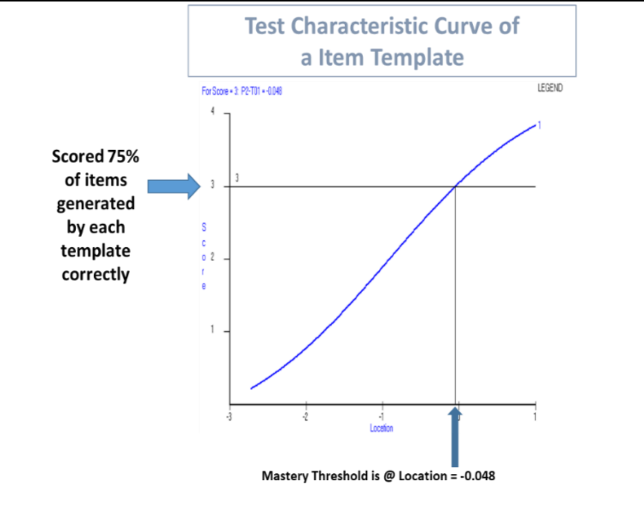
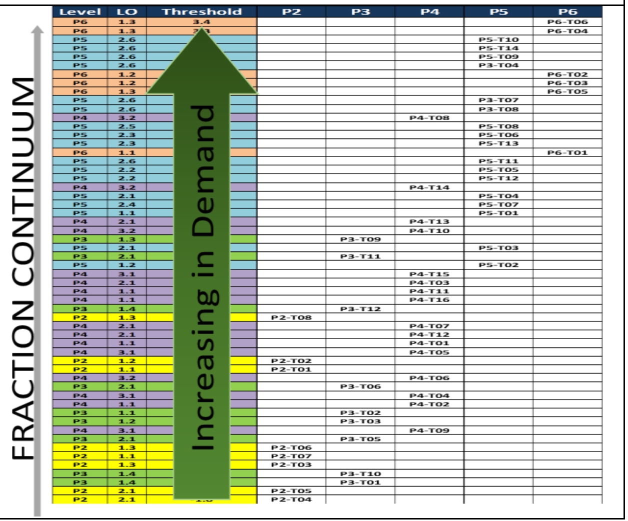


Figure 3: Fraction Continuum across Grade Levels

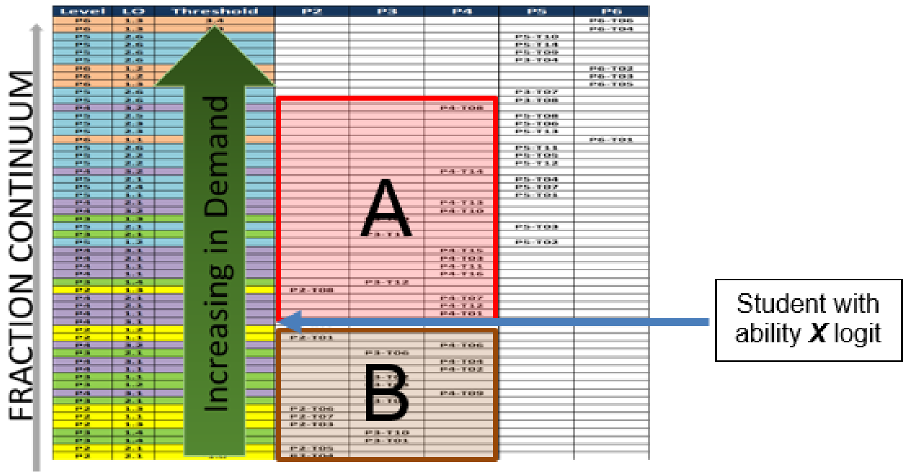


(iii) Repeat the process in (i) and (ii) for all the item templates to obtain the mastery thresholds for all templates.

(iv) Arrange the mastery thresholds of all the item templates in ascending order of demand across the grade levels. This forms the Fraction Continuum as shown in Figure 3.

Step 4: Integrate the Computerised Adaptive Testing engine to efficiently and accurately determine whether a student could meet each LO mastery threshold along the Continuum. With the establishment of the Fraction Continuum, the mastery thresholds of all the item templates are ordered along a continuous logit scale. If the ability of a student is estimated based on the Fractions item bank, then the student's mastery of each item template can be determined by checking against the established Fraction Continuum scale. As shown in Figure 4, a Grade 4 student with an estimated ability ***X*** logit, say, along the Fraction Continuum scale would be more likely to manage the item templates located in Region B and less likely to manage those in Region A, which are deemed to be more demanding for him or her.

Figure 4: Fraction Continuum across Grade Levels (Student Example)

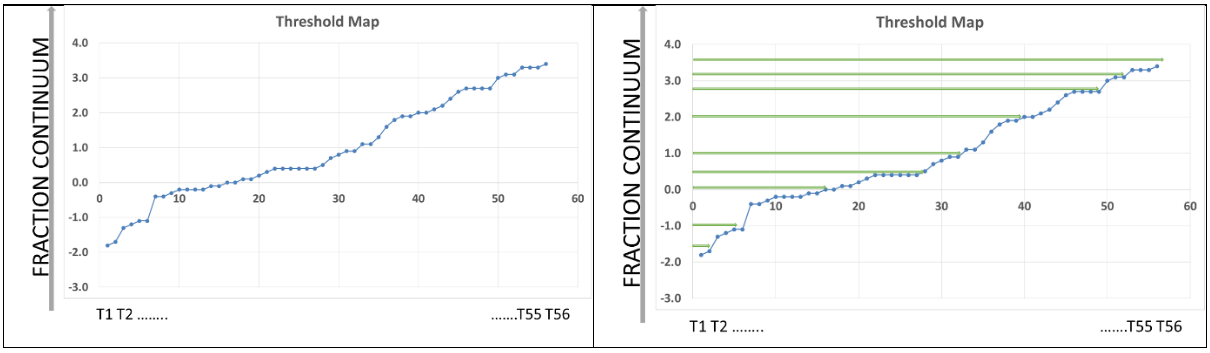


In order to estimate the student's ability accurately and efficiently, the CAT strategy can be deployed to tailor the computerised test to the student's performance during the test administration. In adaptive testing, the test will be customised to select items that contain high information about each student. This will in turn increase the measurement precision to determine his or her ability along the Fraction Continuum, even with the use of a reasonably short test. It is important to note that a student currently studying at Grade 5 may only have a Fractions mastery level at Grade 3. One key advantage of using CAT is that every student's mastery level in Fractions can be identified within a single test administration.

Step 5: Generate students' performance by profiles at the individual, class and school levels. To make the tool more useful for AfL purposes, the next step is to provide teachers with a useful interpretation of their students' performance on the adaptive test. Looking at the Fraction Continuum in Figure 3, one can observe that certain clusters of skills are of similar manageability compared to others. This indicates that statistical methods can be used to separate these clusters so that meaningful qualitative descriptions of students' mastery can be elicited from each cluster.

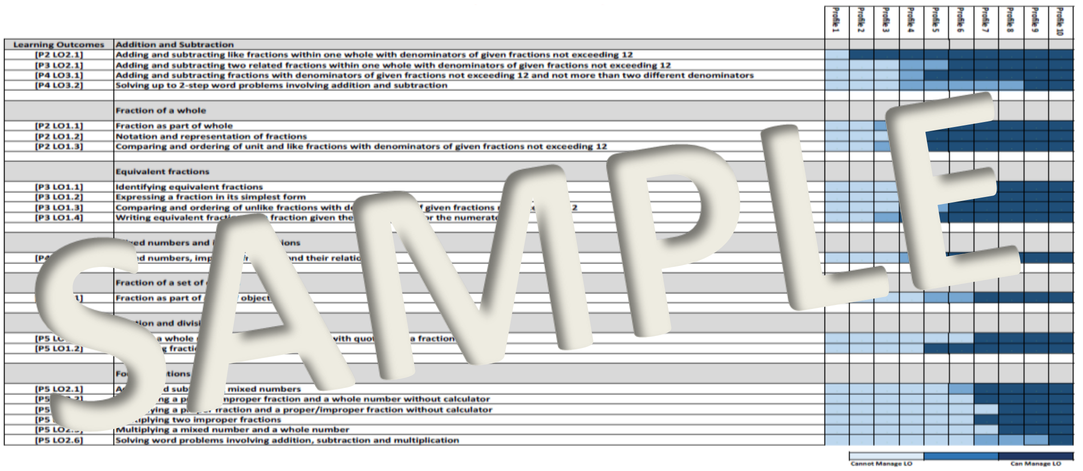
Figure 5 provides a plot of the mastery thresholds of item templates in ascending order of their demand. Observe that there exist naturally occurring breaks between clusters of item template mastery thresholds. These breaks provide a natural demand separation between adjacent clusters. By further ensuring that these adjacent clusters have at least a medium effect size difference, the manifestation of students' performance on the tool across grade levels could be distinctly and meaningfully profiled based on these naturally occurring clusters of item templates. These student profiles can also be aggregated at class and school levels to provide useful overviews for teachers and school leaders to monitor students' learning progress over time.

Figure 5: Threshold Map of the Fraction Continuum



Step 6: Notify and provide teachers with a report describing the LOs that each student can and cannot manage to inform customised interventions for individual students or groups of students. From Step 5, it was found that the manifestation of students' performance on the Fractions tool across grade levels can be distinctly and meaningfully reported as 10 distinct profiles. Based on the number of item templates mastered by students for each LO, their mastery level of each LO is then classified as "Cannot Manage LO", "Can Partially Manage LO" or "Can Manage LO" (expressed in three different shades of colour). The 10 distinct profiles can be distinguished using the degree of manageability of all the LOs in Fractions as shown in the Profile Descriptors in Figure 6.

Figure 6: Profile Descriptors of Computerised Fraction Test



During the implementation, a student will first take the Fractions CAT to ascertain his or her ability along the Fraction Continuum scale. After which, he or she will be assigned to one of the 10 profiles based on his or her ability in Fractions. With the profile descriptors provided, teachers will be able to identify and zoom into the LOs that the student could or could not manage. Finally, teachers can then customise interventions to bridge the learning gaps, targeting at individual students or groups of students.

DISCUSSION ON THE USE OF REPORTS

The approach outlined in the previous section illustrates the process by which CAT can be transformed to serve as a profiling test to inform teachers of the levels at which their students are currently performing and where there may be learning gaps. At the end of each test session, a set of reports is generated for each school, and comprises reports at the school-level, class-level and individual-level. Alongside these reports, an accompanying document (known as Profile Descriptors mentioned in Step 6 of the earlier section) containing detailed descriptors for each profile is also provided as a reference to facilitate the interpretation of the reports. In this section, we will discuss how these reports can be used to support AfL the classroom.

Figure 7a shows a sample school-level report, which gives an overview of the student distribution in a particular school across the different profiles. In the example, the majority of the students in XYZ Primary School fall into Profiles 6 and 7. This report is useful for school personnel such as the Head of Department (HOD), as it could help him or her identify areas of need for the cohort as a whole. In this case, after referring to the profile descriptors, the HOD may decide on further actions such as organising a pedagogical workshop on multiplication and division of fractions, or coordinating the development of additional resources on word problems.

Figure 7b shows a sample class-level report. This overview is again useful for the HOD, but the report is also broken down by class and given to the individual teachers. From here, it is clear that every class has different needs and the teacher is then encouraged to zoom in to the individual level to see how he or she can customise instruction based on the current mastery level of the students. As an example, the teacher of Class C may want to focus on types of problems that are more complex (such as word problems) as the students are mostly in Profiles 7 and 8. The teacher of Class H, on the other hand, may wish to start by spending time to revise and check the students' understanding of the fundamental concepts. Due to the test's adaptive nature, it is likely that some students in Class H did not manage to see or attempt the higher level problems during the test.

Figure 7a: Sample School-Level Report

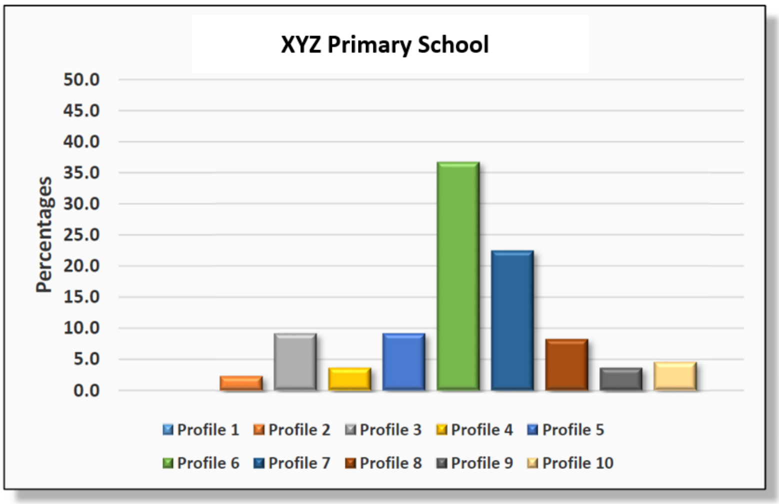


Figure 7b: Sample Class-Level Report

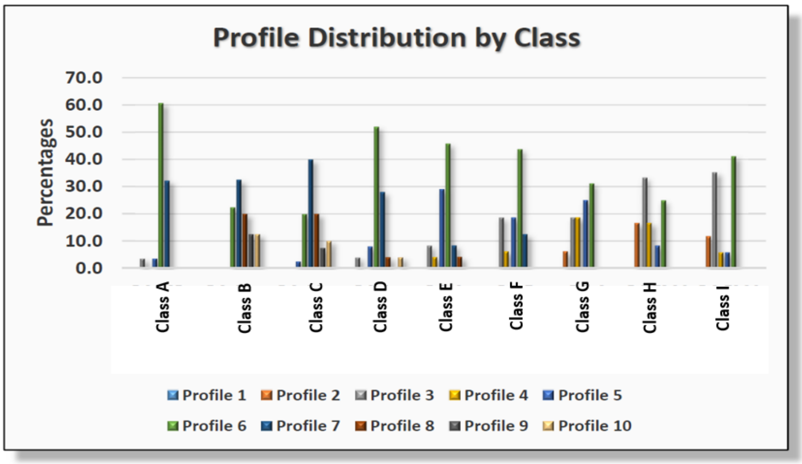


Figure 7c shows a sample individual-level report. This report is still organised at the class-level to give teachers an overview of the class performance, as well as to make it easier for teachers to refer to the individual LOs. In the given example, the teacher can see that most of the students are in Profiles 6 and 7. He or she can be confident that most of the students have mastered the fundamental LOs covered in Primary 3 and Primary 4, but many may find problems under "Fractions and Division" and "Four Operations" less familiar. A whole-class strategy could be to review multiplication and division of fractions, and use a variety of word problems to help students improve on their knowledge application skills. The teacher should also note that Student 8 seems to require extra support in this topic, and hence may consider further remediation actions. Customised actions can also be taken for Student 16 and Student 23, say, to help them close their gaps in the LOs that have not been fully mastered.

Figure 7c: Sample Individual-Level Report

Overall, these reports can help teachers to gain greater insight into their students' strengths and weaknesses in Fractions. The D.E.S.I.G.N. approach draws a link between a student’s ability estimate from CAT and the specific LOs through the development of a Topical Continuum. In the following section, we present some of the feedback that were obtained from primary school teachers in Singapore.

IMPLEMENTATION AND FEEDBACK

In early 2018, over 2000 Grade 6 students from 10 primary schools in Singapore sat the Fractions CAT. The various reports were generated and disseminated to the school teachers within two weeks of the respective test dates. Between the months of March and May, a questionnaire was administered to all teachers, and a series of Focused Group Discussions (FGDs) were conducted in schools to gather teachers’ feedback on the use of the reports.

Teachers fed back through the questionnaire that the reports are easy to interpret, and can provide them with the necessary information to tailor the pace and content of instruction to individual students’ needs. In the FGDs, teachers expressed their appreciation for the level of detail in the reports and generally found them to be informative and useful. Teachers from several schools shared that the reports have helped them identify students for remediation and to target specific LOs during lessons to help students close the gaps. Based on the reports, groups of teachers have also gotten together to discuss pedagogical strategies to teach particular LOs. In one school, the reports were incorporated into a Fractions workshop that was conducted for all teachers. Suggestions to improve the usefulness of the reports were also gathered from the teachers, both in terms of aesthetics and the level of detail in the LOs. These would be taken into consideration when reviewing the test and planning for future launches.

CONCLUDING REMARKS

Through the D.E.S.I.G.N. approach, we have developed an AfL tool which combines computerised adaptive testing with our Topical Continuum to help teachers to elicit evidence on their students’ mastery of a particular topic in Mathematics. Teachers are able to easily interpret the given data and take action when needed. The effectiveness of these intervention measures could be further explored, perhaps through the conduct of a post-test (as one school has done). The invaluable feedback gathered from schools will allow us to review and improve on the current reports to make them more useful for teachers. For instance, the more general LOs in the Mathematics Syllabus could be broken down further in the Topical Continuum so that even more details can be provided to teachers on what their students can or cannot do.

Looking to the future, in a self-directed learning environment, students will be able to take tests anytime and anywhere to ascertain their mastery of learning. A CAT such as the one described in this paper can be expanded to cover multiple topics, and reports can be generated automatically upon completion of the test, showing the student areas that he or she has mastered, as well as specific areas for improvement. In this big data environment, there is also potential for systemic data to be collected, analysed, and used to support procedures such as standard setting and cohort monitoring over the years.

REFERENCES

ACT (2016). *Computer-Adaptive Version of ACT Test to be Administered in International Test Centers Starting in Fall 2017.* Retrieved from https://www.act.org/content/act/en/newsroom/computer-adaptive-version-of-act-test-to-be-administered-in-inte.html

Anderson, C., & Palm, T. (2017). *Characteristics of improved formative assessment practice.* Education Inquiry, 8(2), 104–122.

Assessment Reform Group (2002). *Assessment for Learning: 10 Principles. Research-based principles to guide classroom practice.* Retrieved from https://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf

Bennett, R. E. (2011). *Formative assessment: a critical review.* Assessment in Education: Principles, Policy & Practice, 18(1), 5–25.

Black, P., & William, D. (1998a). *Assessment and Classroom Learning.* Assessment in Education: Principles, Policy & Practice, 5(1), 7–74.

Black, P., & William, D. (1998b). *Inside the Black Box: Raising Standards through Classroom Assessment.* The Phi Delta Kappan, 80(2), 139–148.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). *Working Inside the Black Box: Assessment for Learning in the Classroom.* The Phi Delta Kappan, 86(1), 8–21.

Lee, C., Oh, P. S., Ang, A., & Lee, G. (2014). *Holistic Assessment Implementation in Singapore Primary Schools – Part I: Using Assessment to Support the Learning and Development of Students.* Paper presented at the 40th International Association for Education Assessment Conference, Singapore.

Ministry of Education (2009). *Report of the Primary Education Review and Implementation Committee.* Singapore: Ministry of Education.

Ministry of Education (2012). *Mathematics Syllabus – Primary One to Primary 6 (Implementation starting with 2013 Primary One Cohort).* Retrieved from https://www.moe.gov.sg/docs/default-source/document/education/syllabuses/sciences/files/mathematics_syllabus_primary_1_to_6.pdf

Nortvedt, G. A., Santos, L., & Pinto, J. (2016). *Assessment for learning in Norway and Portugal: the case of primary school mathematics teaching.* Assessment in Education: Principles, Policy & Practice, 23(3), 377–395.

Rezaie, M., & Golshan M. (2015) *Computer Adaptive Test (CAT): Advantages and Limitations.* International Journal of Educational Investigations, 2(5), 128–137.

Seo, D. G., & Choi J. (2018) *Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination.* Journal of Educational Evaluation for Health Professions, 15(14).

Tan, K. H. K. (2017). *Asking questions of (what) assessment (should do) for learning: the case of bite-sized assessment for learning in Singapore.* Educational Research for Policy and Practice, 16(2), 189–202.

Weiss, D. J., & Vale C. D. (1987). *Adaptive Testing.* Applied Psychology: An International Review, 36(3/4), 249–262.

William, D., & Black, P. (1996). *Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment?* British Educational Research Journal, 22(5), 537–548.

ACKNOWLEDGEMENTS

We would like to thank our Singapore Examinations and Assessment Board (SEAB) management team for their support in this study; and our team members Andy Luo Kangshun, Chen-Theng Geak Seng, Chew Kang Wei, Ho Sok Wai Noelle, Jean Phua Yin Chiun, Keith John Lenden-Hitchcock, Lin Yuxin Celeste, Ong Ai Lingg, and Teo Khee Shoon for their contributions