CSAP/PREC

COLLEGE STUDENT ACHIEVEMENT PROJECT

Assessment Development Project:

Final Report

For the

Ontario Ministry of Education

and the

Ontario Ministry of Training, Colleges and Universities

Graham Orpwood and Emily Brown

Preface

The CSAP Assessment Development Project was a team effort and this report reflects the same approach. Chapters 1 and 4 were authored by Graham Orpwood, chapters 2 and 3 by Emily Brown, chapter 5 by Charles Anifowose, 6 by Melanie Christian and Charles Anifowose, and chapter 7 by Ruth Childs and Gulam Khan. The report as a whole was edited by Graham Orpwood and Emily Brown.

Nevertheless, it is our hope that the report as a whole gives the reader a coherent account of the development of the CSAP assessment. The Colleges Mathematics Assessment Program, as it has now become is now ready for implementation. The college administrations are continuing to work together to find an appropriate administrative and business model to enable the use of the assessment both by the colleges themselves and by students and the wider educational community.

Graham Orpwood

Emily Brown

Table of Contents

Chapter 1: Origins of the ADP

The Assessment Development Project (ADP) of the College Student Achievement Project (CSAP) had its origins in 2011, the final year of the College Mathematics Project (CMP). The research conducted by the CMP in that year had included a topical analysis of the foundational, preparatory or remedial mathematics courses that an increasing number of incoming college students were taking in their first semester at college. Either as a result of post-admissions testing or on the basis of their own choice, one in every six incoming students in Fall 2010 were either required or counselled to take a remedial or foundational rather than a diploma-level mathematics course. (Since then, that proportion has risen to over 25%.[1])

This growth in enrolment in foundational mathematics courses led the CMP team to conduct a systematic analysis of the topics being taught in such courses and this analysis forms a major part of the final CMP report.[2] This research showed, somewhat to our surprise, that many of the topics being taught in these courses corresponded to topics found explicitly in the mathematics curriculum from Grades 6, 7 and 8 – fractions, ratios, percentages, and basic algebra – rather than the, later, secondary school mathematics courses on which most college admission requirements are based).

In the deliberative forum that followed this research, the distinction between "mathematics" as a discipline and "numeracy" as a basic requirement for effective participation in work, in society and in further education emerged and became one of the major themes of the CMP final report that year. Recommendations for a multi-faceted provincial "Numeracy Strategy" were outlined with roles for schools, colleges, government and other parts of society. Among the recommendations directed to the college sector were the following two:

- To develop a common numeracy assessment tool to be used by all colleges as part of their admission and placement process for all incoming college students.
- To develop a system-wide college numeracy course for students whose scores on the numeracy assessment show that they need such a developmental course.

Subsequently, in supporting the establishment of the College Student Achievement Project (CSAP) (the successor project to the CMP) in 2013, the Ministry of Training, Colleges and Universities (MTCU) tasked the CSAP team with the implementation of these two recommendations, subject to a feasibility study.

ADP Feasibility Study

This feasibility study was carried out to explore two questions with respect to the development of both a common assessment tool and a set of common learning outcomes and common curriculum for foundational college mathematics: the technical feasibility – can it be done and, if so, how? – and the political feasibility – do the colleges agree that it should be done?

To answer these questions, the CSAP team conducted preliminary research by email survey to identify colleges using some form of assessment and then teleconferences with those colleges on their assessment experiences. The team then prepared a discussion paper outlining the results of

[1] CSAP Project Team, *College Student Achievement Project: Final Report 2014* (in publication).
[2] Graham Orpwood et al. *College Mathematics Project 2011: Final Report* (Toronto: Seneca College: 2012), available on the CMP web site (http://collegemathproject.senecac.on.ca/)

the survey, a model for a proposed assessment, and a set of choices for developing common learning outcomes and curricula[3]. These were then discussed over the following three months both at each college and at the meetings of system-wide committees (including the Heads of Technology, Business and Interdisciplinary Studies, the Ontario College Mathematics Council and the Coordinating Committee of Vice-Presidents Academic) with a view to discovering a consensus on the political questions.

The Feasibility Study report[4] was published in August 2013 documenting the broad endorsement (23 of 24 colleges) of the proposed model for the development of a common assessment tool, equally strong endorsement for the development of common learning outcomes for pre-business mathematics, pre-technology mathematics, and business diploma mathematics (but not for technology diploma mathematics nor for the development of common mathematics curricula).

Following the feasibility study, two projects were initiated as part of the overall CSAP: the Assessment Development Project (ADP) of which this is the final report; and the Learning Outcomes Development Project (LODP), whose final report, has already been published[5]. This report now focuses exclusively on the ADP.

Findings of the Feasibility Study

The email survey and subsequent series of teleconferences undertaken as part of the feasibility study revealed that:

- 14 of the 24 colleges presently used some form of mathematics assessment of post-admission students in all or some academic program areas;
- 5 of these used assessment tests that had been developed in-house;
- 4 used an assessment based on an assessment developed many years ago by members of the Ontario College Mathematics Association;
- 4 primarily used Accuplacer, a proprietary assessment from the United States;
- 1 used Compass, a proprietary tool from Australia.

There was found to be a great variety of uses for the assessment, including student profiling, course exemption, student advisement, and college statistics, among many others. The principle use was as an administrative tool for the college and in many cases, students remained uninformed of their results.

Colleges currently using an assessment were mostly open to developing a common, "made-in-Ontario" assessment and saw benefits of having a common tool in use across the college system. Concerns were expressed that colleges should retain the right to use the results of any assessment in the ways they individually saw fit. However many individuals felt that the time was right to explore a new approach to the assessment of incoming students.

---

[3] Graham Orpwood & Emily Brown. *Assessing Mathematics Skills for College: A Way Forward?* CSAP Discussion Paper. Toronto: Seneca College, 2013 (CSAP web site: http://csap.senecacollege.ca).

[4] Graham Orpwood & Emily Brown. *Developing Mathematics Assessment, Learning Outcomes and Curriculum for the Ontario College System: Final Report of a Feasibility Study*. Toronto: Seneca College, 2013 (CSAP web site: http://csap.senecacollege.ca).

[5] Patricia Byers, *Bridging the Mathematics Gap through Learning Outcomes - Final Report*. Toronto: Seneca College, 2014 (CSAP web site: http://csap.senecacollege.ca).

A brief telephone survey of leading colleges in other provinces of Canada was also undertaken. This revealed that many colleges were in a state of transition with regard to assessment and the general feeling was expressed that a common mathematics assessment tool would be beneficial nationally.

CSAP Assessment Principles

With this background, the CSAP team proposed a set of four principles to underpin the future development of a system-wide approach to assessment of incoming college students. These were strongly supported by the colleges and are therefore restated here as the basis for the program.

1. *The primary purpose of the assessment should be to support students' learning.*

Before embarking on a system-wide approach to assessing students' mathematics skills, there needs to be a consensus on its basic purpose. Both the educational research literature and Ontario assessment policy has evolved significantly in recent years and now clearly favours a student-centred system where what is known as "assessment for learning" takes precedence over "assessment for administrative purposes".

CSAP has adopted a concept of assessment for learning comprising five basic features:

- The active involvement of students in their own learning;
- The provision of effective feedback to students;
- Adjusting teaching to take into account the results of the assessment;
- The need for students to be able to assess themselves;
- Recognition of the profound influence that assessment has on the motivation and self-esteem of students, both of which are crucial influences on learning[6].

Placing a primary focus on students' learning does not diminish an assessment's direct value to colleges. A good mathematics skills assessment can assist colleges in placing students into appropriate courses in their first semester, it can enable institutions to use the results as baseline data for program or course evaluation, and it can enable the early identification of students likely to require additional support in college. It can also provide the basis for facilitating a constructive conversation with the elementary and secondary school community about colleges' expectations of students in relation to mathematics knowledge and skills.

2. *The focus of the assessment should be on high levels of competence on basic numeracy skills.*

The CSAP system-wide mathematics skills assessment is not equivalent to an evaluation of students' grade 12 achievement. The tests in current use are already assessments of more basic mathematics skills and, as has been noted earlier, the CMP research showed that the preparatory mathematics courses offered at most colleges are focused more on basic numeracy skills than on topics taught in grades 11 and 12.

While the mathematical knowledge and skills assessed by the CSAP assessment are of a basic nature, the level of competency required of students entering college programs is high. A 50% pass mark in calculations involving percentages or fractions, for example, is insufficient for students entering a

[6] Adapted from: Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.

college program that prepares them for employment in technology, business or health care. Rather, students in such programs should be able to demonstrate a high degree of competence in basic mathematics and, where necessary, *without* the use of technological aids such as calculators.

3. *The assessment should be developed and operated according to high standards of quality.*

Over the past several years, standards for the development and use of assessment tests have been developed both in Canada and internationally.[7] These standards should guide the development and use of a college mathematics assessment. Selected aspects of these are as follows:

- Assessment must be seen not as a (test) document alone, but as a complete process that includes the conduct of the assessment and the uses made of the results.
- Assessments should be aligned with expectations for student learning and instructional opportunities and provide for feedback to students as an essential design element.
- Assessments should respect students' cultural and linguistic diversity and be free of bias.

CSAP has adhered to these quality guidelines throughout the development of the college mathematics assessment.

4. *The assessment should be as cost-effective as possible*

While the assessment system needs to be of high quality, its costs – whether to the students, the colleges, or the system as a whole – need to be taken into account and the assessment system needs to be as efficient as possible. Developing the system involves much more than the development of the tests themselves and as the project proceeds, many practical factors need to be taken into account. Now that the assessment development is complete, the college system is seeking to establish a system based on the same principle for their ongoing use and further development.

The CSAP Assessment Model

Based on these four principles, the CSAP developed a model to guide the development of the assessment. This model, which was first outlined in the discussion paper and refined in the feasibility study final report, has evolved further as the ADP has progressed with consultations among stakeholders.

Its central features are as follows.

[7] For example: *Principles for fair student assessment practices for education in Canada* (1993) (Edmonton: Joint Advisory Committee, Faculty of Education, University of Alberta); Joint Committee on Standards for Educational Evaluation (JCSEE). *Classroom assessment standards* (5th draft) (2013). In publication.

Two types of test

The college mathematics assessment actually comprises two distinct types of test: a basic numeracy test (yielding a single numeracy score); and a full diagnostic test (yielding a numeracy score and sub-scores on each of nine mathematics topics). All tests, along with reporting and remedial systems are available in both English and French.

Multiple modes of use

The assessment system as developed has multiple modes of use. The first two of these (originally called formal and informal) were described in the feasibility study report while the third and fourth have emerged during the development project itself.

1. The first and original mode is under the supervision of a college (or an academic unit within a college) and used to assess post-admission students. At the discretion of a college, all or some students entering business or technology programs could be required to take the assessment, with reports being provided both to the students and to the college.
2. The second mode of use, also envisioned in the feasibility study, is by secondary school students or adults who are considering applying to a college program and who wish to determine if they have the mathematics skills expected of incoming college students and to undertake remedial work if necessary. In this mode, reporting would only be to the students themselves.
3. As the ADP developed, a third mode of use emerged as college faculty and secondary school teachers asked if they could use the tool as a prior learning assessment for students new to their particular mathematics course. In this mode, reports would go both to the students and to the faculty member or teacher, who could use the test to identify students with serious weaknesses and in need of special support and also to modify their own teaching to take into account topics which significant numbers of students found challenging.
4. Finally, as a variety of other institutions became aware of the college mathematics assessment, inquiries about its use by a university business school, two university faculties of education and other groups began to appear.

Computer-based delivery from a central platform

In order to provide multiple consistent and reliable assessments, the test system is computer-based. Each individual test is uniquely constructed from a pool of items whose psychometric characteristics have been established. In addition, each test module is linked to a remedial self-instructional unit so that students, whose score on any given topic is lower than desired, can immediately seek support. In order to provide students with universal access to the assessment system, both the assessment and remedial systems are mounted on a web-based technology platform accessible from any internet-connected device. The details of the technology platform are discussed further in chapter 5.

The ADP Work Plan

Work on the ADP was authorised by the MTCU in August 2013 and over the following 18 months, the following tasks and sub-tasks were undertaken. Each of these is discussed in more detail in the following chapters of this report.

Task A: Curriculum and Assessment Framework (see chapter 2)

- Analysis of college course outlines
- Development of draft framework
- Validation of framework with colleges and schools

Task B: Assessment Items (see chapter 3)

- Collection and classification of items from existing sources
- Development of new items
- Pre-testing and analysis of items
- Development of French language items

Task C: Test design (see chapter 4)

- Consultation over test design
- Development of test design

Task D: Technology Platform (see chapter 5)

- Consultation with MTCU, CCVPA and (now) OCAS: Ongoing
- Support Development of RFP: In 2014
- Ongoing collaboration between CSAP, OCAS, and AMS provider: in 2014

Task E: Remedial Modules (see chapter 6)

- Collection of instructional materials
- Development/refinement of remedial modules
- Development of French language remedial modules

Task F: Field Trials (see chapter 7)

- College field trials
- Secondary school field trials
- Analysis of field trial data
- System modifications

Task E: Governance and Management Protocols (work in progress)

- Selection of lead institution
- Drafting of implementation plans
- Approval processes

ADP Development Partners

In completing the development of this assessment project, the members of the CSAP team would like to acknowledge the major roles played by many partners in this project. The following were of particular importance.

- The colleges and, in particular, the mathematics faculty who contributed course outlines and assessment items to enable to project to begin and to all who facilitated the field trialling of the assessments by over 4000 students in 16 colleges during Fall 2014.
- Secondary schools throughout Ontario, 112 of which participated in either the pre-testing of items in the spring of 2014 or the field testing of the final tests in the fall of 2014 by over 6000 students. Their feedback was of enormous value in refining both the items and the overall test system.
- Vretta Inc. which constructed the technology platform for the assessment and provided draft remedial modules to enable this project to incorporate this feature.
- Dr Ruth Childs and Mr Gulam Khan of OISE/UT whose analyses of the item pre-test and the field trial of the assessment enabled the ADP to meet its quality standards.
- The ADP/LODP Advisory Committee, with members of the college and secondary school community with expertise in mathematics and assessment for their advice and support.
- The CSAP steering committee for its ongoing support and advice.

Chapter 2: Assessment Framework

The assessment framework is the foundation of the development and classification of test items, the basis for the test and reporting design, and the key to linking the remedial instructional modules to the assessment. It sets out the project's selection of content topics and of the skills students are expected to display in relation to those topics. It was therefore the first and central element in designing the CSAP assessment.

As resources for constructing the CSAP Assessment Framework, we employed the following:

- The topical analysis of foundational mathematics courses conducted by the College Mathematics Project (CMP) team in 2011 and reported in its final report[8];

Numeracy topics in foundational mathematics courses: Pre-business

College	Order of Operations	Fractions	Decimals	Percentages	Ratio & Proportion	Algebra	Exponents
A	Y	Y			Y	Y	Y
B	Y	Y	Y	Y		Y	
C	Y	Y	Y	Y	Y	Y	
D	Y	Y	Y	Y	Y	Y	Y
E	Y	Y	Y	Y	Y	Y	Y
F	Y	Y	Y	Y	Y	Y	
G	Y	Y	Y	Y	Y		
H	Y	Y	Y	Y	Y	Y	Y
I	Y	Y	Y	Y	Y	Y	Y
J	Y	Y	Y				Y
K	Y	Y	Y	Y	Y		
Total	11	11	10	9	9	8	6
Percent	100%	100%	90.9%	81.8%	81.8%	72.7%	54.5%

Numeracy topics in foundational mathematics courses: Pre-technology

College	Order of Operations	Fractions	Decimals	Percentages	Ratio & Proportion	Algebra	Exponents
A	Y	Y			Y	Y	Y
B	Y	Y	Y		Y	Y	Y
C	Y	Y	Y		Y	Y	
D	Y	Y	Y	Y	Y	Y	Y
E	Y	Y	Y	Y	Y	Y	Y
F	Y	Y			Y	Y	Y
G	Y	Y	Y	Y	Y	Y	
H	Y	Y	Y		Y	Y	
I	Y	Y	Y	Y		Y	Y
J	Y	Y	Y	Y	Y	Y	Y
K	Y	Y	Y	Y	Y	Y	Y
L	Y					Y	Y
M	Y	Y	Y	Y		Y	Y
N	Y	Y				Y	Y
O	Y	Y	Y	Y	Y	Y	Y
P	Y	Y	Y	Y	Y	Y	Y
Q	Y	Y	Y	Y	Y	Y	Y
R	Y				Y	Y	Y
Total	18	16	14	10	14	18	15
Percent	100%	88.9%	77.8%	71.4%	77.8%	100%	83.3%

- Course outlines for remedial and foundational mathematics courses submitted by colleges as part of the ADP and LODP;

[8] Graham Orpwood et al. *College Mathematics Project 2011: Final Report* (Toronto: Seneca College, 2012).

- Textbooks in common use for foundational mathematics courses in colleges;
- The monograph "Curriculum Frameworks for Mathematics and Science" developed for the Third International Mathematics and Science Study (TIMSS).

From the first three of these, we developed a draft list of topics that represented topics common to pre-business and pre-technology mathematics courses. It was recognised from the start that, in the time and with the resources we had available, we could not create an assessment that would cover *all* the topics regarded as important for *all* college program areas. Rather by focusing on those topics most frequently encountered in foundational mathematics courses, it was hoped that the CSAP assessment would be of value to the greatest number of colleges and college faculty. With the knowledge that the assessment would be modular in nature, it should be noted that the framework could be expanded or adapted at any time in the future to suit the needs of a particular college or program to include modules which focused specifically on concepts for business, technology or programs not included in the CSAP ADP work at this time, such as health, and the skilled trades, including topics such as geometry and trigonometry.

This list of nine mathematics topics and sub-topics forms the content dimension of the assessment framework.

1. Whole Numbers
2. Arithmetic
3. Integers
4. Decimals
5. Fractions
6. Ratio and Proportion
7. Percents
8. Basic Algebra
9. Measurement

Mathematics achievement, however, is more than knowing basic mathematical topics and performing simple operations. Mathematics courses include skills above and beyond such basic goals: application of mathematics concepts, problem solving, analysis and more complex reasoning, and so on. And a rich assessment requires that this dimension of mathematics learning be represented in the assessment framework also. We therefore used the concept of "performance expectations" as described by Robitaille and colleagues for the Third International Mathematics and Science Study (TIMSS)[9] as the means for identifying this second dimension of the assessment framework.

Three categories of performance expectation were identified to form the second dimension of the CSAP assessment framework:

A. **Knowing** (including recalling, recognizing, computing, retrieving, measuring, classifying, ordering)

---

[9] David Robitaille (ed.). *Curriculum Frameworks for Mathematics and Science* (TIMSS Monograph No. 1). Vancouver: Pacific Educational Press, 1993.

B. **Applying** (including selecting, representing, modelling, implementing, routine problem solving)
C. **Reasoning** (including analyzing, generalizing, specializing, integrating, synthesizing, justifying, non-routine problem solving)

When the content dimension and the performance dimension are put together they correspond to the learning outcomes and objectives for which the assessment is designed to be a measure and the remedial modules support student learning. This two-dimensional matrix representing the assessment curriculum framework is shown below.

	CSAP Assessment Framework	***Performance Expectations***		
		A **Knowing**	**B** **Applying**	**C** **Reasoning**
	Content Topics and Sub-topics	Recall, Recognize, Compute, Retrieve, Measure, Classify, Order	Select, Represent, Model, Implement, Solve routine problems	Analyze, Generalize Specialize, Integrate, Synthesize, Justify, Solve non-routine problems
1	**Whole Numbers**			
1.1	Place value, reading and writing numbers			
1.2	Equality/inequality			
1.3	Rounding			
1.4	Absolute value			
2	**Arithmetic**			
2.1	Addition			
2.2	Subtraction			
2.3	Multiplication			
2.4	Division			
2.5	Exponential notation			
2.6	Order of operations			
2.7	Prime numbers			
2.8	Factors and multiples			
2.9	Scientific notation			
3	**Integers**			
3.1	Adding and subtracting negatives			
3.2	Multiplying and dividing negatives			
3.3	Exponents with negatives			
4	**Decimals**			
4.1	Place value, reading and writing numbers			
4.2	Arithmetic operations with decimals			
4.3	Rounding			
5	**Fractions (rational numbers)**			
5.1	Types of fractions, equivalent fractions, conversion			

5.2	Addition and subtraction			
5.3	Multiplication			
5.4	Division			
5.5	Order of operations			
6	**Ratio and proportion**			
6.1	Creating ratios			
6.2	Ratios in simplest form			
6.3	Calculations involving ratios			
6.4	Rates			
6.5	Proportions			
7	**Percents**			
7.1	Converting between percent, fraction and decimal			
7.2	Calculate amount given base and percent			
7.3	Calculate percent given amount and base			
7.4	Calculate base given percent and amount			
7.5	Calculate percent change			
8	**Basic Algebra**			
8.1	Variable expressions			
8.2	Monomial operations			
8.3	Binomial operations			
8.4	Polynomial operations			
8.5	Factoring			
8.6	Solving equations			
9	**Measurement & Unit Conversion**			
9.1	Mass and length			
9.2	Metric-imperial conversions			
9.3	Area and volume (capacity)			

The CSAP assessment framework in its two-dimensional form, has provided us with a coding system which was used to clearly identify each assessment item by the topic and sub-topic it addressed and the performance expected of the student responding to it. For example, an item designed to assess a student's ability to apply their knowledge (Performance Expectation B) of fractions (content topic #5), using addition of fractions (sub-topic 1), would be tagged 5.1 B. This coding was critical both in evaluating the items (see chapter 3) and in designing the overall tests (see chapter 4). The matrix also provides a visual template for displaying and identifying the numbers of items related to each topic and performance expectation.

Following its development, the framework was validated by consultation with the CSAP ADP Advisory Committee and a variety of college organisations including Ontario College Mathematics Council (OCMC), Ontario College Mathematics Association (OCMA), Heads of Technology (HOT), Heads of Business (HOB) and Heads of Interdisciplinary Studies (HOIS). The Learning Outcomes Development Project team judged it to be broadly consistent with their framework, while pointing out that their outcomes were differentiated by program (pre-business and pre-technology) whereas the ADP framework was designed to reflect both.

Chapter 3: Assessment Items

Assessment Item Collection

As soon as the ADP was initiated, colleges were asked to contribute current course outlines and assessment tools for foundational (pre-business, and pre-technology) mathematics courses. As a first step, assessment items, which were currently in use by colleges and which fit the curriculum framework, were extracted and tagged for inclusion in the ADP assessment item pool. Next, items were sourced from the public domain in an effort to populate the cells in the assessment framework and were also adapted and tagged.

Each assessment item collected was given two tags which identified it with one or more content topics and sub-topics and with one level of performance expectation. This allowed for an ongoing count of assessment items in each cell of the assessment framework to ensure good coverage, and to identify topic and performance areas which needed more assessment items due to poor representation. It should be noted that most of the items collected to this point were in the domain of the 'knowing' (A) performance expectation, with some in the 'applying' (B) performance expectation domain, leaving substantial gaps in the curriculum framework, particularly for the 'reasoning' (C) performance expectation.

The assessment framework mapping was examined after all these assessment items were tagged and the cells were further populated with items written specifically to fill any gaps existing in assessment framework and ensure adequate coverage.

Item Types: Selected Response v Constructed Response

The majority of the assessment items both from colleges and from the public domain were of a selected response (multiple choice) type. This is not surprising since on a paper-and-pencil test such items can be scored easily by a computer. However selected response items also have serious disadvantages: firstly, such items do not directly assess a student's ability to conduct a mathematical operation or solve a problem; rather that ability is *inferred* from the student's ability to select the correct answer from a limited number (usually 4 or 5) of choices. Secondly, since random selection of answers has a 20% or 25% chance of being correct (depending on the number of choices), overall scores on tests made up of selected response items are inevitably inflated to some degree.

A constructed response item, by contrast, requires that a student actually performs the operation required by the question and enter it into a blank space on the screen. In a paper-and-pencil test, this means that papers must be scored manually, which is why such items are used less often in large-scale assessment. However in an on-line assessment, the computer can be programmed to accept as correct, a range of responses (such as 43, 43.0, 43.00) where the range of acceptability can be determined following an item's pre-testing. This conversion relied on the strength of the underlying assessment system to perform multiple layers with multiple accepted responses so as not to be biased towards marking items as incorrect for small discrepancies in expected input.
Thus, constructed response items are preferred where the correct answer can be entered fairly simply. Items where the correct answer was more complex and where a student's successful response might be inhibited by their ability to work comfortably with the technology interface (such

as in the case of an algebraic expression or an exponent), were usually left in selected response format.

Item Digitization

Items were written into the online platform using a template system for the text and notation, and used randomized algorithms to produce of a variety of numeric values. This provided the parameters which would determine which variants could be drawn into the assessment.

Variant Selection

A large number of the assessment items require at least some level of computation. As the assessment was designed, certain aspects needed to be closely attended to and parameters set for possible variants to reflect the following concerns.

1. Realistic quantities. Many of the items on the assessment were contextualized in real life situations and quantities that would be realistic within each of those settings needed to be determined.

2. Ease of computation. The assessment is designed to be completed without the use of a calculator. Many items have implicit calculations which must be done to arrive at the correct answer and without careful selection of variants, the item could put an overly large emphasis on computation, taking time and attention away from the concept that is being tested. Further, variants were selected such that students who have a high level of computational or mathematical understanding and flexibility would find the questions quite reasonable to solve without a calculator. Rather than be frustrated with the lack of a calculator, these students might see the solution of a question simplified by factoring, for example.

For the purpose of the field testing, where the psychometric properties of each assessment item were being determined, only one variant of each assessment item was used.

Adaptation for Francophone Students in Ontario

During the digitization process, each assessment item was assigned a unique identification number. These identification numbers were used to link English items with their adapted French counterparts. The English items were adapted to French by francophone mathematics teachers in the Ontario college system, ensuring that both the technical mathematics language as well as other expressions and forms of speech were recognizable to the unique francophone college student population in Ontario.

Calculators

After lengthy discussions with college mathematics faculty and college administrators, it was agreed that the tests should be designed to be completed *without the use of calculators.* College faculty have argued that, while calculators have many valuable benefits, incoming college students should be able to demonstrate their ability to undertake simple calculations without dependence on the use of electronic devices. However, not allowing students to use calculators during the field trials did raise concerns from teachers and students in the secondary school system, as well as in discussions with members of our advisory committee who represented the K-12 system, who noted that students writing province wide Education Quality and Accountability Office (EQAO) assessments are allowed to use calculators throughout starting in grade 6. While this decision is somewhat controversial it has been revisited several times during the development process and reinforced on each occasion.

Pre-Testing of Items

All items in the pool were then tested in secondary school classrooms during the spring of 2014. Teachers of grades 11 and 12 college level mathematics courses (MBF3C, MAP4C and MCT4C) were recruited and authorisations obtained from school board superintendents and school principals. Each student in the participating classes was provided with an online consent form, and in some cases paper informed consent forms as well, where required by the school board, enabling those who did not wish to participate, to withdraw before beginning the test. The first "tests" themselves were not comprehensive tests covering the overall framework but rather sets of items somewhat randomly assembled to enable all items to be pre-tested by a minimum of 50-100 students. One carefully selected anchor block of 8 items was included in all test forms to establish a baseline for comparison of student achievement regardless of test paper and thus to enable analysis of item difficulty and other psychometric properties of each item. In addition to responding to the questions, students were asked to provide information on their gender and mother tongue to enable analysis of the items from these demographic perspectives. They were also asked for an overall qualitative summary evaluation of the assessment and the items and teachers were asked to evaluate each item for appropriateness. Specifically, teachers were asked if the item was too easy, too difficult, assessed an untaught concept, was in an unfamiliar context, or contained unfamiliar vocabulary. Teachers could provide as many of these descriptors per item as they saw fit, or none at all. Each participating school was given a $500 donation following completion of their evaluations.

The student responses to these item pre-tests were collected by the Vretta technology platform (see chapter 5 for further details) and submitted to experts at OISE/UT for detailed psychometric analysis. Further details of their analytical techniques and final item analysis is provided in chapter 7 but the analysis of the pre-test data provided data for all items. This included p-values (measures of difficulty level from 0-1 where 0 represents nobody obtaining a correct response and 1 represents all students obtaining a correct response) and point-biserials (a measure of the correlation between correct responses on a given item with overall achievement on the test). Following this analysis, each item's data was examined carefully and items were selected based on p-values falling within a 0.1-0.9 range and point-biserials being positive and as high as possible. Items failing to satisfy these criteria were either deleted or reconstructed prior to their being tested again in a later field trial.

The qualitative data from students and teachers was also reviewed at this stage. Further feedback on the assessment items was also provided by the CSAP ADP advisory committee members, individuals from both the Ministry of Education, and the Education Quality and Accountability Office. Based on the vast amount of feedback obtained, the assessment items were sorted into three categories; those which would be kept in their current form, those which needed modification, either to mitigate any perceived sensitivities (gender, language, culture), or to remove any ambiguity which surfaced from examining groupings of incorrect responses, and lastly, those which would be discarded.

The assessment items which was retained, unchanged, and those items which were modified, formed the pool from which assessment items were pulled to create the beta tests which were used in the final round of field tests with students currently in first year programs in college, or currently enrolled in either university or college pathway mathematics courses.

A summary of the numbers of items available for inclusion into the field trial tests in relation to the assessment framework is shown below.

	Performance Objectives		
Knowledge Objectives	**A. Knowing**	**B. Applying**	**C. Reasoning**
1. Whole Numbers	14	9	7
2. Arithmetic	32	12	4
3. Integers	22	3	12
4. Decimals	6	9	4
5. Fractions	12	18	14
6. Ratio and proportion	13	8	14
7. Percents	4	12	24
8. Basic Algebra	6	8	10
9. Measurement	6	13	12

Chapter 4: Test Design

A valid and reliable test is more than the sum of its items. The design of the test and the method of constructing test forms was an important consideration for the ADP and we sought advice from our psychometric experts in undertaking it.

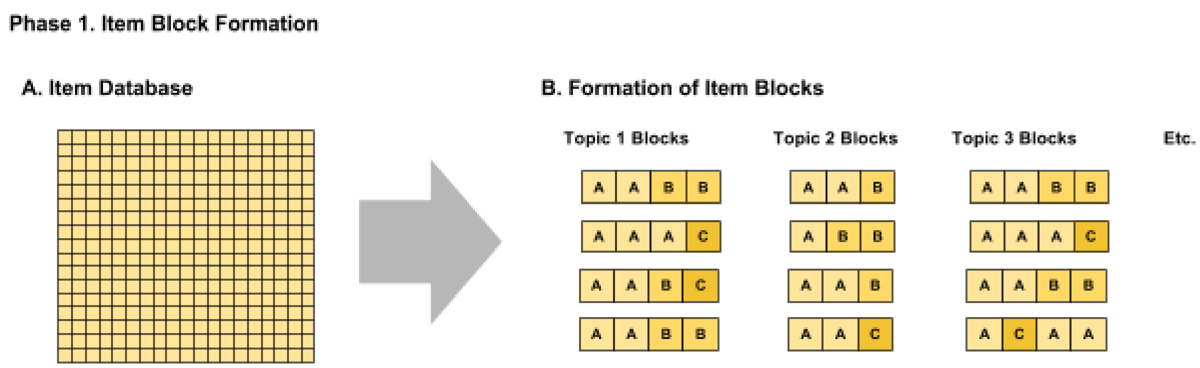
In creating the test design, the following criteria needed to be satisfied:

- Each test taker should receive a unique but equivalent test form
- Each test form should be equally representative of the assessment framework
- Each test form should be equally difficult
- Tests should contain sufficient items for reporting purposes but be short enough to complete in a reasonable length of time
- Diagnostic tests should contain enough items to enable reliable reporting of both overall scores and topic-based sub-scores

Accordingly the test forms are constructed using blocks of items in two phases as follows:

1. In phase 1, items for a given content topic are sorted into blocks comprising from 3 to 5 items, based on the following principles:
    a. Each block of items should have as close as possible the same average p-value (i.e. blocks are of approximately equal difficulty;
    b. The range of content tags (i.e. the range of sub-topics) of the items within each block should be as broad as possible;
    c. No more than half of the items in any block should have an A (knowing) performance tag.

This phase can be illustrated as follows, where items are shown as small squares with their performance tags (A, B or C) shown:



2. In phase 2, tests of two basic types – basic numeracy and full diagnostic – are made up as follows:
    a. A basic numeracy test form comprises *one* block of items, selected at random, from each set of topic blocks; this will contain 36 items (9 topic blocks x 4 items/block)

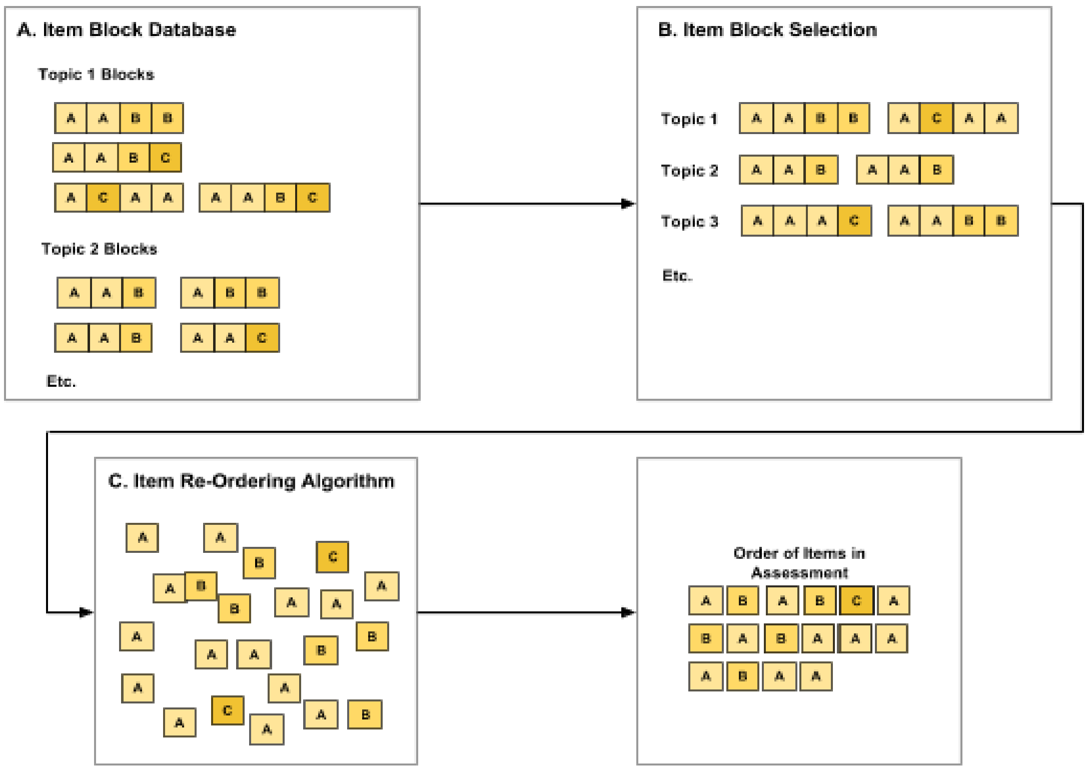
which can be completed in 45-60 minutes. It will yield a single "basic numeracy" score;

b. A full diagnostic test form comprises *two* blocks of items, selected at random, from each set of topic blocks; this will contain 72 items (2 x 9 topic blocks x 4 items/block) which can be completed in 90-120 minutes. It will yield a single "basic numeracy" score *plus* a sub-score for each of the 9 content topics;

c. Following the selection of the blocks for each test form, the sequence of items within that form is changed to a random sequence to avoid the effect of a topic where the test taker is weak having an impact on their overall achievement.

Phase 2 is illustrated below:

Phase 2. On-Demand Form Creation



Further information about the creation and on-line delivery of test forms to users is provided in chapter 5.

Chapter 5: Technology Platform

The platform on which the CSAP assessment was mounted for its development was created by Vretta Inc. and comprises a series of inter-related systems and databases. The following should be understood as a conceptual rather than a technical description of the platform elements. For example, what is described here as the "item database" is in reality made up of several actual databases.

- Item database
- User database
- Test management system
    - Test creation and delivery
    - Scoring and reporting
    - Data analysis

Item Database

This database contains records relating to every test item received or developed for the ADP. It can be thought of as comprising five sections, as follows:

- Development history
- Basic properties
- Pretest results
    - Psychometric analysis
    - Qualitative evaluation
- Block assignment for field trial
- Field trial results

User Database

This database contains records of all individual test-takers, along with their group affiliations where appropriate. Note that, for the purposes of the ADP, including both item pretesting and the full field trials, all test takers participated anonymously and so those fields relating to their identification were unused. In addition, the various "modes of use" (see chapter 1) of the assessment system imply correspondingly different administrative arrangements for individual test-takers and that adds a level of complexity to the way that "user" information is recorded. The following account is a somewhat simplified version of the actual database, therefore.

This database includes the following sections:

- User identification
    - Unique Identification Number (UIN) – generated by the system
    - *Name of test taker (not used in ADP trials)*
    - *School or college ID (not used in ADP trials)*
    - *Email (if not part of a group) (not used in ADP trials)*
    - Gender of test taker
    - Language of test (E or F)
    - Language of test taker (other than E/F)

- Group identification (where appropriate)
    - Name of institution
    - Type of institution (college/school)
    - School board affiliation (where appropriate)
    - Institutional language (E or F)
    - Group identification (e.g. MCT4C class, Faculty of Business, etc)
    - Group contact person
    - Email address
- Test results

Test Management System

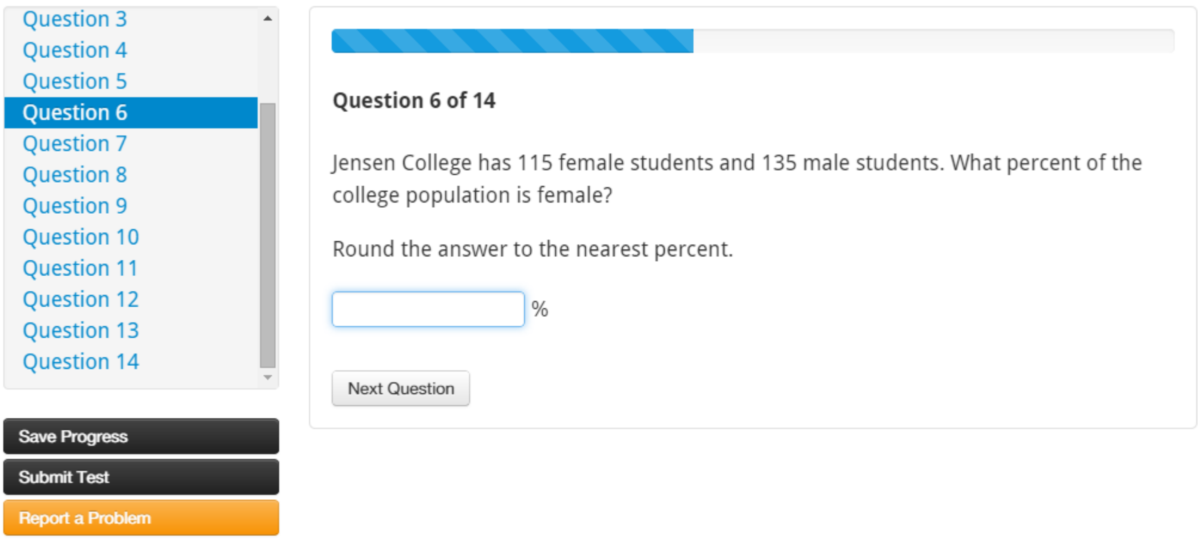
The central role of the technology platform was its capacity to create and deliver unique tests on demand, to receive, score and report on test-takers' responses, and to collate analytical data both on each test item and on the assessment as a whole.

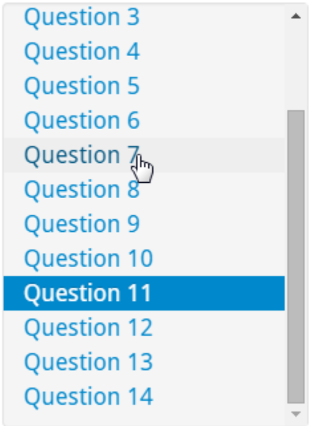
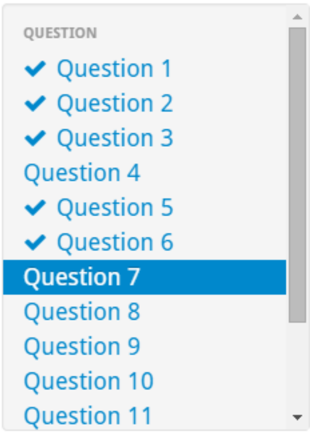
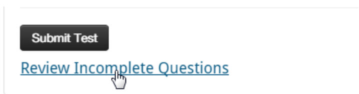
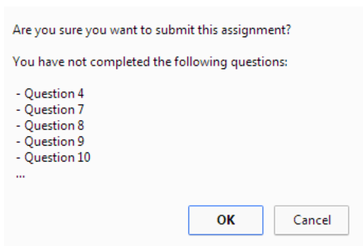
Test Creation and Delivery

Test "forms" are constructed of either 36 items (basic numeracy test) or 72 items (full diagnostic test) in either English or French (according to the order received) based on the test design criteria described in chapter 4. The sequence of items delivered to the test taker is randomized in order to avoid a negative response to the overall test potentially created by the test-taker encountering a complete block of items on a topic which the test taker finds particularly difficult.

The interface for the assessment has been designed to feel non-threatening and familiar for students who may not be used to computerized assessment. To this end, the project team felt it important to have the process feel analogous to the standard format of paper-based assessments where examinees can freely navigate between the questions in the assessment to get an overview of the scope, skip questions and return to them at a later time, and have a single submission event (rather than a submission of each individual question, which can lead to heightened levels of stress among examinees).

Following are illustrative screenshots from sample tests.



<table>
<tr>
<td>
</td>
<td>All the items in the assessment are listed in order, so that examinees can skip ahead or move back to any item in the assessment. This includes items that have already been responded to.</td>
</tr>
<tr>
<td>
</td>
<td>Items for which the examinee has entered a response are marked with a checkmark, making it easy to identify what has and has not been completed and reducing the chance of an involuntary non-response.</td>
</tr>
<tr>
<td>
</td>
<td>Once the examinee has reached the end of the assessment, he/she is provided with a final opportunity to review any of the responses they have entered before the assessment is submitted for evaluation, including any items which have specifically been left blank.</td>
</tr>
<tr>
<td>
</td>
<td>A warning message is displayed to students if they attempt to submit their assessment with incomplete questions.</td>
</tr>
</table>

Scoring and Reporting

Included in the item database are records of correct responses to each question. In the case of selected response (multiple choice) items, these will be the identity of the correct answer. In the case of constructed response (write-in) items, these may include a range of responses regarded as acceptable. For example, in response to the question "43.5 x 2 =   ", both 87 and 87.0 could be judged as equally correct. This use of more than one acceptable answer is of particular importance when both French and English numbering styles may be used, in which 11 013,5 and 11,013.5 may both be considered acceptable.

For each test-taker, the number of items answered correctly is expressed as a percentage and reported as such, along with the percentage of items not answered (the non-response rate). In the case of a basic numeracy test, these overall scores are reported. In the case of a full diagnostic test, these are supplemented by scores for each of the nine topics that make up the assessment framework. An illustrative screenshot of an individual test-taker's report on a full diagnostic test is shown below.

College Student Achievement Project
Mathematics Assessment - Field Trial

Full Diagnostic Test - Student Report

Results of your mathematics assessment:

	% correct	
Overall	**68%**	
Whole Numbers	100%	Congratulations!
Decimals	89%	Congratulations!
Measurement	83%	Congratulations!
Integers	67%	Remediation Available
Percent	64%	Remediation Available
Algebra	63%	Remediation Available
Arithmetic	60%	Remediation Available
Ratio	53%	Remediation Recommended
Fractions	50%	Remediation Recommended

% Non-Response	16%

Note that the test-taker is offered a "green light" and "Congratulations!" when the topic score is 80% or higher, an "amber light" and "Remediation Available" for a topic score of between 60% and 80%, and a Red light" and "Remediation Recommended" for a topic score below 60%. This report is provided to each test-taker immediately on submission of the test and the remediation suggestions are hyperlinked to the appropriate remediation module so that test takers can go directly there.

In the case of group tests, reports are also made available on a secure web site to the contact person whose name and email address is associated with the group (such as a school or college mathematics teacher or administrator). A partial sample of such a group report is shown below.

		Topic Scores									
	Overall Score	Whole Numbers	Arithmetic	Integers	Decimals	Fractions	Ratio	Percent	Algebra	Measurement	Non-Response Rate
Group Total	63%	73%	61%	77%	68%	61%	62%	51%	59%	57%	8%
Students (52)											
31731	34%	64%	10%	33%	10%	10%	57%	44%	37%	22%	59%
30965	37%	82%	10%	78%	56%	48%	8%	30%	39%	0%	15%
31737	37%	63%	60%	38%	56%	20%	0%	11%	47%	45%	0%
31743	38%	45%	38%	50%	11%	8%	33%	11%	55%	73%	19%
30968	40%	64%	50%	57%	78%	50%	43%	8%	13%	20%	35%
31724	41%	58%	30%	67%	25%	48%	31%	56%	35%	23%	2%
31740	41%	90%	60%	64%	40%	40%	15%	25%	38%	17%	34%
31736	42%	75%	70%	14%	33%	57%	29%	11%	45%	27%	34%
31732	43%	67%	70%	44%	20%	25%	47%	30%	37%	55%	45%
30963	44%	64%	30%	64%	57%	50%	38%	11%	40%	45%	4%
30964	45%	50%	50%	100%	60%	57%	0%	22%	34%	64%	18%

Data Analysis

Following both the item pre-test and the full ADP field trials, psychometric analysis of the test results was carried out (see chapter 7) and the results of this for each item added to the item database along with any other data collected for the item. This cumulative database of information can then be available whenever a review and revision of the assessment is undertaken. A sample item report following the item pre-test is appended here.

Item ID (source)	Item Variant
2183	4

Include	Exclude

Determine the lowest common denominator of: $\frac{1}{6}, \frac{1}{8}, \frac{1}{28}$

____

Item Tags:

2.8.A 5.1.A

Appears in Form Blocks:

FT1-B8

Psychometric Analysis:

Number of Examinees Attempting	IRT Difficult (b) Parameter	CTT Difficulty (p) Value	Item-Test (point-biserial) Correlation	NR Proportion (not included in number attempting)	Score=0 Proportion	Partial Score Proportion	Score=1 Proportion
80	2.62	0.05	0.34	0.22	0.95		0.05

Quantile 1: Score=1	Quantile 2: Score=1	Quantile 3: Score=1	Quantile 4: Score=1	Quantile 5: Score=1
0.00	0.00	0.00	0.00	0.21

Teacher Questionnaire Tally:

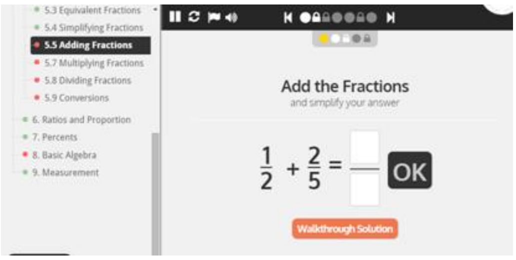
Appropriate	Too difficult	Too easy	Unfamiliar context	Unfamiliar vocabulary	Not previously taught
8	0	0	0	0	0

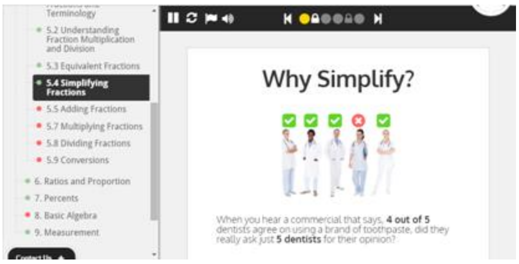
Chapter 6: Remedial Modules

The remedial modules are aligned to the nine topics of the assessment framework (as outlined in chapter 2) and are further subdivided into lesson modules, each encapsulating a specific aspect of the broader topic, as illustrated here.

Upon first accessing the remedial modules, students are immediately presented with a summary of areas of their strengths and weaknesses as determined from the assessment. Topics in which the assessment indicated the student has achieved satisfactory performance are marked with green lights, otherwise they are marked with red lights (as in the assessment reports shown in chapter 5).

The student can click on any module to display its lesson modules and access the relevant content.





Features of the lesson modules

The interactive lesson modules have the following features:

1. Voice track which interactively describe the action on the screen
2. Mastery-based lock system which guides students from introductory elements to high levels of performance, with completion of the module being achieved upon reaching the end of the scenes in the module
3. The ability to pause and replay sequences
4. Introductory videos which present complex or abstract concepts in a familiar way
5. Interactive learning tools which deconstruct problems and provide students with flexibility in thinking through problems, while giving them the ability to make mistakes (and learn from those mistakes)
6. Specialized input and feedback mechanisms for various numeracy concepts that use common errors as a teaching opportunity
7. Interactive walkthroughs of complex processes and algorithms

8. Mastery-based exercises which serve as checkpoints in the student's learning, and serve to boost confidence and the growth mind-set
9. Tracking tools which log data, optionally anonymized, on student responses and interactions across all the modules

The CSAP Remedial Modules employ an adaptive learning methodology to respond to student remediation needs at the post-secondary level. The process starts with students completing the calibrated CSAP assessment to identify strengths and weaknesses in the nine mathematics topics. Each of the 9 topics is further divided into a set of sub-topics. These are highlighted as problematic (red dot) or as satisfactory (green dot) depending on a student's achievement. The interfaces encourage a linear progression through the modules from fundamental concepts (e.g., Whole Numbers, Arithmetic) to higher level concepts (e.g., Algebra, Percentages, and Measurement). Meanwhile, students are given the freedom to explore the modules in whatever way makes the most sense for them, giving them the sense of control which will prove important as they continue with their post-secondary studies.

Upon the completion of all of the remedial modules, students should be better prepared to undertake a first-year mathematics course in an Ontario community college. To this end, the remedial modules are designed to convey the necessary knowledge, develop the necessary skills, and support the necessary positive attitudes which students will need for college success.

Student Demographics

The project team consulted with educators across the college system to develop a working profile of the students who would use the remedial mathematics modules. Various factors may contribute to a lack of success in a remedial program, and the CSAP project team worked to develop a solution which was sensitive to the diverse student demographic.

Students may enter their college studies as recent Ontario secondary school graduates, as mature students, or as international students. Each of these three groups contain considerable internal diversity, but also embody a few characteristics which were considered to ensure that the resources were developed to support their needs.

Students arriving as recent Ontario graduates have received instruction using the current mathematics curriculum and are assumed to have successfully completed a mathematics course in the preceding two school years. (Mathematics is only a required course up to grade 11 and students may be accepted into college programs with those minimal prerequisites having not taken a mathematics course in their final year). Still, the homogeneity of students' abilities in this category should not be assumed. They arrive with considerable differences in their mathematics abilities (with clear demarcations between students who studied in the academic as opposed to the applied streams of the mathematics curriculum and with varying degrees of success, as evidenced by the CSAP research). The pedagogical considerations by which these students were taught is also unknown.

Mature students, by comparison, may have been away from an academic setting for several years, but tend to come with considerably more work experience which may have added relevance to their mathematics education, such as managing the finances of a household, or being immersed in contextual mathematics on the job and to their desire to succeed in their chosen college programs.

International students represent one of the most diverse groups attending Ontario colleges. For many, English or French was not the language of instruction when they originally underwent their mathematics schooling, and the language may still represent a general communication hurdle in explaining concepts or describing situations. The most fundamental mathematical terminology (e.g., numerator, denominator, factorization, product, and sum) may be unfamiliar or have slightly different meanings. Number formats used in Canada may also differ from those formats used in their home countries (e.g., the use of lacks in India). Finally, methods used for solving basic math problems may be unfamiliar. This required the project team to keep math methodology generalizable enough that someone coming from a different context of mathematics education would gain from understanding the underlying methods for performing operations and transformations in arithmetic and algebra, as opposed to being confused by static solutions which do not conform to their understood way of doing things.

Math Anxiety

For all students, we can be nearly certain that a certain amount of mathematics education has been a requirement for a large part of their education. Students who are not comfortable with mathematics will nonetheless be exposed to it and evaluated on it on a regular basis. From this experience, many students come to see mathematics as a subject which devalues their intelligence or worth as a person. This social phenomenon, often referred to as *math anxiety*, has been measured in various contexts and the project team's communication with partnering colleges hinted at a widespread challenge in attitude. The remedial system makes use of various techniques to alleviate math anxiety and build lasting confidence in mathematics, with some of general principles described below:

1. Effective pedagogy can make unfamiliar ideas simple and usable. All activity in the lessons is prefaced with an opportunity to gain an intuition for the concept that is being used to solve problems.
2. Students are provided with an opportunity to start with straightforward, tangible examples to gain confidence in their abilities before proceeding to more challenging areas. Students however have the option to skip over some of the material in order to demonstrate their mastery at key points in the lesson.
3. Many students don't realize that trial and error can be a valid approach to certain mathematical problems (despite the fact that it is included as a stated expectation in the Ontario Mathematics curriculum). The remedial modules provide students with tools which allow for guided trial-and-error exploration which develops a student's sense of number as well as demonstrating it as a valid approach in certain circumstances.
4. Students are instructed in using the most effective techniques for perform computations and solving problems using pencil and paper methods. However, where these techniques

may be complicated for students without the necessary background, we also expose them to simpler computer-assisted techniques to develop a good intuition of the problem.

5. For computational algorithms and problem solving processes which contain many steps (e.g., order of operations calculations), students are provided with interfaces which allow them to work through the various steps through the digital interface, allowing the system to gently correct any errors that are made through the multi-step process. In traditional remedial programs, students are forced to complete the entire procedure on their own, leaving many opportunities to make errors and become frustrated.
6. Wherever it can be done, while retaining the item's validity, students would be provided with more than one attempt at answering a question, so that if the first attempt is incorrect due to a minor mistake, the student can self-correct and recuperate their progress.

Prior Knowledge

Students often have some level of familiarity (albeit superficial) with all of the concepts, and may be unmotivated to study such basic concepts which they require for success in their post-secondary studies. Careful design of the mastery-based resources ensure that students can see the value of what they are learning as quickly as possible and focus on those areas which are most relevant to their own needs. Items were also carefully selected to ensure

The project team consulted with experts on the secondary school mathematics curriculum on the ADP/LODP Advisory Committee to identify topic areas which may be taught differently than in traditional college math classes. This also helped the project team on developing interactions which follow the latest pedagogical research insights (much of which is reflected in the latest curriculum revision). This consultation highlighted differences in the methods used for performing basic operations such as addition, subtraction, multiplication, and division. There is also some terminology that is traditionally used when solving algebraic problems (such as *cancelling* terms, which was removed in a prior revision of the Ontario mathematics curriculum) which has been phased out of practice in the Ontario secondary school curriculum.

Voiced Animated Activities

For many students, weak reading skills contribute to difficulties in learning mathematics. This is one of the big barriers in any self-paced remediation program aimed at college students. A common practice in Ontario colleges, involves the use of paper-based remedial workbooks which require the student to rely solely on written explanations of descriptions and scenarios.

The remedial modules make use of voice to guide students through concepts while holding their attention with a human touch. Use of audio frees the student to stay focused on the task at hand without having to read various snippets of text which are available to guide the learning process.

Real-World Examples

One of the largest obstacles for students learning mathematics is in seeing the relevance of the subject matter to one's own life. Although there is a place for focusing on mastering the abstract and universal notions of mathematics through its notation and other abstract representations, many students are coming to the system without the prerequisite experience to see any value in such activities. The project team engaged in dialogue with foundational math teachers to see what works with their students, and proceeded to produce rich multimedia experiences which made mathematics concepts relevant.

Real world scenarios are used for introductions to concepts so that the parameters of ideas can be framed in more tangible terms. These scenarios are also used as a form of activity / assessment of student learning throughout the modules.

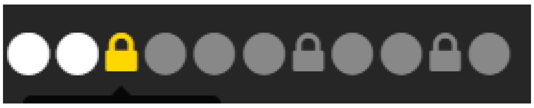
The modules contain various examples of ethno-mathematics to provide a broad context for the use of mathematics in the real-world. This reinforces a larger vision for not just mathematics, but of the diversity of cultures and people surrounding the student.

Considerations for Language Learners

The development team designed the modules to support the needs of language learners (e.g., ELL). This was largely accomplished by combining multiple representations which allows students to develop their knowledge of terminology in a natural way. Simple descriptive language is used in combination with formal mathematics terminology and textual descriptions are combined with text and visualizations to ensure maximum clarity of communication.

Mastery-Based Learning

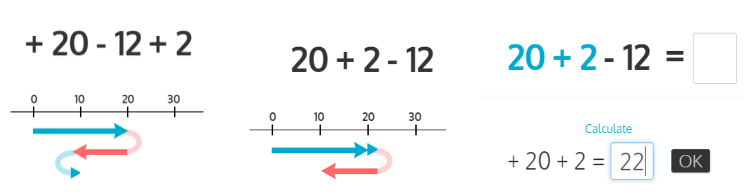
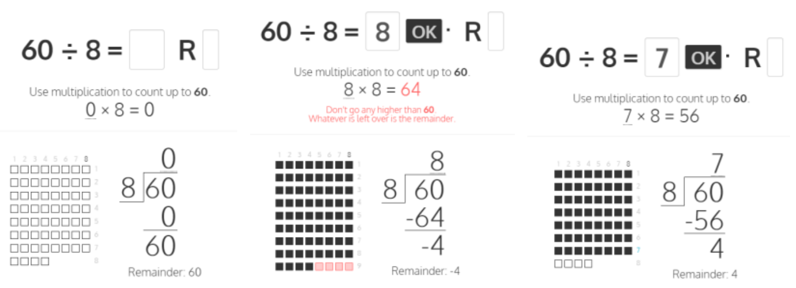
The remedial resources are built according to mastery based learning principles. This is both to support the proper development of ideas with the student, ensuring they have a solid foundation for every new concept that they are presented with, but also provide a more meaningful measure of progress to students. In the diagram below, the circle represent optional instructional scenes and the locks represent key checkpoints where students must demonstrate knowledge in order to proceed. Once all of the locks in a module have been "unlocked", the student is provided a completion mark for the module.



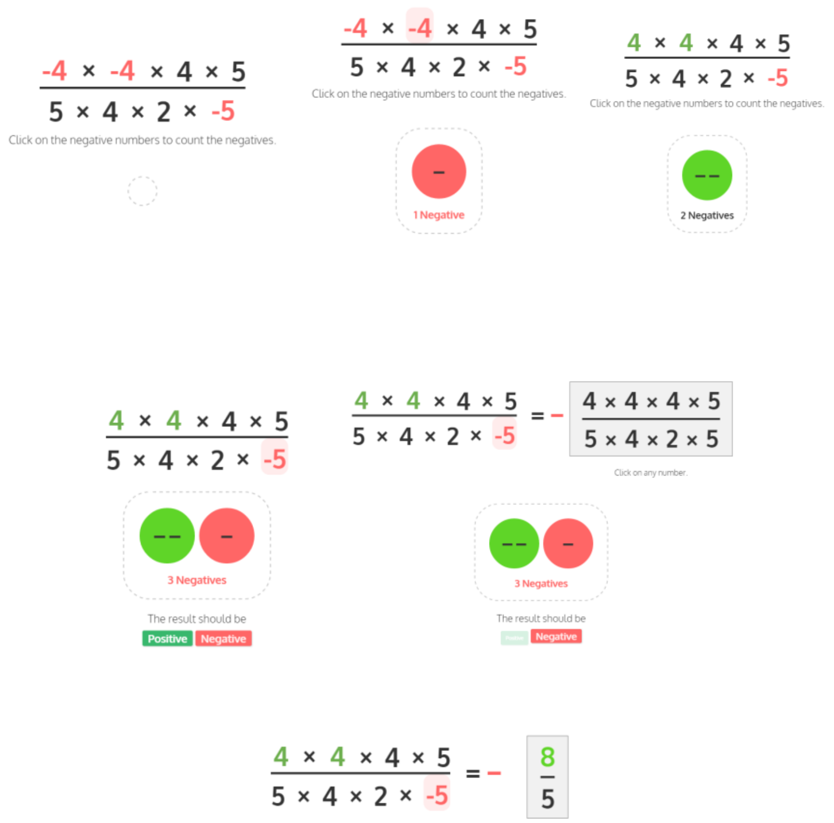
A subordinate level lock-based interface is presented to students when they are completing exercises within the system to provide finer granularity of progress over variations of the same activity. This allows students to skip over simpler variant in order to demonstrate mastery of more complicated versions.

Key Topical Pedagogical Interfaces

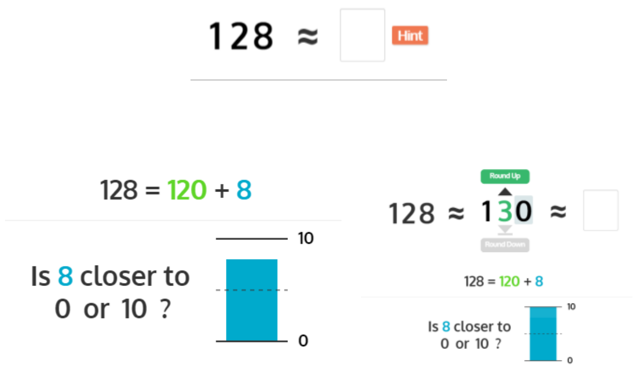
This section describes the interfaces that have been developed to support student learning in the various topical strands covered by the CSAP remedial resources.

Addition and Subtraction Using Decomposition	Compute 
Long Division using Multiplication	Compute the division 

<table>
<tr>
<td>Finding the Difference using Decomposition</td>
<td>
Evaluate and simplify

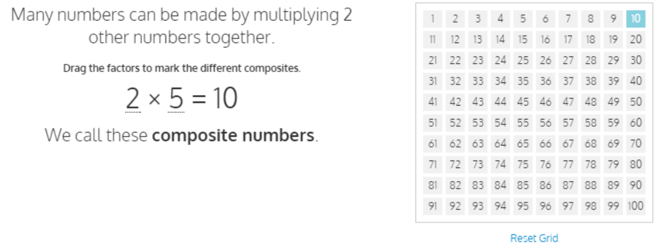


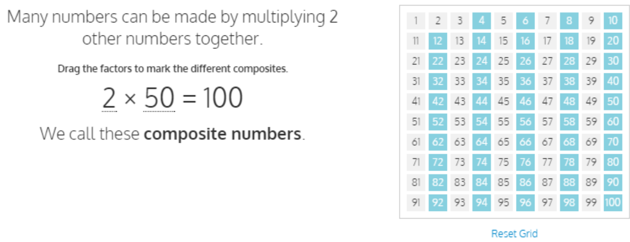
</td>
</tr>
<tr>
<td>Decomposition of the Rounding Algorithm</td>
<td>
Round to the nearest ten

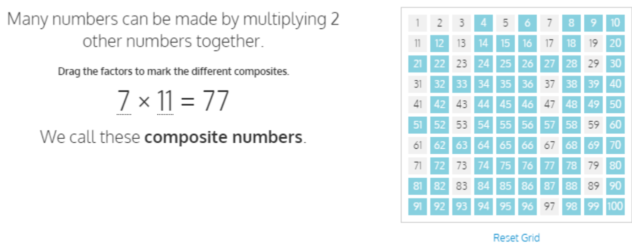


</td>
</tr>
</table>

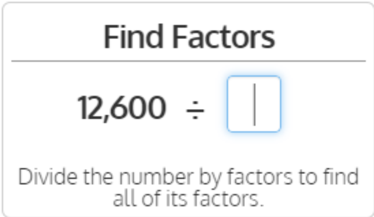
Prime Numbers	Identify composite numbers and prime numbers

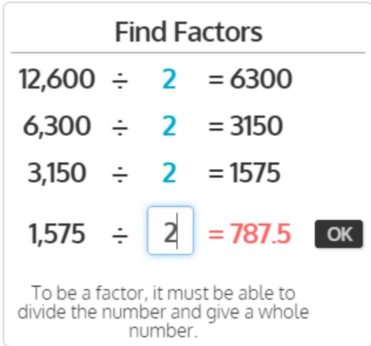


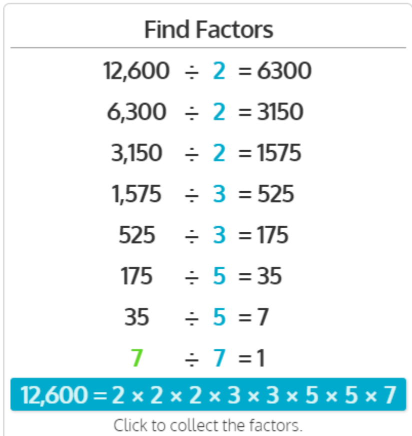


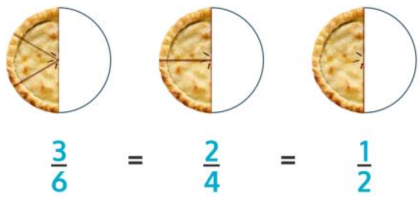
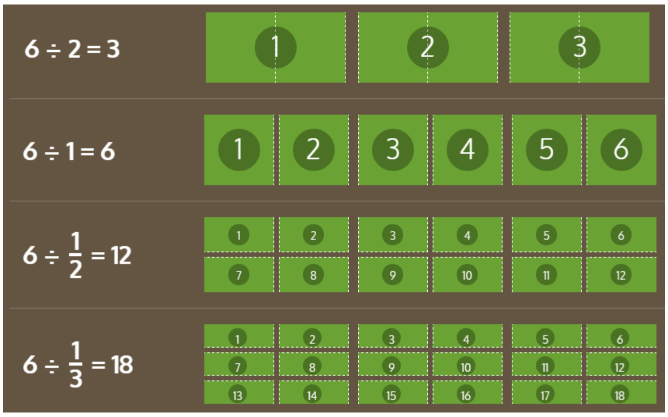
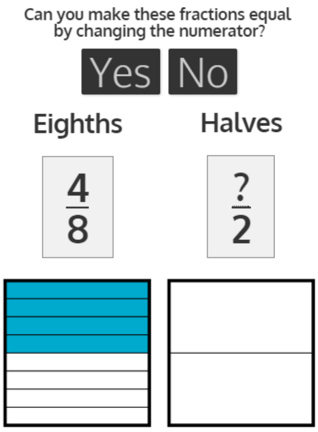
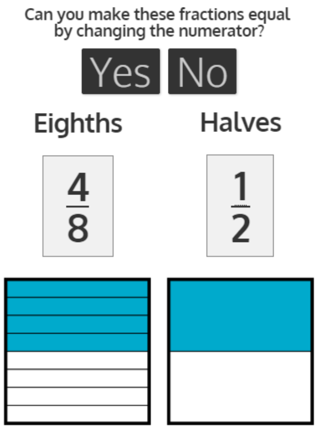


Prime Factorization	Find the prime factors of a number

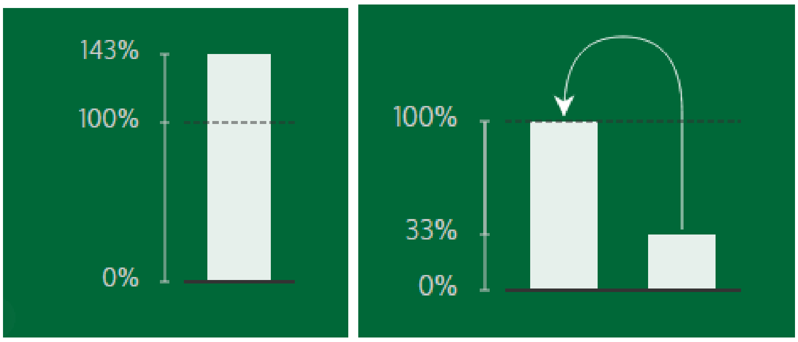






Multiple Representations of Fractions	Assorted selection: representation of fractions    

Order of Operations	Evaluate the operations $\frac{2+3\times4-4}{5+4}$ $\frac{2+3\times4-4}{5+4}$ $\frac{2+3\times4-4}{5+4}$ $=\frac{2+12-4}{5+4}$ $=\frac{14-4}{5+4}$ $=\frac{10}{5+4}$ $=\frac{10}{9}$
Grouping Like Terms	Group the like terms $4a+b+3a+2b$ $4a$ $+b$ $+2b$ $+3a$ $4a$ $+3a$ $+1b$ $+2b$

<table>
<tr>
<td>Percentage Bar</td>
<td>Calculate a percent increase of 43%... of 33%

</td>
</tr>
</table>

Modes of Use

The remedial modules are designed to be used in combination with CSAP Assessment and to support in the learning of the concepts in that test. This has proven to be an effective pairing for both pre-entry remediation which students might complete at home or in “boot camp” sessions hosted by their home colleges.

A limited number of students used the resources as a part of their first year mathematics course. For those students who were enrolled in remedial courses, this provided an opportunity to go above and beyond the classroom experience for a personalized learning experience which augmented their learning in the course. Other students made use of the resources while engaged in higher-level math courses. In these cases, students were generally required to complete the remedial program in the first few weeks of class to create a solid base for learning.

Chapter 7: Analyses of Field Test Data

The purpose of this chapter is to discuss in detail the implications of the statistical analyses conducted on the complete set of field test data including suggestions for future directions for the ADP project.

The results of the cumulative statistical analyses conducted over the past year hold notable implications for form development, test dimensionality, and cut score positions. Additionally, the results of these analyses identified items within each topic that are near the borders between the red, amber, and green levels of performance and so might be useful for communicating with test takers about the knowledge and skills required to meet or exceed these scores.

In our meeting on November 26, 2014, several important questions were raised in regard to the test:

- How relatively difficult are the items?
- What can be said about the dimensionality of the test given that the analyses to date have been based on the assumption that the test is unidimensional (i.e., measures overall mathematical knowledge and skills)?
- Are some items differentially difficult (and so, potentially biased) by gender or first language?
- How can the multi-topic items best be used?
- How many items in each area are needed for reliable results?
- Where should the borders (cut scores) between the red and amber and between the amber and green levels of performance be?

Item Difficulty

The item response theory (IRT) analysis provides an estimate of item difficulty, *b*, that is independent of the ability level of the examinees who happen to have taken the item – in other words, where the *p* value from CTT tells us what proportion of examinees in the particular field test group answered the item correctly, the *b* value from IRT has already adjusted for differences in ability between the field test groups. In these analyses, the simplest IRT model, estimating only item difficulty, was used (this model is mathematically equivalent to another common model, the Rasch model).

In IRT analysis, the item characteristic curve (ICC) is a logistic regression line that relates the probability of an examinee correctly answering an item to the trait or ability measured by that item. Figure 1 provides some example item characteristic curves of varying degrees of difficulty. The *b* parameter can be interpreted as the ability level at which 50% of the examinees with that ability would answer the item correctly – or a single examinee with that ability would have a 50% probability of answering the item correctly. Items with lower *b* values are easier and items with higher *b* values are more difficult.

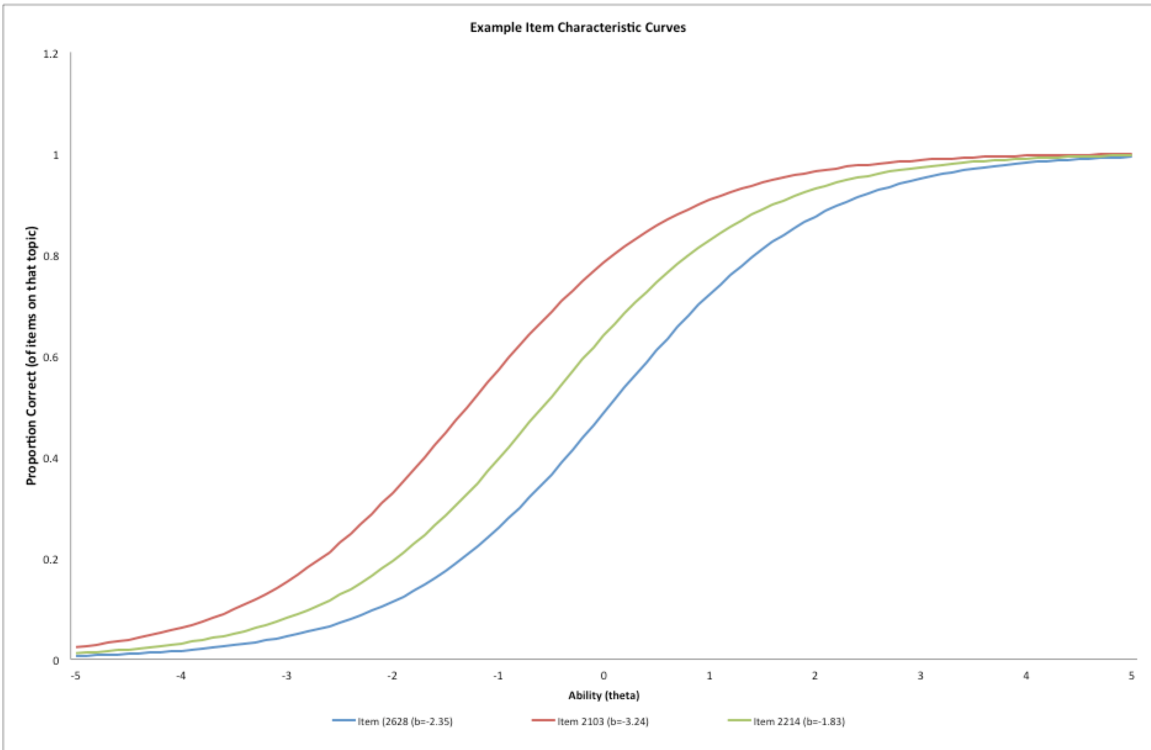


Figure 1. Example item characteristic curves.

Dimensionality Analyses

The test is designed to measure knowledge and skills in nine topic areas. An important question to consider is whether examinees' performances are consistently strong or weak across all topic areas, which would suggest there might be one underlying dimension of mathematics knowledge and skill, or if examinees' performances vary across the topics, suggesting there may be multiple dimensions. Unfortunately, the relatively small number of examinees who have provided responses to each of the items in the pool prevented us from using the traditional approaches of exploratory or confirmatory factor analysis to investigate the dimensionality of an instrument.

Instead, test dimensionality was explored indirectly through examining the infit mean square (hereafter referred to as infit) statistics derived from the item response theory (IRT) analyses using Construct Map Version 4.6 beta (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2008). The infit for an item is defined as the ratio of the variance of the observed residuals over the variance of the expected residuals. When the infit is close to 1, the observed residuals of the items vary as expected. Large infit values are indicative of more observed variance than expected, meaning that the item does not measure the underlying ability as well as other items. Moreover, items with large infit values contribute less to measurement of the overall ability of interest and should be flagged for removal or correction. Very small infit values are equally problematic as they indicate that these items do not show enough variation for meaningful measurement over a broad range of ability levels.

Dimensionality analyses were performed in two ways: (1) items for all topics together and (2) items for each topic separately. If the infit statistics for the items were much smaller when the topics were analyzed separately, that would suggest that the topics that showed this pattern measured a dimension that is different from the overall dimension. While the infit statistics for the separate analyses were a little smaller, the difference was not great. (It is worth noting that, because of the relatively small number of responses per item and the sparseness of the data, the infit statistics did not suggest very good model fit in either analysis.)

Infit mean squares are influenced by unexpected patterns among more average observations. Adams and Khoo (1996) suggest a range of 0.75 and 1.33 as acceptable infit values. The infit values obtained from the first analysis ranged from 1.46 to 18.1. Infit values obtained by running the items by topic yielded a range of 1.29 to 11.79, with one item falling within the acceptable range.

Differential Item Functioning

Differential item functioning (DIF) analysis is typically used to identify test items that are differentially difficult for respondents who have the same level of knowledge, skill, or ability but differ in ways that should be irrelevant to their performance on the test such as gender or native language (Miller, 2010).

In order to determine if some items were differentially difficult (and potentially biased) by gender and or first language, DIF analyses were conducted using DIFAS 5.0 (Penfield, 2007). Unfortunately, in spite of running for several hours, DIFAS was unable to compute the indices needed for DIF. We also re-ran the DIF analysis only on those items that received the largest number of responses, unfortunately to no avail. We suspect that this is due to the sparseness of the data. As such, we recommend conducting DIF analyses once additional data are collected to look for evidence of potential bias.

Number of Items per Topic

The standard error (the blue lines in Figures 2 and 3) of the ability estimates were computed based on the results of the IRT analyses. In addition, the IRT analyses yielded test characteristic curves ((TCCs, obtained by summing each of the ICCs across the entire continuum of ability level; the red lines in Figures 2 and 3) showing the relationship between the proportion correct and the IRT ability scale. Figure 2 shows the standard error of the ability estimates and the test characteristic curve for all 37 items that measure only Topic 5.

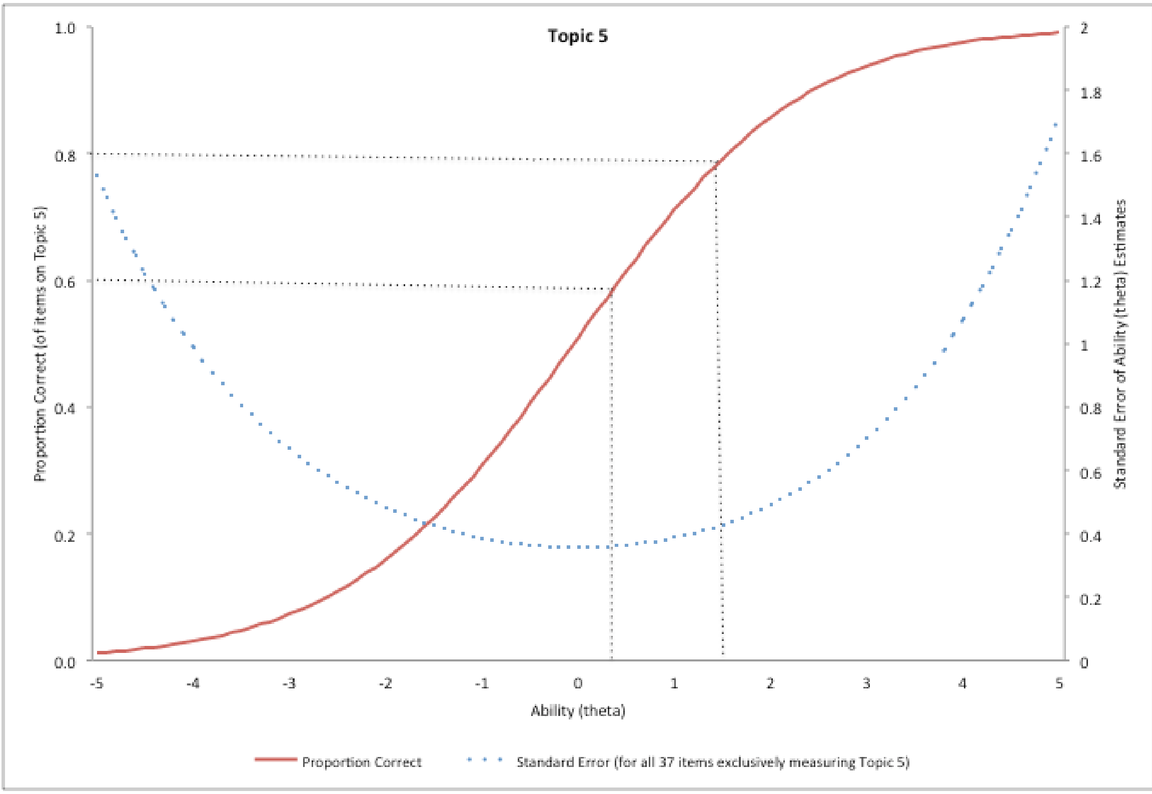


Figure 2. Test characteristic curve and standard error of ability estimates for the 37 items measuring Topic 5.

Figure 3 shows the standard ability estimates for 8 items and 4 items. Standard error of ability estimates progressively decreases, as more items are included in Topic 5. This trend is also observed among the other topics. The standard error is indicative of how certain we can be when estimating the examinee's placement on the ability (theta) scale where large standard errors means less certainty. As the lines for the easiest and the hardest items show, it is possible to select where on the ability scale precision is most important and then choose items of appropriate difficulties to minimize the standard error around that ability level.

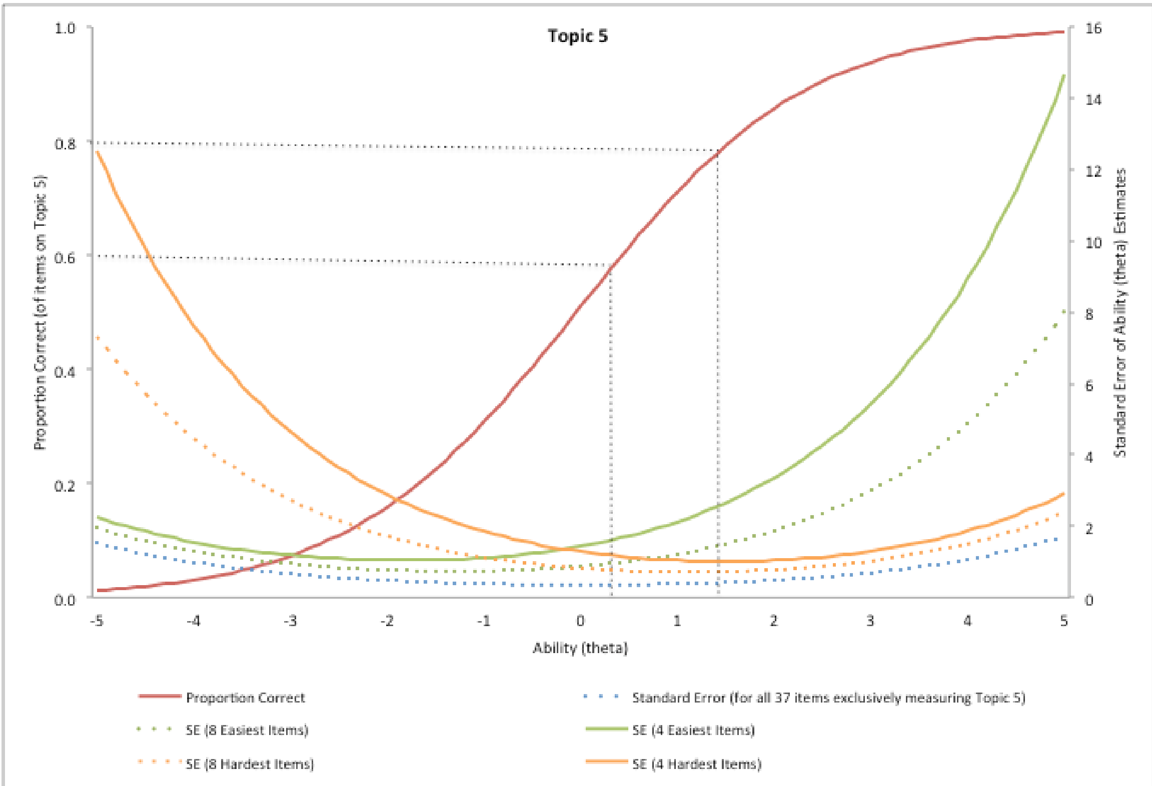


Figure 3. Test characteristic curve and standard error of ability estimates for different combinations of items measuring Topic 5.

Cut Scores

Establishing the accuracy of the test is essential in order to determine cut scores. In its use as a pass or fail test for qualification purposes where there is a single cut score, an efficient test can be developed by selecting items whose difficulty is near that of the cut score in order to have a smaller standard error near the cut score. For this test, the cut scores were first determined on the 'Proportion Correct' scale, and then translated to the IRT scale using the test characteristic curve.

For example, Figure 4 shows that in Topic 1, items ranging from a difficulty level of -0.5 to 0.5 are most useful given the required 0.6 - 0.8 proportion correct range and that at least 8 items at this difficulty level are required to keep the standard error of the ability estimate at a minimum.

Figure 4 also shows that the reliability of the ability estimates decreases (i.e., there are higher standard errors) for more difficult items. This is a trend not limited to only items within Topic 1. This has been consistently observed over all 9 topics.

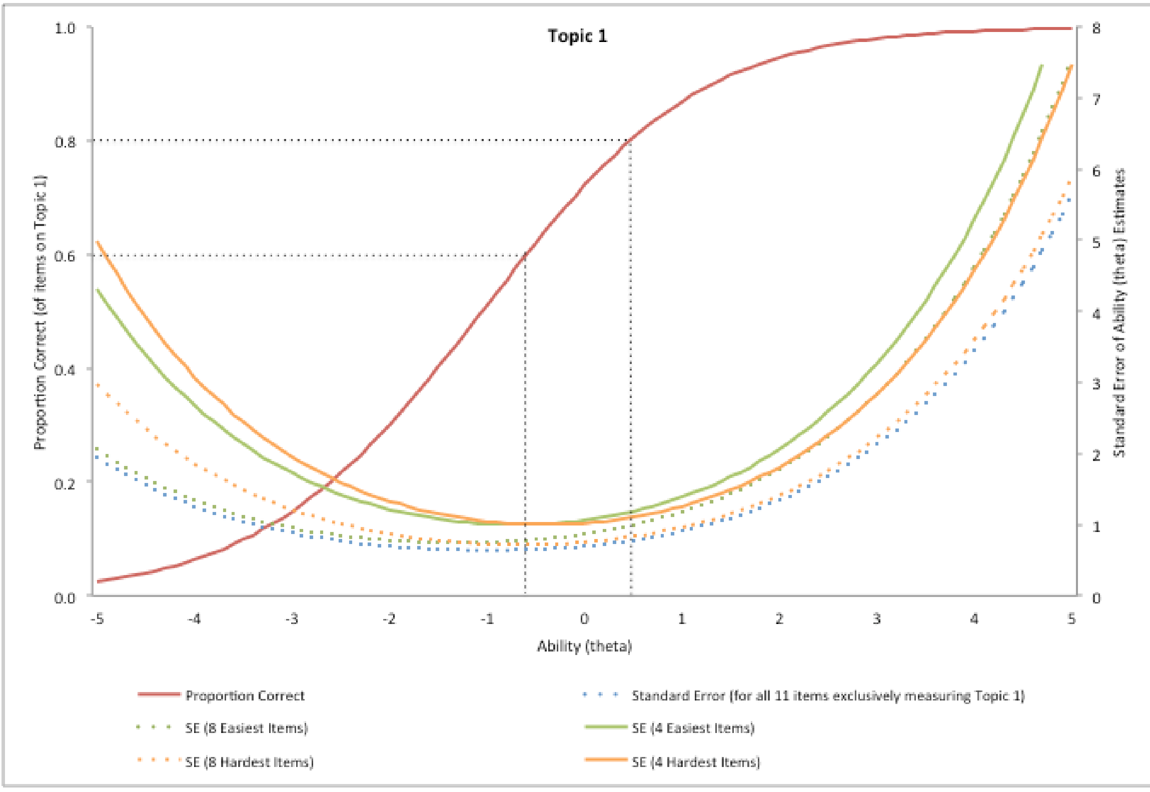


Figure 4. IRT scale with proportion correct and standard error of ability for items in Topic 1.

Future Directions/Considerations

Taken together, the analyses have shown the need to collect more data to be able to conduct the more complicated analyses of dimensionality, DIF, and CDA.

While the wide range of infit values suggests that at least some of the items should be reviewed for their content validity, it is important to mention that infit analyses were used as a proxy measure of dimensionality in place of the more conventional approach of exploratory and confirmatory factor analysis that would provide richer information about the degree of unidimensionality of the items.

The conventional approach of exploratory and confirmatory factor analysis would provide a detailed analysis of the factor structure or the dimensionality of the test. DIF analyses can also be conducted with a large enough set of data to have enough respondents per item and sufficient overlap of item sets between people or randomly equivalent groups of respondents per item. This would also provide information about items that may be biased due to gender or native language. Finally, a larger set of data would allow for the development of cognitive diagnostic models (CDMs) and a Q-matrix allowing cognitive diagnostic assessment (CDA), informing the diagnostic purpose of the test. CDMs combine responses across items to determine the profile of skills associated with each examinee through computing the probability that each examinee has each set of skills needed to correctly answer an item. The Q-matrix consists of 1s and 0s and depicts the skills each item is intended to measure (see

Figure 5). Specifically, CDA would provide an examinee with detailed information about the particular skill or skills they may lack based on the response to an item. This information could then be used to plan remedial training.

Item ID	Number of Topics	Topics								
		1	2	3	4	5	6	7	8	9
I1988	2	1	1	0	0	0	0	0	0	0
I1989	1	0	1	0	0	0	0	0	0	0
I1990	1	0	0	0	0	0	0	1	0	0
I1991	2	1	0	0	0	1	0	0	0	0
I1992	3	0	0	0	1	1	0	1	0	0
I1993	1	0	0	0	0	0	0	1	0	0

Figure 5. An example of a Q-Matrix.

Furthermore, the results of the cut score analyses revealed the need for a greater number of items within the red to green range of ability levels. As such, a next step would be to develop item variants of current items that lie within the red to green range.

Another possible future direction is to develop a validity argument (Kane, 2002), specifying how the analyses conducted to date – and those planned for the future – support the interpretation and use of results from this test.

The results of these analyses also may have been affected by motivation, fatigue, and boredom of the examinees. Specifically, the results may be underestimating the average ability level of the examinees. As such, the results of these analyses should be considered with these factors in mind.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest: Interactive item analysis*. Melbourne, Australia: Australian Council for Educational Research.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41. DOI: 10.1111/j.1745-3992.2002.tb00083.

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2008). *ConstructMap v4.4.0: Quick start guide.* Berkeley, CA: Berkeley Evaluation & Assessment Research (BEAR) Center.

Miller, T. (2010). Detecting differential item functioning and differential step functioning due to differences that *should* matter. *Practical Assessment, Research & Evaluation, 15*(10), 1-13.

Penfield, R. D. (2007). *DIFAS 4.0 user's manual*. Retrieved March 8, 2008, from http://www.education.miami.edu/facultysites/penfield/index.html.