Difficulty Study: Field test of *Insights 6*

Research Institution: Trent University

Principal Investigator: Dr. Catherine D. Bruce

Research Team: Carolyn Brioux, Dr. Catherine Bruce, Tara Flynn, Dr. Lynn Kostuch

Final Report: March 23, 2013

Title: Difficulty Study: Field test of Grade 6 Success Resource

Research Institution: Trent University
Principal Investigator: Dr. Catherine D. Bruce
Research Team: Carolyn Brioux, Dr. Catherine Bruce, Tara Flynn,
Dr. Lynn Kostuch

Final Report: March 23, 2013

I. BACKGROUND

This research study involved a brief field test of a bank of mathematics assessment items, developed by Pearson, that are similar to those on EQAO Grade 6 mathematics tests, the Ontario standardized test for student achievement. As the dominant metric for all elementary schools in Ontario, EQAO tests and related results impact school, district and provincial planning, and provides a pulse of how students are achieving in mathematics. Pearson Canada asked the Trent Mathematics Education Research Team (Dr. Catherine D. Bruce and her research assistant Tara Flynn) to field-test a bank of Grade 6 test items to determine whether test items generated by Pearson were of equal difficulty to the Grade 6 EQAO test of 2011. As additional points of interest, the research team was interested in whether the items were helpful to teachers in making instructional decisions, and/or predicting student achievement. The field test was conducted in northern Ontario in one district school board in the fall of 2012 over a 6 week period. This report summarizes the results of the difficulty study including: methods, data collection and analysis, findings and recommendations.

II. METHODOLOGY

Overall Methodology of the Difficulty Study

Pearson Canada generated a bank of test items in parallel to the 2011 EQAO test items with efforts to ensure that the level of difficulty of each parallel item was equal to that of the source item from the EQAO test. In order to determine whether the newly generated Pearson items were of the same level of difficulty as those administered through EQAO, researchers from Trent University tested the items using a 'difficulty study' methodology. In this case, we used classical test theory to determine test equivalence. The definition of equivalence is multi-dimensional in this study because there were three interconnected types of equivalence:

a. Equivalence of the overall mathematics content coverage - the same types of content and strand concepts that have been mapped to the Grade 6 curriculum are similarly addressed on the Pearson generated items as those of the 2011 EQAO assessment

b. Equivalence of the length of the tests, so that each test had the same number and type of questions (the same number of multiple choice, open response and closed response items) and equivalent format of the tests.

c. Equivalence of each item across the test forms: an item by item analysis to ensure that each EQAO question is operating in parallel with two Pearson generated items and is of the same difficulty to complete.

Classical test theory has developed procedures for matching content coverage (judges match the items to a test blueprint) and researchers use statistical procedures for measuring whether differences in test scores are statistically significant. Authoritative texts on test theory that support the research methods of this study are: 1) Linn, R. L., & Miller, M. D. (2005). Measurement and assessment in teaching, 9/E. Upper Saddle River, NJ: Pearson; 2) Schmeisser, C. B., & Welch, C. J. (2006). Test development. R. L. Brennan (Ed.), Educational Measurement. 4E (pp. 307-335). Westport, CT: American Council on Education/Praeger.

The central research question was as follows: Are test items generated by Pearson, as parallel EQAO test items, comparable in difficulty? [Do students perform equally well on the Pearson items as on the EQAO items with no instruction or interventions?]

Researchers hypothesized that

i) test items generated as parallel EQAO test items would be comparable in difficulty, with some possible differences when contexts were distinct (e.g., temperature changes versus height of snow changes);
ii) responses to these test items would support teachers in providing students with explicit feedback and instruction for improved learning;
iii) should the test items prove to be of the same difficulty, scores on test items administered prior to EQAO may help to predict achievement on EQAO (with appropriate item weighting strategies that mirror those of EQAO);
iv) students from different regions of Ontario may not respond in the same ways to EQAO type questions.

Collaboration

The research team partnered with the Lakehead DSB consultant team to ensure smooth participation of teachers and students and to provide some professional development for teacher participants.

Test Development

Members of the research team reviewed Pearson generated items for mathematics content and match to EQAO items from the 2011 provincial assessment. In cases where the research team were not confident that the match or the math content was close enough, the research team made suggestions to the Pearson team for revisions. These revisions occurred over a two-week period via electronic communications. Once the test items were revised, the research team designed 6 test forms as follows: Test A: 8 multiple choice items, 1 closed response item, 2 open response items from the EQAO 2011 assessment.

Test B: 8 multiple choice items, 1 closed response item, 2 open response items generated by Pearson deemed to be equivalent to the EQAO 2011 assessment items selected for Test A.
Test C: 8 multiple choice items, 1 closed response item, 2 open response items generated by Pearson deemed to be equivalent to the EQAO 2011 assessment items selected for Test A.
Test D: 5 open response items from the EQAO 2011 assessment.
Test E: 5 open response items generated by Pearson deemed to be equivalent to the EQAO 2011 assessment items selected for Test D.
Test F: 5 open response items generated by Pearson deemed to be equivalent to the EQAO 2011 assessment items selected for Test D.

Once the tests were format-generated by the Trent research team, the Pearson team then reviewed each test in detail to ensure that all diagrams and wording were exact and accurate. Revisions were made by the research team and reviewed a second time by the Pearson team. Once approvals were in place, the research team set the test for scanning purposes and began to prepare materials for implementation of the difficulty study field-test.

III. METHODS OF THE DIFFICULTY STUDY IMPLEMENTATION

Tests were delivered to the Lakehead DSB in person by two Trent researchers who then provided the teacher participants with professional development related to EQAO activity in general and to the difficulty study in particular. This included analysis of EQAO results in the district and discussion about mathematics programming, instruction and testing practices. (Dr. Bruce's Powerpoint slides available upon request)

The teachers then took the assessment packages with them back to class for administration. This test administration window was approximately 5 weeks in the Fall of 2012 (from October 24 to November 27, 2012).

At the PD session, each participating teacher received a package containing:

1. A letter explaining procedures and thanking them for their time and involvement;
2. Parent/guardian information and consent letters;
3. Student information/assent letters;
4. A student ID tracker sheet;
5. Class sets of two tests.

All photocopying was prepared for teachers (no copying costs required by teachers). A cover letter provided with each packet explained the procedures for test administration in detail. Each test was assigned an ID number and students were not required to write their names on the tests. The tests were pre-packaged as Day 1 (with an assortment of the two tests) and Day 2 (with an assortment of the two tests) to ease the distribution of tests - the teacher did not have to figure out which student would do which test on which day.

Teacher participants were instructed to administer the tests to their students on two days back-to-back with no instruction between the test administrations. For each class, students were divided in half. On Day one in Class I for example, one half of the class responded to items on Test A and one half completed Test B. On Day 2, and with no instruction between, Class I students were switched so that those who wrote Test A on day one wrote Test B and students who wrote Test B on day one wrote Test A. Each student was only be expected to take two tests (Test A and B or Test A and C with 8 multiple-choice and 3 open-response items per test OR Test D and E or Test D and F). The tests required between 20 minutes and 60 minutes to complete.

Implementation Schedule:
Cluster One

Class I		Class II	
Day 1	Day 2	Day 1	Day 2
½ class: Test A	½ class: Test B	½ class: Test A	½ class: Test C
½ class: Test B	½ class: Test A	½ class: Test C	½ class: Test A

Cluster Two

Class I		Class II	
Day 1	Day 2	Day 1	Day 2
½ class: Test D	½ class: Test E	½ class: Test E	½ class: Test F
½ class: Test E	½ class: Test D	½ class: Test F	½ class: Test E

Data Set (focus on Tests AB and AC)

Grade seven students at nineteen schools in the northern Ontario school district completed two of three different math assessment forms consisting of a mix of multiple choice and open-ended questions. The three assessments were named Assessment A, Assessment B and Assessment C.

To review, Assessment A comprised eight multiple-choice items and three open-response items from the 2011 EQAO mathematics assessment. Assessment B contained eight multiple-choice items and three open-response items designed to match the content and structure of the items from Assessment A. Assessment C contained a third set of eight multiple-choice items and three open-response items, also designed to match the content and structure of Assessment A.

The study design assigned names to the two groups of students, Group AB (N = 202 students) and Group AC (N = 204 students). Assessment A was common to both groups, so that each of the two groups answered the same questions on that form. Group AB then completed Assessment B. Group AC completed Assessment C. The test design is detailed in Figure 1.

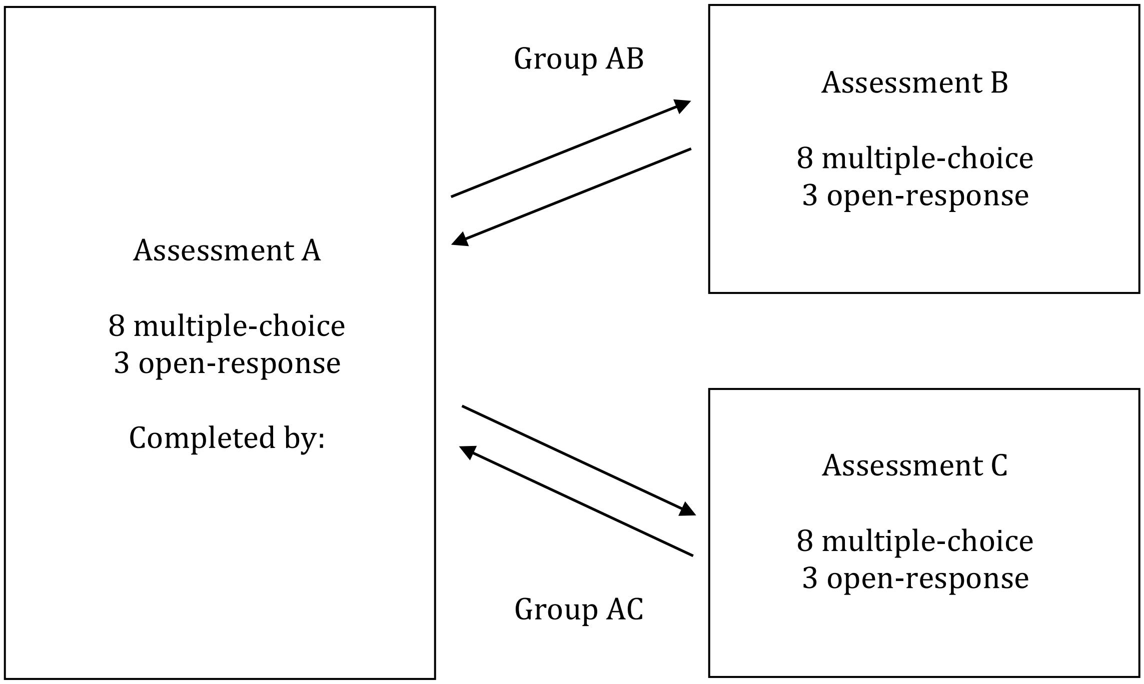


Figure 1: Difficulty study field test design

Two hundred and two (202) students were initially assigned to Group AB. One hundred and sixty-five students completed both Assessment A and B (82%). The students who did not complete both assessments were removed from the sample.

Two hundred and four (204) students were initially assigned to Group AC. One hundred and twenty-six students completed both Assessment A and C (62%). The students who did not complete both assessments were removed from the sample. The initial and final sample numbers are detailed in Table 1.

Table 1.
Initial and final sample sizes for Groups AB and AC

Group	Initial number of students assigned to the group	Number of students who did not complete one or both assessments	Final sample size
AB	202	37	165
AC	204	78	126

Scoring of open-response items

Upon completion of the tests, all data were returned by the teachers to the research team at Lakehead DSB in pre-labeled envelopes on November 28, 2012. On this same day, the teacher participants engaged in moderated marking training led by Dr. Bruce and then began scoring student test responses.

The scoring training session took the following form:

Part 1. Training Input Session

Goal: mark open and closed items on at least 33 tests in the morning and 33 tests in the afternoon

- Importance of 1000 item landmark to increase reliability across teachers and schools in terms of scoring and reporting

Scoring Practice Training for Tests A, B, C:

1. Look at items 9, 10 & 11 on Test B (try these tasks);
2. Look at rubric and anchor papers for item 9;
3. Look at one student test response (photocopied from a teacher – all markers mark the exact same student response);
4. Each teacher marks this response independently and *then* discusses with a partner
5. Whole group discussion
6. Repeat for item 10 and item 11

Discussion of possible student issues:

Item 9: A question about changing temperature. Students may struggle with the issue of “warmer than what?” - What is the reference point? Should the student pick zero? This is an additional step not required with the changing snow task which is the comparable EQAO item because there is a ‘ground’ level. And will students think that they should return to the point of origin or does each point become the next point of reference?

Item 10: There may be an interpretation where only Alice’s set of coins are considered, if students don’t notice “all in one” – they are still doing the probability but only for Alice not the entire range of coins. Each type of coin has a different value. This may operate as an additional distractor because the student must ignore this information. This distractor is not part of the EQAO parallel item and may make the coin version of item 10 slightly more challenging.

Item 11: The Pentagon task asks students to label angles, sides, etc but the anchor coded 40 does not have labels. The markers found this problematic. The team determined that since EQAO marked this at code 40, so would they. The team also developed a more explicit way of marking this item. The students must attend to four things to generate the complete diagram:

1. make a pentagon (closed figure with 5 sides)
2. make one side length as specified
3. make one right angle
4. make one obtuse angle with specified degrees

If the student did any one of these things accurately, they received a code 10. With two parts accurate they received a code 20, with 3 – code 30 and if all 4 things were accurate, the response was coded 40.
Teachers observed that it is highly language based. If students didn't read carefully, errors could be introduced.

Additional Clarifications:

- The participants discussed the difference between Codes and Levels (Code 30 is not necessarily the same as level 3 in classroom assignments).
- As per EQAO scoring, if the student had a correct answer but no work shown, the response received a code 20.

Part 2. Scoring session
Scoring of tests A, B, C (19 class sets x 2 tests) (approximately 66 tests per marker)
Folders were distributed with the scoring/recording sheets and teachers started in pairs to mark question 9. The markers selected to stay focused on one question at a time.

During this scoring period, the Trent research team and one of the consultants circulated through the room to trouble shoot, answer questions and concerns and to observe the way teachers were marking. If the research team or consultant noticed a teacher marking inaccurately, there was an informal consultation with that teacher and their marking partner. The teachers sat in pairs and checked in with one another regularly (approximately every four papers).
The whole marking team proceeded to mark items 10 and 11 in the same manner.

Scoring Reliability Methods
The marking reliability methods were employed over two days. On day one of marking, three people conducted reliability checks by circulating and discussing with pairs and bringing the pair to same scores (throughout the scoring process) – as described above. These three people also conferred with one another to check their own scoring skills and returned to the scoring samples and rubric repeatedly for verification.

On marking day two, three people were involved in a random re-scoring process. Two of these people were proven reliable scorers from day one of the marking. The second was one member of the research team. This reliability team randomly pulled every fifth paper for the re-scoring of 160 papers.
Procedure:

1. For each class set of papers, 1 in every 5 papers was redrawn for scoring.
2. The reliability team rescored these randomly selected papers in a blind fashion. That is, they did not know what the day one scores were. The day two scoring team recorded their scores on a separate score sheet each time.
3. Then the random selected paper scores of the reliability score team were compared to the scores of the first scoring team. The threshold of accuracy was very high: If there was a discrepancy of more than 1 level in any 3 items within the set of papers, the papers were deemed unreliable and the whole set of papers were re-scored. Similarly if more than one fourth of the items were

scored with a difference of one level, the whole class set of papers was also set aside for re-scoring.

4. In this process, it was determined that14 of the 16 sets of papers were consistently scored. This is a remarkable level of reliability (87.5% rate of agreement) – indicating that the scoring training was of high quality and that the scoring guides were helpful in making decisions about scores.
5. Two class sets did not meet reliability standards. In one set of papers, there were 3 scored items that had a discrepancy of more than 1 code (e.g., the day one score team marked a student response at code 10 but the day 2 score team marked the same student response at code 30). This entire set of papers was then pulled from the files, and every single paper in the set was re-scored by the reliability team with discussion amongst the team members whenever a response was difficult to score.
6. In the second set of papers that were scored unreliably, there were approximately 15 instances where a student response was scored with a difference of one level by the scoring team and the reliablity team. The entire set of papers was pulled from the files and re-scored by the reliability team.
7. For these two sets of papers that were scored unreliably by the first marking team, the scores from the reliability team were used as the final scores.

Data cleansing and preparation

Student assessment forms were collated upon completion of scoring and shipped to Trent for analysis. All scoring forms were scanned using Teleform© software version 10.4.1 and the resulting assessment data were imported into IBM© SPSS Statistics v.21. Multiple-choice responses were scanned according to a four-point scale corresponding to the response options for each question. Multiple choice data was then recoded such that 1 = correct answer and 0 = all else, where all else represented an incorrect answer or a blank response. Open-response item results were recorded according to the four level scale from the scoring rubric used to assess each open response question. Open response data were recoded as 1 = at or above provincial standard (level 3 or level 4) and 0 = all else, where all else represented a response scored at level 2 or below, or a blank response.

Questionable/unclear response choices were flagged by the software and manually verified (e.g., change of option choice) and the data exported to SPSS for statistical analysis. The data for each assessment was exported to data files Atest.sav, Btest.sav, and Ctest.sav. Each data file was reviewed for anomalies. The data for A-B sample and A-C sample was then merged on std_id. The SPSS data file Atest.sav contained data for all students who completed assessment A; therefore, data file Btest.sav and data file Ctest.sav were used as the key data files to add the student's assessment A data to the same student who completed assessment B or assessment C.

IV. DATA ANALYSIS

Item Matching

Assessment items on forms A, B and C were reviewed and matched according to curricular content and item type. Table 2 shows the curricular content from the three assessments, the item type where MC = Multiple-Choice and OR = Open-Response, and the corresponding item numbers from each of the three assessments. For example, the item that measured multiplication of units was a multiple-choice question numbered as question 1 on Assessment A, whereas that same content was measured by question 3 on Assessment B and by question 5 on Assessment C.

Analysis was conducted on matched pairs only. For example, Assessment A question 1 was compared to Assessment B question 3 because the same group of students completed both questions. Assessment A question 1 was compared with Assessment C question 5 for the group AC analysis.

Table 2.
Matched item series from Assessments A, B and C

Curricular content	Item type	Assessment A	Assessment B	Assessment C
Multiplication of units	MC	1	3	5
Building a geometric pattern	MC	2	7	2
Top view of a 3-D geometric figure	MC	3	1	4
Ratio	MC	4	6	8
Reading graphic information	MC	5	8	6
Translating data from a table to a graph	MC	6	2	7
Length of a line segment	MC	7	5	3
Calculating difference in measures	MC	8	4	1
Changes with units of measurement	OR	9	9	9
Probability	OR	10	10	10
Geometric angles and line segments	OR	11	11	11

In each of the two samples the same students completed the same items, therefore, a Paired Sample T-tests analysis was used to determine how each item on Assessment A correlated to the comparable items on B or C and the degree of difficulty of the items.

AB group analysis

The percentage of correct answers for each of the matched questions was compared for the AB group. Figure 2 shows the percentage of correct student responses for each of the matched items from Assessment A and Assessment B.

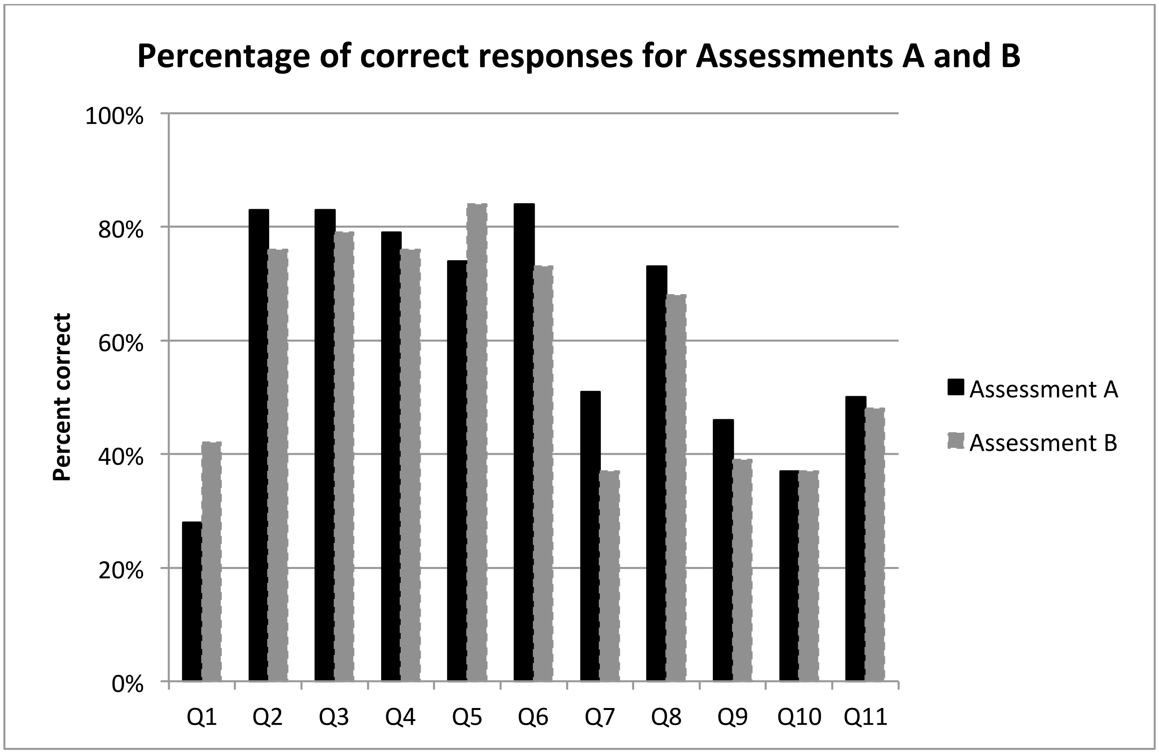


Figure 2: Percentage of correct responses for matched items in group AB

A quick visual interpretation of Figure 2 might lead some observers to the conclusion that there is only one matched item pair of the same difficulty. This is a conclusion based on the graph showing that Q10 has exactly equal results for Assessment A and Assessment B. While this is true in the purest interpretation of item agreement, it is more important to see that most of the matched pairs are within a few percentage points of each other. Overall, all the item pairs are well matched with the exception of Q1 and Q7. Descriptive statistics, such as comparing percentage results on two items, are certainly informative, but can be enhanced with additional analysis, as per below, in order to further investigate whether items are equal.

Reliability is the ability of an instrument to consistently measure constructs. The correlation between two forms of an assessment that are made of matched items is known as parallel-forms reliability. Cronbach's alpha, measured from 0 to 1, is the most common form of scale reliability. By convention a lenient cut off for exploratory research is 0.40 and ***all AB item pairs meet the criteria.*** Table 3 shows the scale reliability estimates for AB multiple choice item pairs.

Table 3.
Cronbach's alpha reliability estimates for AB multiple choice item pairs

AB item pairs	Estimates of scale reliability
A1 and B3	α=.40
A2 and B7	α=.56

A3 and B1	α=.80
A4 and B6	α=.59
A5 and B8	α=.45
A6 and B2	α=.46
A7 and B5	α=.45
A8 and B4	α=.68

A paired-samples *t* test is an inferential statistical technique used to compare the means or averages of groups to determine if there is a significant difference between them. The results of a *t* test report the mean for each item along with the standard deviation (SD) for that item. The *t* test also yields a *p*-value for statistical significance, a value which is often quoted but perhaps not well understood. Traditionally a *p*-value of less than 0.05 (reported as $p<.05$) is considered significant. However, in recent years, researchers have moved away from reporting *p*-values in terms of 'less than' or '<' to reporting the actual calculated value. This is because computer software technology allows for more exact calculations and has removed the necessity of consulting statistical tables for interpretative ranges.

A paired-samples *t* test was conducted on each item pair to evaluate whether students were more successful on Assessment A or on the matched item on Assessment B. T-test indicated that the mean score for four item pairs (A1-B3, A5-B8, A6-B2, A7-B5) were significantly different. These items are denoted as having statistically significant difference through the use of an asterisk (*) in the table. The results for each AB item pair are shown in Table 4.

Table 4.
T-test results for AB multiple choice item pairs

Variable Name	Mean	SD	t-test statistic
A1	.29	.45	t(152)=-2.78, *p*=.006*
B3	.42	.49	
A2	.82	.38	t(162)=1.93, *p*=.055
B7	.76	.43	
A3	.83	.38	t(161)=1.21, *p*=.226
B1	.80	.40	
A4	.80	.40	t(162)=1.22, *p*=.224
B6	.75	.43	
A5	.74	.44	t(162)=-2.57, *p*=.011*
B8	.83	.37	
A6	.84	.37	t(163)=3.04, *p*=.003*
B2	.73	.45	

A7	.51	.50	t(161)=2.27, *p*=.024*
B5	.38	.49	
A8	.75	.44	t(154)=1.63, *p*=.106
B4	.69	.46	

AC group analysis

The percentage of correct answers for matched questions was compared for the AC group. Figure 3 shows the percentage of correct student responses for each of the matched items from Assessment A and Assessment C.

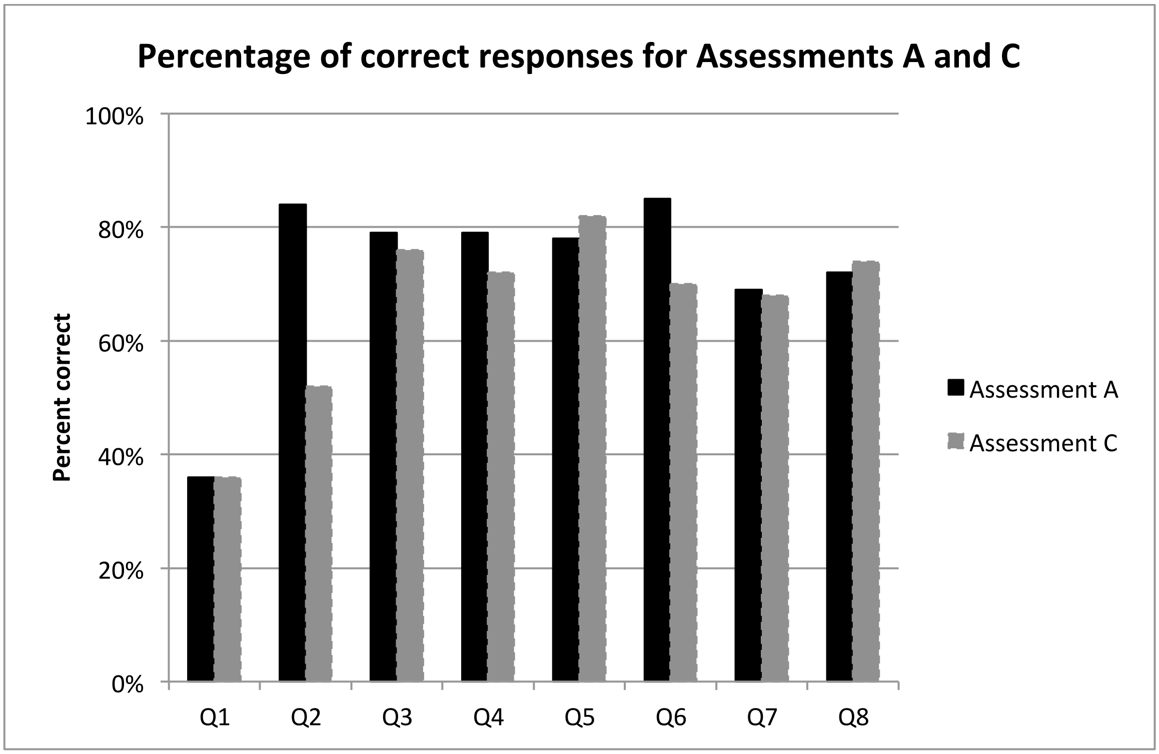


Figure 3: Percentage of correct responses for matched items in group AC

A quick visual interpretation of Figure 3 might lead some observers to the conclusion that there is one matched item pair of the same difficulty (Q1), however most of the matched pairs are within a few percentage points of each other with the exception of Q2 and Q6. Again, descriptive statistics, such as comparing percentage results on two items, were enhanced with additional analysis in order to further investigate whether items are equal.

To review, reliability is the ability of an instrument to consistently measure constructs. The correlation between two forms of an assessment that are made of matched items is known as parallel-forms reliability. Cronbach's alpha, measured from 0 to 1, is the most common form of scale reliability. By convention a lenient cut off for exploratory research is 0.40. Three AC item pairs, as noted with an asterisk in the table, do not meet the criteria and could be re-examined. Table 5 shows the scale reliability estimates for AC multiple choice item pairs.

Table 5.
Cronbach's alpha reliability estimates for AC multiple choice item pairs

AC item pairs	Estimates of scale reliability
A1 and C5	α=.71
A2 and C2	α=.42
A3 and C4	α=.83
A4 and C8	α=.29*
A5 and C6	α=.28*
A6 and C7	α=.20*
A7 and C3	α=.46
A8 and C1	α=.52

A paired-samples *t* test was conducted on each item pair to evaluate whether students were more successful on Assessment A or on the matched item on Assessment C. T-test indicated that the mean score for certain item pairs (A2-C2, A4-C8, A6-C7) were significantly different (as noted by the asterisk *). The results for each AC item pair are shown in Table 6.

Table 6.
T-test results for AC multiple choice item pairs

Variable Name	Mean	SD	t-test
A1	.34	.48	t(115)=-.41, *p*=.685
C5	.36	.48	
A2	.84	.36	t(121)=6.48, *p*=.001*
C2	.53	.50	
A3	.78	.41	t(124)=.83, *p*=.408
C4	.76	.42	
A4	.80	.40	t(117)=-2.40, *p*=.018*
C8	.90	.30	
A5	.78	.41	t(123)=-.87, *p*=.386
C6	.82	.38	
A6	.84	.36	t(124)=2.93, *p*=.004*
C7	.70	.45	
A7	.69	.46	t(122)=.16, *p*=.870

C3	.68	.47	
A8	.72	.45	t(122)=-.54, *p*=.592
C1	.74	.44	

V. DISCUSSION

This study duplicated the scoring procedures EQAO employs to ensure that assessment results are valid and reliable. All open-response items were scored by trained scorers. The multiple-choice items were captured by a scanner, operated by an experienced research technician. The analysis was completed by a researcher / teacher with a PhD in education and research methods who is familiar with EQAO items, the Ontario Curriculum, and the behaviour of item pairs on multiple choice and open response assessments.

The items developed in this project relate directly to items used on previous provincial assessments and measure how well students are achieving selected expectations from *The Ontario Curriculum*. The developed items include both performance-based, open-ended response items and multiple-choice questions through which students demonstrate what they know and can do in relation to the selected curriculum expectations.

The design allowed for the comparison of student results on matched assessment items. Results from the analysis of multiple-choice items show that most item pairs are of equal difficulty, however several item pairs could be further investigated.. These questionable item pairs were flagged during reliability analysis and during a comparison of means (*t* test) analysis. The items pairs that could be further examined are listed in Table 7, along with an indication of whether or not the potential misalignment was easily identifiable.

Table 7.
Item pairs that were significantly different in this study

Curricular content	Item pair	Potential misalignment identified?
Multiplication of units	A1 and B3	yes
Reading graphic information	A5 and B8	yes
Translating data from a table to a graph	A6 and B2	yes
Length of a line segment	A7 and B5	yes
Building a geometric pattern	A2 and C2	yes
Ratio	A4 and C8	yes
Reading graphic information	A5 and C6	yes

Translating data from a table to a graph	A6 and C7	Not clear - further field testing may be needed

The analysis phase of this study incorporated descriptive and inferential statistical techniques in order to determine statistical significance. Test of statistical significance are commonly used to provide a salient and definitive conclusion. As a result, Table 7 may contain some disappointing information. However, the data in this study was analyzed with a simultaneous review of the item pairs on all assessments. Each math question on all three assessments was physically completed by the research analyst in order to understand the behaviour of the item statistics in the context of the curricular content and item structure. As a result, possible reasons for the difference in item pair difficulty were hypothesized and are suggested in Table 8. In every case, the problem is very easily remedied.

Table 8.
Item pairs with suggestions for revisions

Item pair	Notes for possible item revision
A1 and B3 Multiplication of units	Step two of the problem should be re-examined for differences in number combinations resulting in the difficulty of dividing two digits into four digits
A5 and B8 Graphic information	A5 graphic pattern requires counting (1:2; 2:4; 3:6) B8 has an easier graphic pattern (3:3; 6:6, 9:9)
A6 and B2 Translating data	A6 legend is below graphic and inside graph frame B2 legend is beside graphic, not in graphic frame
A7 and B5 Length of a line segment	A7 response choices include measurements that are larger and smaller than the segment B5 line is shorter than all response options, making the response decisions different
A2 and C2 Build a geometric pattern	A2 pattern adds a complete shape C2 pattern adds two sides to an existing shape
A4 and C8 Ratio	Indeterminate
A5 and C6 Graphic information	See A5 and B8
A6 and C7 Translating data	A6 graphic has vertical bars C7 graphic has horizontal bars

This study has yielded statistically significant results, resulting in suggested revisions to the item pairs. The practical significance of this study should also be briefly discussed. The item development and testing of the assessment questions mirrored provincial protocol in every way possible. The content assessed directly relates to the Ontario Curriculum students follow in their classrooms. Student results from the assessment items can be examined by teachers in order to identify student strengths and areas of need, which in turn can be used to tailor instruction. Whether or not these items pairs were determined to be 'equal' through significance testing, there is no doubt they have practical significance. With relatively little revision, they should be even better matched.

VI. OPEN/ NON-MULTIPLE CHOICE ITEMS: Q 9,10 & 11

Data preparation

Student assessment forms were marked by moderated marking teams, and the resulting assessment data were entered into IBM© SPSS Statistics v.21. Data were entered according to a four-point scale corresponding to the response options for each question (e.g. Scores of 10, 20, 30 and 40). Responses that were illegible were coded "I" and blank responses were coded "B". Open response data was then recoded such that 1 = at or above the standard (scores of 30 & 40) and 0 = not attaining standard (Scores of 10, 20, B and I).

Assessment items on forms A, B and C were reviewed and matched according to curricular content and item type. Table 9 shows the curricular content for the three open response questions from each of the assessments.

Analysis was conducted on matched pairs only. For example, Assessment A question 9 was compared to Assessment B question 9 because the same group of students completed both questions. Assessment A question 9 was compared with Assessment C question 9 for the group AC analysis.

Table 9

Matched item series for OR (open response) questions for Assessments A, B and C

Curricular content	Item type	Assessment A	Assessment B	Assessment C
Changes with units of measurement	OR	9	9	9
Probability	OR	10	10	10
Geometric angles and line segments	OR	11	11	11

AB group analysis

The percentage of correct answers for each of the matched open response questions was compared for the AB group. Figure 4 shows the percentage of correct student responses for each of the matched open response items from Assessment A and Assessment B.

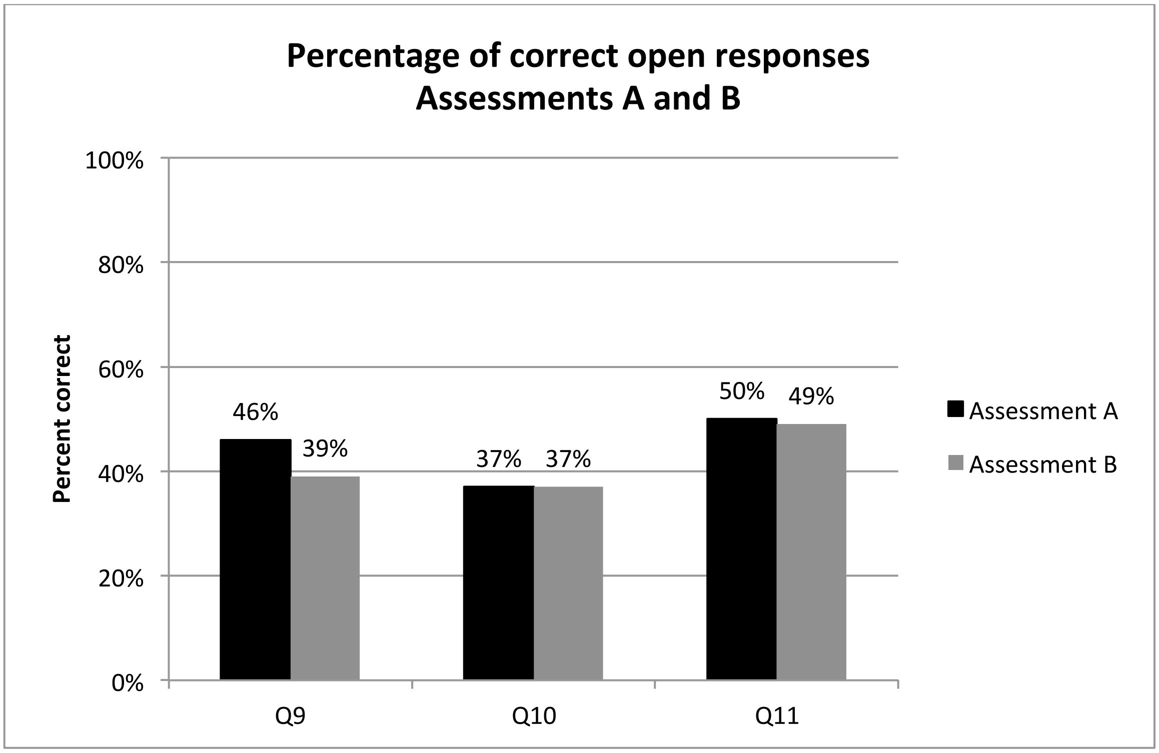


Figure 4: Percentage of correct open responses for matched items in group AB

A quick visual interpretation shows there is one exact matched item pair (Q10) for which students attained the same percentage of success (as measured by a Level 3 or 4, or 'attaining the standard') and one item pair that had a discrepancy of only 1%. However, further analysis, beyond a simple visual interpretation, is required to determine item equivalence.

Reliability is the ability of an instrument to consistently measure constructs. The correlation between two forms of an assessment that are made of matched items is known as parallel-forms reliability.

Cronbach's alpha, measured from 0 to 1, is the most common form of scale reliability. By convention a lenient cut off for exploratory research is 0.40 and all AB item pairs meet the criteria, *meaning the item pairs reliably and comparably measure the mathematical constructs*. Table 10 shows the scale reliability estimates for AB open response item pairs.

Table 10
Cronbach's alpha reliability estimates for AB open response item pairs

AB item pairs	Estimates of scale reliability
A9 and B9	α=.81
A10 and B10	α=.73
A11 and B11	α=.71

A paired-samples *t* test is an inferential statistical technique used to compare the means or averages of groups to determine if there is a significant difference between them. The results of a *t* test report the mean for each item along with the standard deviation (SD) for that item. The *t* test also yields a *p*-value to denote when there is a significant statistical difference between items. A *p*-value of less than 0.05 (reported as *p*<.05) is considered to be a statistically significant difference.

A paired-samples *t* test was conducted on each of the open response item pairs to evaluate whether students were more successful on Assessment A or on the matched item on Assessment B.

T-test results indicate that the mean scores for all item pairs were not significantly different, as students demonstrated a similar level of performance on each of the matched item pairs. The results for each AB item pair are shown in Table 11.

Table 11
T-test results for AB open response item pairs

Variable Name	Mean	SD	t-test statistic
A9	2.38	1.31	t(155)=.550, *p*=.583
B9	2.33	1.25	
A10	2.32	1.24	t(151)=.564, *p*=.573
B10	2.27	1.23	
A11	2.63	1.19	t(135)=-.605, *p*=.546
B11	2.68	1.21	

AC group analysis

The percentage of correct answers for each of the matched open response questions was compared for the AC group. Figure 5 shows the percentage of correct student responses for each of the matched items from Assessment A and Assessment C.

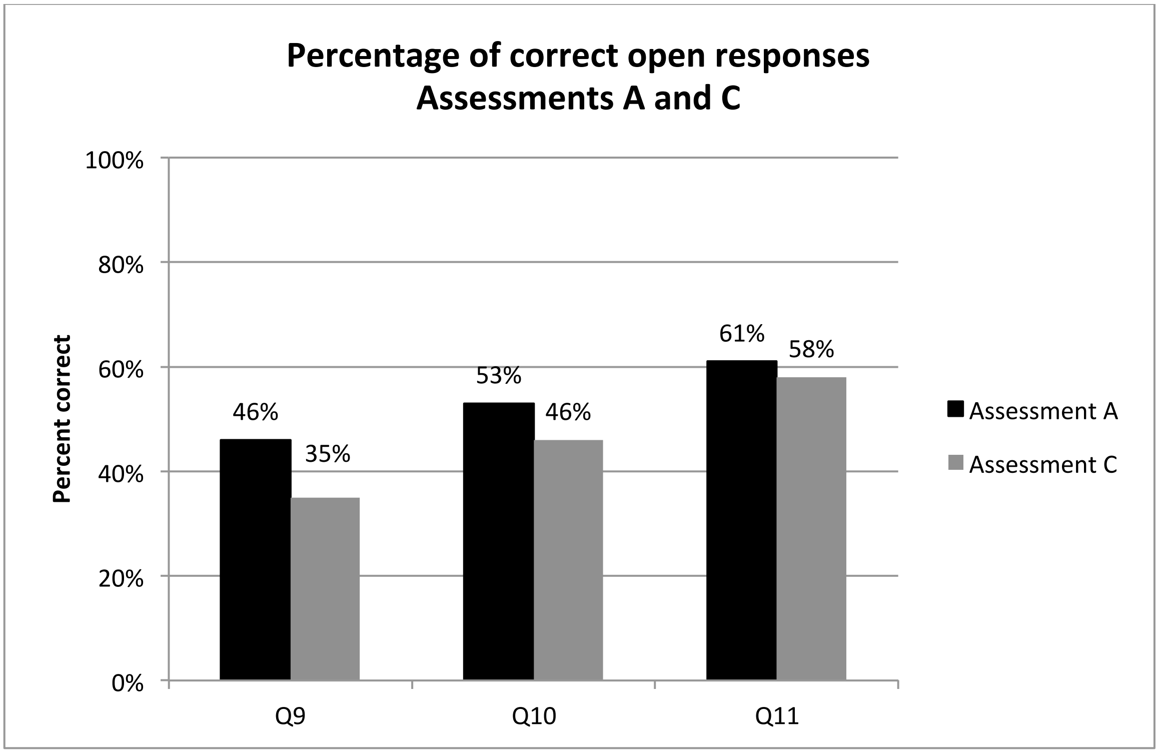


Figure 5: Percentage of correct open responses for matched items in group AC

A quick visual interpretation shows that there are not matched open response pairs for which students attained the same percentage of success. This is a conclusion based on the graph in Figure 3 showing that none of the pairs have an equal percentage of students attaining the standard (Level 3 or 4) for Assessment A and Assessment C. However, once again further analysis, beyond a simple visual interpretation, is required to determine item equivalence.

Reliability is the ability of an instrument to consistently measure constructs. The correlation between two forms of an assessment that are made of matched items is known as parallel-forms reliability.

Cronbach's alpha, measured from 0 to 1, is the most common form of scale reliability. By convention a lenient cut off for exploratory research is 0.40 and all AC item pairs meet the criteria. Table 12 shows the scale reliability estimates for AC open response item pairs.

Table 12

Cronbach's alpha reliability estimates for AC open response item pairs

AC item pairs	Estimates of scale reliability
A9 and C9	α=.64
A10 and C10	α=.83
A11 and C11	α=.77

A paired-samples *t* test is an inferential statistical technique used to compare the means or averages of groups to determine if there is a significant difference between them. The results of a *t* test report the mean for each item along with the standard deviation (SD) for that item. The *t* test also yields a *p*-value to denote when there is a significant statistical difference between items. A *p*-value of less than 0.05 (reported as *p*<.05) is considered to be a statistically significant difference.

A paired-samples *t* test was conducted on each of the open response item pairs to evaluate whether students were more successful on Assessment A or on the matched item on Assessment C.

T-test results indicate that t the mean scores for one item pair (A9-C9) were statistically significantly different.

The mean scores for two item pairs (A10-C10 and A11-C11) were not significantly different, as students demonstrated a similar level of performance on each of the matched item pairs. The results for each AC item pair are shown in Table 13.

Table 13

T-test results for AC open response item pairs

Variable Name	Mean	SD	t-test
A9	2.58	1.20	t(115)=2.77, *p*=.007
C9	2.28	1.08	
A10	2.67	1.23	t(110)=1.501, *p*=.136
C10	2.53	1.24	
A11	3.03	1.25	t(102)=1.643, *p*=.103
C11	2.85	1.26	

Discussion

This study duplicated the scoring procedures EQAO employs to ensure that assessment results are valid and reliable. All open-response items (Q 9, 10 & 11) were scored by trained scorers with stringent scoring reliability measures in place. The analysis was completed by a school board researcher / teacher who is familiar with

EQAO items, the Ontario Curriculum, and the behaviour of item pairs on open response assessments.

The items developed in this project relate directly to items used on previous provincial assessments and measure how well students are achieving selected expectations from *The Ontario Curriculum.* The developed items include performance-based, open-ended response items through which students demonstrate what they know and can do in relation to the selected curriculum expectations.

The design allowed for the comparison of student results on matched assessment items. Results from the analysis of open response items show that one item pair should be further investigated. The item pair was flagged by a comparison of means (*t* test) analysis. The pair is listed in Table 14, along with an indication of whether or not the potential misalignment was easily identifiable.

Table 14
Open response item pairs that were significantly different in this study

Curricular content	Item pair	Potential misalignment identified?
Measurement change	A9 and C9	yes

The analysis phase of this study incorporated descriptive and inferential statistical techniques in order to determine statistical significance. Tests of statistical significance are commonly used to provide a salient and definitive conclusion. Each math question on all three assessments was physically completed by the research analyst in order to understand the behaviour of the item statistics in the context of the curricular content and item structure. As a result, possible reasons for the difference in item pair difficulty were hypothesized and are suggested in Table 15.

Table 15
Item pairs with suggestions for revisions

Item pair	Notes for possible item revision
A9 and C9 Measurement change	A9 uses cm to measure cumulative change in snow depth over 10 days. This involves addition and subtraction of cm from a base level of 15 cm on the first day and could be visually observed over time using a vertically oriented metre stick or ruler. C9 uses degrees Celsius to measure change in classroom temperature between a certain time period each day. The base level temperature is not easily determined and the cumulative change is abstract and would not be observable on a thermometer.

This study has yielded statistically significant results and determined that five out of six open response item pairs are equivalent. The sixth pair was examined and a hypothesis presented as to why the Pearson developed item was not a good match. It would be valuable for developers of test items to examine what is being asked of students on this item to further consider where understanding broke down for students between items A9 and C9.

The practical significance of this study should also be briefly discussed. The item development and testing of the assessment questions mirrored provincial protocol in every way possible. The content assessed directly relates to the Ontario Curriculum students follow in their classrooms. Student results from the assessment items can be examined by teachers in order to identify student strengths and areas of need, which in turn can be used to tailor instruction. Whether or not these items pairs were determined to be ‘equal’ through significance testing, there is no doubt they have practical significance.

VII. RECOMMENDATIONS

This difficulty study has revealed that in fact, most test items generated by Pearson Canada were of the same level of difficulty as those on the 2011 EQAO math assessment, with some minor variances. Overall, this suggests a proof of concept that Pearson should continue with the current process of item generation in order to develop a larger bank of EQAO comparable items.
The caution is for Pearson and test designers to pay very close attention to details of:

a. geometric design similarities and accuracies;
b. location and orientation of graphics and information for each question;
c. level of cognitive load of the task when numbers are altered or the context is changed (including level of abstraction required to make sense of the question).

APPENDICES

1. Tests A, B, C in PDF format
2. Teacher report sample
3. Responses to Feedback from Pearson based on Draft 1

Appendix 1

Assessments A, B and C

DO NOT STAPLE FORMS

ASSESSMENT A

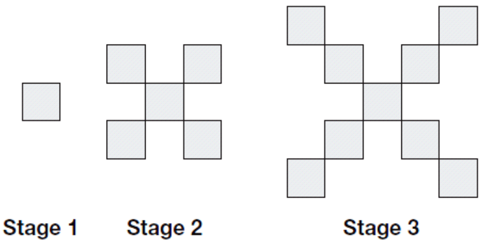
Name:________________________________

Mark an ☒ in the box next to your answer.

1. Every week, Danny eats 540 grams of cereal. Over 8 weeks, he finishes a total of 12 boxes of cereal. Each box contains the same amount of cereal. How many grams of cereal are in each box?

□360 □810 □4320 □ 6480

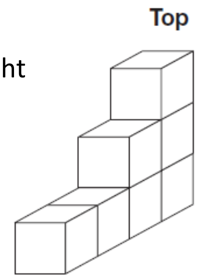
2. Manny uses tiles to build the geometric pattern shown.



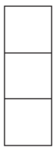
Which of the following represents the number of squares in Stages 4, 5 and 6 of Manny's pattern?

□17, 24, 31 □13, 17, 24 □13, 17, 21 □12, 16, 20

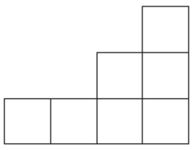
3. The three-dimensional figure at the right was built using cubes.



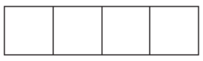
What is the top view of this figure?



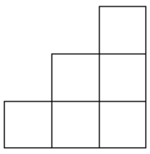
□ A



□ B



□ C



□ D

4. A recipe for a fruit drink uses 1 litre of cranberry juice, 2 litres of grape juice and 3 litres of orange juice. Which of the following could be represented by the ratio 3:2?

□ grape juice to orange juice

□ orange juice to grape juice

□ grape juice to cranberry juice

□ cranberry juice to grape juice

5. The graph below shows a relationship between the number of tasks Cole completes and the number of tokens he earns.

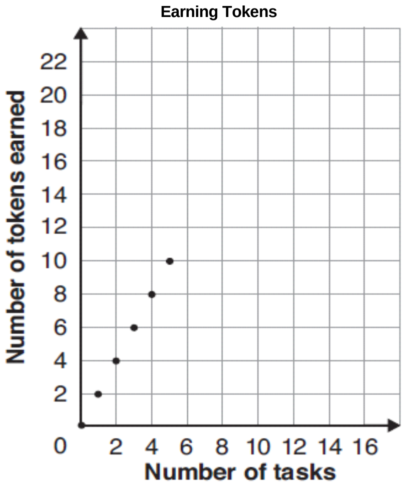
According to the pattern shown on the graph, how many tasks must Cole complete to earn 16 tokens?

□ 6

□ 8

□ 16

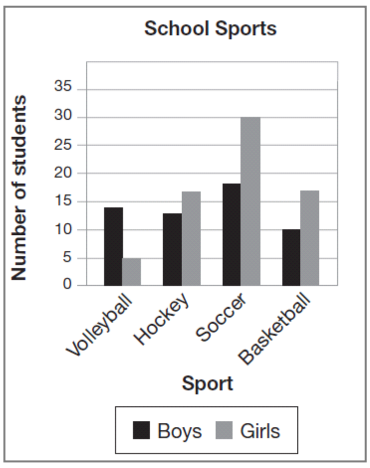
□ 32



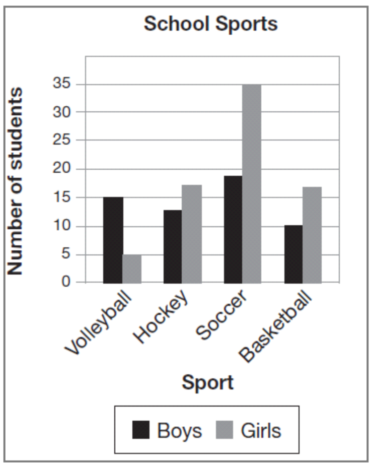
6. The table below shows data about participating in school sports.

Sport	Number of boys	Number of girls
Volleyball	14	5
Hockey	13	17
Soccer	18	35
Basketball	13	17

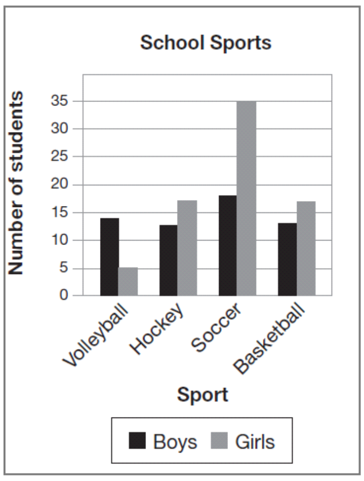
Which graph represents this data?



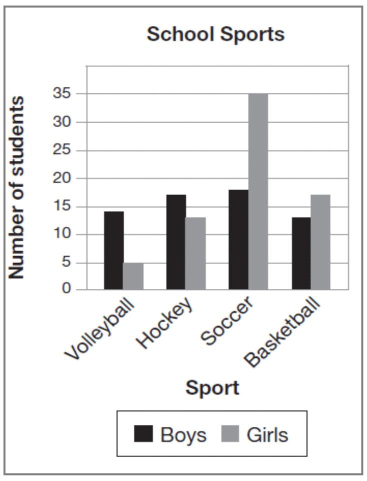
□ A



□ B

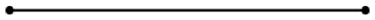


□ C



□ D

7. Consider the line segment below.



Which of the following is closest to its length?

□3.7 cm □4.2 cm □47 mm □ 57 mm

8. The amounts of water in two containers are shown in the table below.

Container	Amount of water (L)
A	0.967
B	1.02

What is the difference between the amounts of water in the containers?

□ 0.053 L □0.865 L □1.947 L □1.987 L

9. The table below shows the changes in the amount of snow on the ground over 10 days.

Day	Change
1	15 cm new
2	7.5 cm new
3	no change
4	4.5 cm melted
5	3.5 cm melted
6	4 cm melted
7	no change
8	12 cm new
9	2.5 cm new
10	8 cm new

Ali estimates that the total change is an increase of 30 cm.
Nadia estimates that the total change is an increase of 25 cm.

Which student makes a more accurate estimate? □ Ali □ Nadia

Justify your answer.

10. Dakota and Bryan count their coloured paper clips and record the results in the table below.

Colour	Dakota	Bryan
Red	14	18
Yellow	7	9
Blue	6	5
White	17	20

They put all of the paper clips in a box.
Dakota chooses one paper clip from the box without looking.

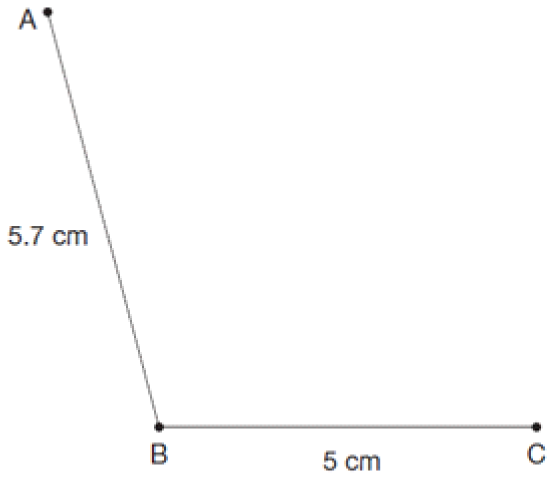
Determine the probability that Dakota chooses a red paper clip.

Show your work.

11. Use the line segments AB and BC below to construct pentagon ABCDE with the following properties:

- a right angle at point C
- an angle that measures 110° at point A
- a side of 4.7 cm

Label all angles and sides with their measures.



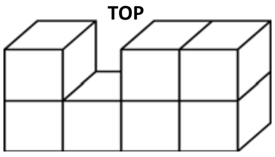
ASSESSMENT B

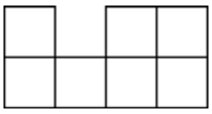
Name:____________________________________

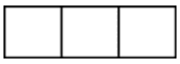
Mark an ☒ in the box next to your answer.

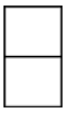
1. Here is an object made with cubes.

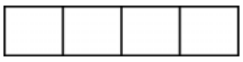
Which diagram represents the top view of this object?




□ A


□ B


□ C

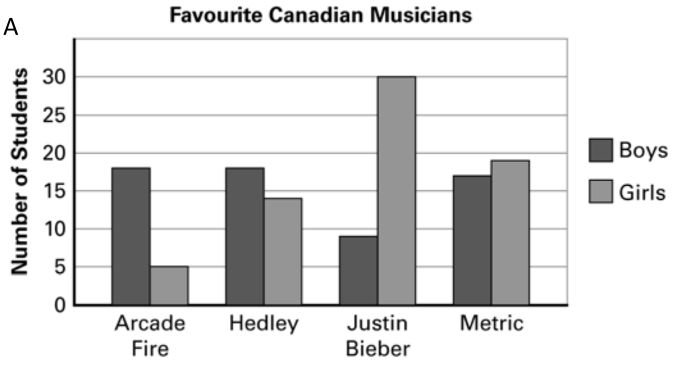

□ D

2. Grade 6 students in a school were surveyed about their favourite Canadian musician. The results are shown in this table.

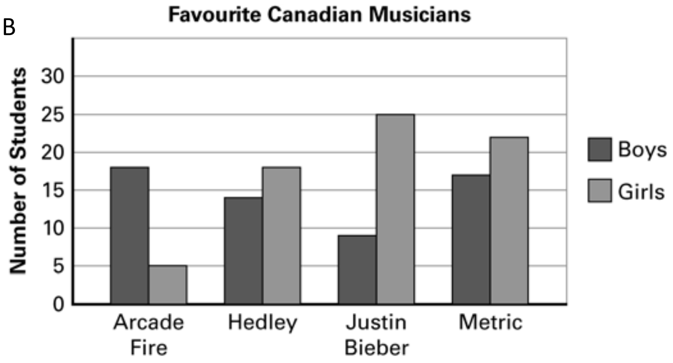
Musician	Number of boys	Number of girls
Arcade Fire	18	5
Hedley	14	18
Justin Bieber	9	30
Metric	17	19

Which graph represents these data?

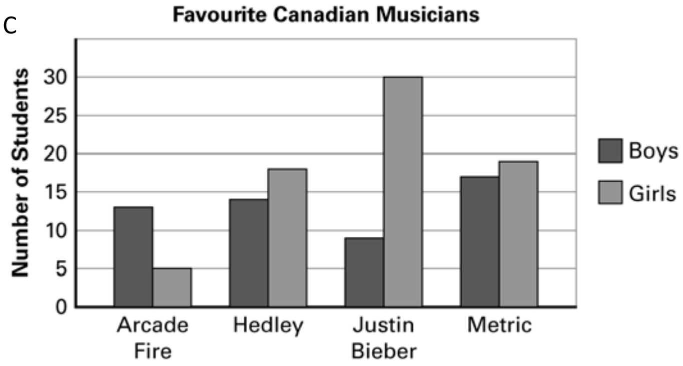
□ A



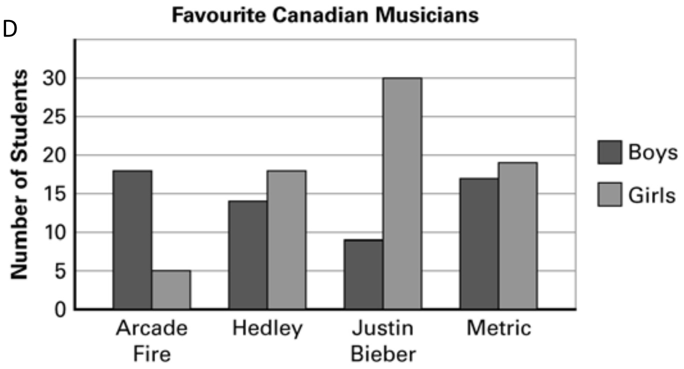
□ B



□ C



□ D



3. Hamida drinks 480 mL of juice each week. In 15 weeks, she drinks 12 cartons of juice. All the cartons have the same amount of juice. How many millilitres of juice are in each carton?

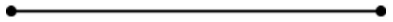
□ 384 □ 600 □ 5760 □ 7200

4. The amounts of juice in two jugs X and Y are shown in the table. What is the difference between the two amounts?

Jug	Amount of juice (L)
X	0.895
Y	1.04

□ 1.935 L □ 0.791 L □ 1.855 L □ 0.145 L

5. Which of the measures below is closest to the length of this line segment?



□ 4.6 cm □ 61 mm □ 51 mm □ 5.6 cm

6. Claire mixed 2 litres of pineapple juice, 5 litres of orange juice, and 7 litres of water to make a fruit punch for the class party. What could the ratio 7:2 represent?

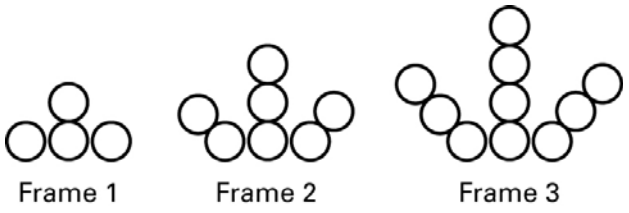
□ pineapple juice to water

□ water to orange juice

□ orange juice to pineapple juice

□ water to pineapple juice

7. Jake used counters to create this pattern.



The pattern continued. How many counters were there in Frames 4, 5, and 6 of Jake's pattern?

□ 12, 15, 18 □ 13, 16, 22 □ 13, 16, 19 □ 15, 18, 21

8. This graph shows the relationship between the number of books Alex bought at a yard sale and the cost of them.

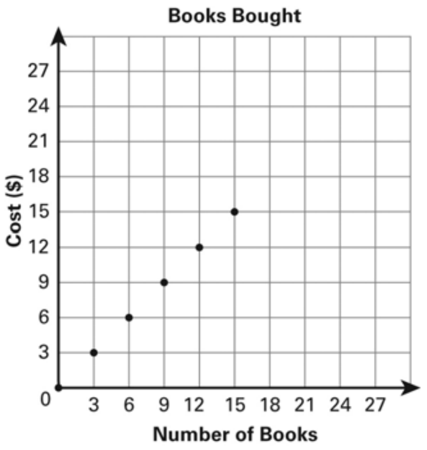
Assume the pattern on the graph continues. How many books could Alex buy for $24?

□ 24

□ 27

□ 30

□ 21



9. This table shows the changes in the temperature of the water in a swimming pool over 10 weeks.

Week	Change
1	8°C warmer
2	9.5°C warmer
3	3.5°C colder
4	no change
5	2.5°C colder
6	6°C colder
7	11°C warmer
8	no change
9	5.5°C warmer
10	7°C warmer

Trudy estimated that the total change is an increase of 25°C.
Len estimated that the total change is an increase of 30°C.

Which student made the better estimate? □ Trudy □ Len

Justify your choice.

10. Alice and Remy counted the coins in their charity boxes. They recorded their results in this table.

Coin	Alice	Remy
Penny	6	2
Nickel	14	6
Dime	13	19
Quarter	7	13

Alice and Remy put all the coins into one box.
Without looking, Alice picked one coin from the box.

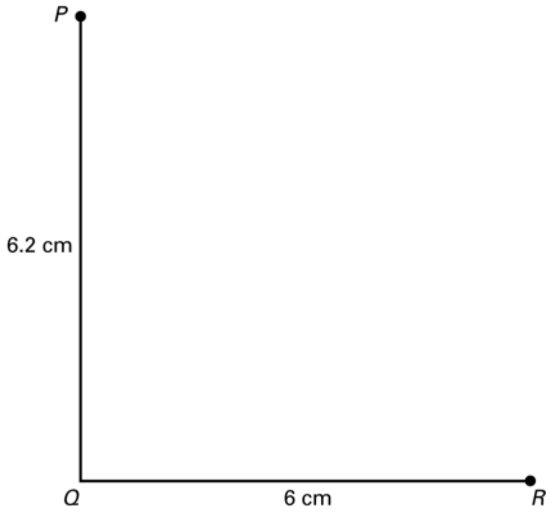
Determine the probability that Alice picked a dime.

Show your work.

11. Sides PQ and QR of pentagon PQRST are shown.
Construct pentagon PQRST with these properties:

- an angle of 120° at point P
- a side of length 3.6 cm
- a right angle at point R

Label all angles and sides with their measures.



DO NOT STAPLE FORMS

ASSESSMENT C

Name:___________________________________

Mark an ☒ in the box next to your answer.

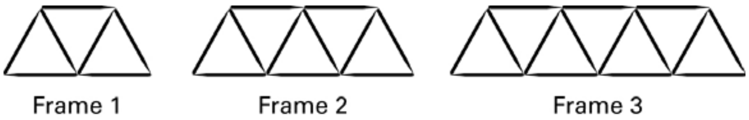
1. The amounts of soup in two bowls P and Q are shown in the table.

Bowl	Amount of soup (L)
P	0.876
Q	2.03

What is the difference between the two amounts?

□ 2.846 L □ 0.673 L □ 1.154 L □ 2.906 L

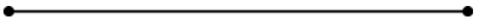
2. Mario used toothpicks to create this pattern.



The pattern continued. How many toothpicks were there in Frames 4, 5, and 6 of Mario's pattern?

□ 17, 19, 21 □ 9, 11, 13 □ 19, 22, 26 □ 19, 23, 27

3. Consider this line segment.

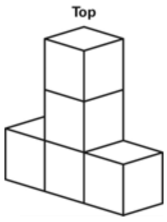


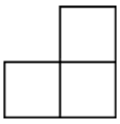
Which of the measures below is closest to the length of this line segment?

□ 5.1 cm □ 5.6 cm □ 61 mm □ 66 mm

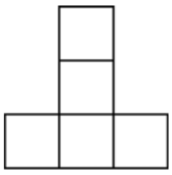
4. Here is an object made with cubes.

Which diagram represents the top view of this object?

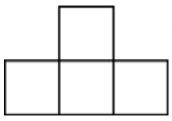




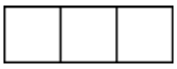
□ A



□ B



□ C



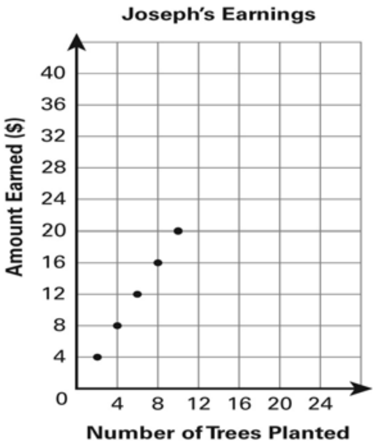
□ D

5. Jason eats 360 grams of waffles each week. In 9 weeks, he eats 12 boxes of waffles. The boxes of waffles have the same amount. How many grams of waffles are in each box?

□ 270 □ 480 □ 3240 □ 4320

6. This graph shows the relationship between the number of trees Joseph plants and the amount he earns.

Assume the pattern shown on the graph continues. How many trees must Joseph plant to earn $32?



□ 32

□ 12

□ 14

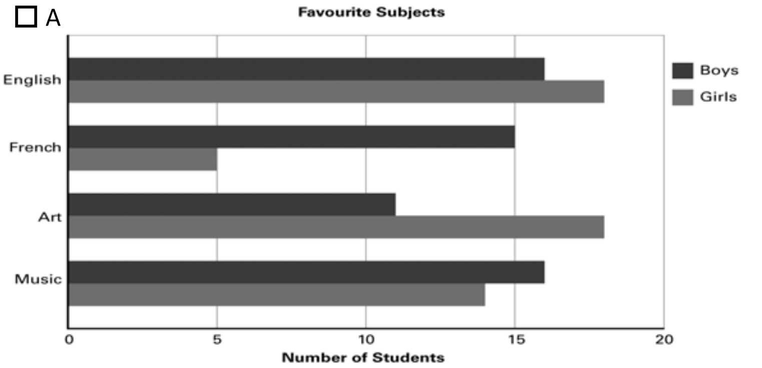
□ 16

7. Grade 6 students in a school were surveyed about their favourite subject. The results are shown in this table.

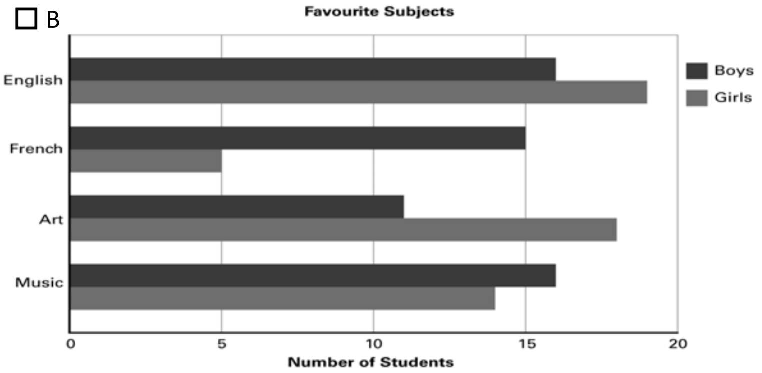
Subject	Number of boys	Number of girls
English	16	18
French	15	5
Art	11	18
Music	16	14

Which graph represents these data?

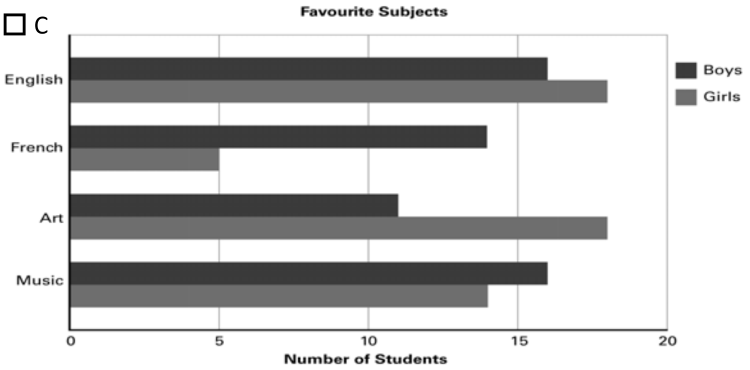
□ A



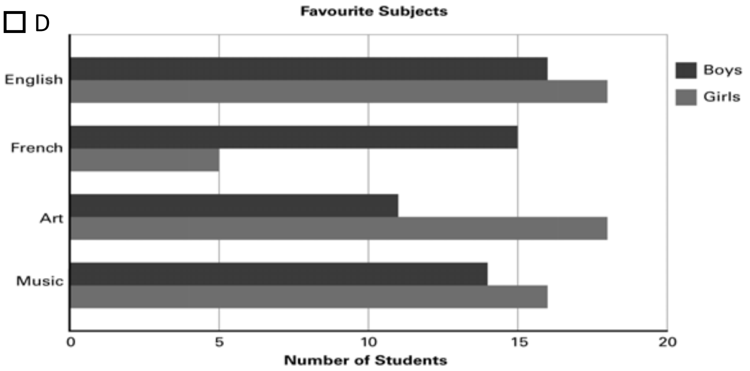
□ B



□ C



□ D



8. Emily mixed 2 litres of lemonade, 3 litres of grape juice, and 4 litres of orange juice to make a fruit punch. What could the ratio 4:3 represent?

□ lemonade to grape juice

□ orange juice to grape juice

□ orange juice to lemonade

□ grape juice to lemonade

Draft

9. Students in a Grade 6 class measured the temperature in their classroom at 9 a.m. and 3 p.m. each day for 10 days. This table shows how the temperature changed each day.

Week	Change
1	9˚C warmer
2	3.5˚C colder
3	no change
4	6˚C colder
5	5.5˚C warmer
6	9˚C warmer
7	2.5˚C colder
8	7˚C warmer
9	no change
10	11˚C warmer

Lisa estimated that the total change is an increase of 30°C.
Graeme estimated that the total change is an increase of 25°C.

Which student made the better estimate? □ Lisa □ Graeme

Justify your choice.

10. Tess and Jon counted their coloured counters.
They recorded their results in this table.

Counter	Tess	Jon
Blue	6	4
Black	17	19
Green	5	8
Red	12	13

Tess and Jon placed all the counters in one bag.
Without looking, Jon picked one counter from the bag.

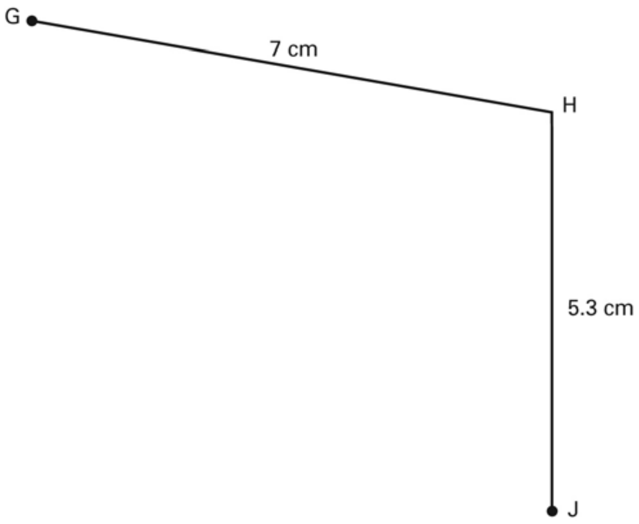
Determine the probability that Jon picked a black counter.

Show your work.

11. Sides GH and HJ of pentagon GHJKM are shown.
Construct pentagon GHJKM with these properties:

- a right angle at point J
- a side of length 3.3 cm
- an angle of 95°

Label all angles and sides with their measures.



Appendix 2

Teacher Report Sample

Field Test of Grade 6 Mathematics Success Resource
Assessment A-B Sample
Class Summary

Students in grade 7 at 19 schools in one school district were selected to complete two Grade 6 math assessment forms which consisted of a mix of multiple choice and open-response questions. The first form Assessment A was made up of questions asked on a previous EQAO test given in the Spring of 2011, the second assessment form B and C were questions drawn from an independent question bank that were considered to be of equal difficulty to similar questions on form Assessment A. The total sample assigned to the study were divided into two groups each completing Assessments A and B or Assessments A and C.

Below are the summaries of results for students in your class that participated in the Assessment A-B administration and completed both assessments. Students who completed only one of the assessments were removed from the study.

Table 1 provides a summary of the mean value[1] for each item topic for your class (N=22), the overall mean for the Assessment A-B sample (N=165), and the overall mean for students in the Assessment A-C sample (N=126) on a comparable item topic. The mean value is the achievement score or the proportion of students who correctly answered the item. It can range from 0.0 to 1.0. It is calculated as the sum of all scores divided by the number of students. If all students answer the question correctly, the achievement score is 100% or 1.00. The mean value represents the difficulty index or mastery index of the item topic. The criterion for evaluating the achievement score as a difficulty index are a mean value of 0.62 the item is of optimum difficulty level between high achievers and low achievers, greater than 0.90 the item may be too easy, and less than 0.20 the item may be too difficult or is measuring something other than what the item was intended to measure)[2]. The criterion as a mastery index the mean value should be 0.90 or higher (e.g., 90% of the students mastered the item topic).

Table 1.
Item Topic Mean Score for Assessment A, B, and C

	Assessment A			Assessment B			Assessment C	
		Mean			Mean			Mean
Item Topic	Q#	**Class**	All	Q#	**Class**	All	Q#	All
calculating ml or grams	1	**.10**	.27	3	**.40**	.41	5	.34
growing patterns	2	**1.00**	.83	7	**.80**	.75	2	.52
representing figure views	3	**.60**	.81	1	**.70**	.79	4	.75
representing ratios	4	**.80**	.79	6	**.90**	.75	8	.86
predicting using continuous line graphs	5	**.60**	.73	8	**.70**	.83	6	.81
selecting graphical representations of data	6	**.90**	.84	2	**.70**	.73	7	.70
measuring line segment	7	**.60**	.50	5	**.50**	.37	3	.67
calculating decimal amounts (litres)	8	**.70**	.72	4	**.70**	.65	1	.73
estimating cm or temp using table data	9a	**.80**	.73	9a	**.80**	.66	9a	.64
explaining how differences were calculated	9b	**.41**	.46	9b	**.45**	.39	9b	.35
determining probability and explaining	10	**.55**	.38	10	**.59**	.37	10	.46
constructing and labeling a pentagon	11	**.55**	.50	11	**.41**	.48	11	.58

[1] Using the answer key, the multiple choice item data were recoded into new variables such that 1=correct and 0=all else, open-ended items were recoded as correct and incorrect such that 1=3 or greater and 0=all else.

[2] On a four-alternative, multiple-choice item, the random guessing level is 1.00/4=.25; therefore, the optimal difficulty level is .25+(1.00-.25)/2=.62; on a true-false question, the guessing level is (1.00/2=.50) and, therefore, the optimal difficulty level is .50+(1.00-.50)/2=.75

Tables 2 and 3 provide the individual student scores for your class. The multiple choice questions were scored such that 1=correct and 0=incorrect, and the open response questions were scored using a rubric scoring guide such that 10=lack of understanding of concept and 40=thorough understanding of concept. A mean score was calculated to show the average score achieved for the three open response items.

Table 2.
Assessment A - Individual Student Scores (N=22)

	Multiple Choice Questions										Open-response Questions				
Student	calculating ml or grams	growing patterns	representing figure views	representing ratios	predicting using continuous line graphs	selecting graphical representations of data	measuring line segment	calculating decimal amounts (litres)	estimating cm or temp using table data	**Score /8**	explaining how differences are calculated	determining probability and explaining	constructing and labeling a pentagon	**Score /120**	**Mean Score**
Alaska	0	1	0	1	1	1	1	1	1	**7**	20	30	10	**60**	20.00
Alyssa	0	1	0	1	0	1	1	1	1	**6**	20	10	20	**50**	16.67
Amanda	0	1	1	1	1	0	1	1	0	**6**	10	40	20	**70**	23.33
Brian	0	1	0	0	1	0	0	0	1	**3**	10	10	10	**30**	10.00
Cameron	1	1	0	1	1	1	1	0	0	**6**	10	10	30	**50**	16.67
Charlie	0	1	0	1	1	0	1	0	1	**5**	30	40	30	**100**	33.33
Cuyler	0	1	1	0	1	1	0	1	1	**6**	40	30	20	**90**	30.00
Drewe	0	1	1	1	1	1	1	1	1	**8**	40	20	40	**100**	33.33
Dylan	0	1	0	0	1	1	0	1	0	**4**	0	0	0	**0**	0.00
Emilie-Jade	0	1	0	0	1	1	1	1	1	**6**	10	40	10	**60**	20.00
Jenna	1	1	1	1	1	1	0	1	1	**8**	30	40	30	**100**	33.33
Joshua	0	0	1	1	1	1	1	1	1	**7**	10	40	30	**80**	26.67
Kody	1	1	1	1	1	1	0	1	1	**8**	40	10	0	**50**	16.67
Makenzie	0	1	1	1	1	1	1	0	1	**7**	10	10	40	**60**	20.00
Matthew	0	1	1	0	1	0	0	1	0	**4**	0	0	10	**10**	3.33
Natalie	1	1	1	1	1	1	1	0	1	**8**	40	40	30	**110**	36.67
Rayna	0	1	1	1	0	1	1	1	1	**7**	40	40	40	**120**	40.00
Riley	0	1	1	1	1	1	0	0	0	**5**	30	40	20	**90**	30.00
Sydney	0	1	1	1	0	1	0	1	1	**6**	20	40	40	**100**	33.33
Tesha	1	1	0	1	0	1	0	0	1	**5**	10	10	30	**50**	16.67
Thomas	0	0	1	1	1	1	0	1	1	**6**	40	40	30	**110**	36.67
Walker	0	1	1	0	0	1	0	1	1	**5**	10	10	30	**50**	16.67

Table 3.
Assessment B - Individual Student Scores (N=22)

	Multiple Choice Questions										Open-response Questions				
Student	calculating ml or grams	growing patterns	representing figure views	representing ratios	predicting using continuous line graphs	selecting graphical representations of data	measuring line segment	calculating decimal amounts (litres)	estimating cm or temp using table data	**Score /8**	explaining how differences are calculated	determining probability and explaining	constructing and labeling a pentagon	**Score /120**	**Mean Score**
Alaska	0	1	1	1	1	1	1	1	1	**8**	40	40	20	**100**	33.33
Alyssa	0	1	0	1	1	0	0	0	1	**4**	20	10	20	**50**	16.67
Amanda	0	1	1	1	1	0	0	0	0	**4**	10	30	30	**70**	23.33
Brian	0	0	0	0	0	0	0	1	1	**2**	20	30	10	**60**	20.00
Cameron	0	1	0	1	0	1	1	1	0	**5**	10	10	20	**40**	13.33
Charlie	0	1	0	1	1	1	1	1	0	**6**	30	40	30	**100**	33.33
Cuyler	0	1	1	1	1	1	0	1	1	**7**	40	30	20	**90**	30.00
Drewe	1	1	1	1	1	1	0	1	1	**8**	40	20	40	**100**	33.33
Dylan	0	1	0	1	0	1	0	1	0	**4**	10	20	10	**40**	13.33
Emilie-Jade	0	1	0	1	1	0	0	0	0	**3**	10	30	20	**60**	20.00
Jenna	1	1	1	1	1	0	0	1	1	**7**	40	40	30	**110**	36.67
Joshua	1	1	1	1	1	0	0	0	0	**5**	10	30	40	**80**	26.67
Kody	1	1	1	1	1	1	1	0	0	**7**	30	10	0	**40**	13.33
Makenzie	0	1	1	1	1	1	1	1	0	**7**	10	10	40	**60**	20.00
Matthew	0	1	1	0	1	1	0	0	0	**4**	0	10	0	**10**	3.33
Natalie	1	1	1	1	1	1	1	1	1	**9**	40	40	40	**120**	40.00
Rayna	1	1	1	1	1	1	0	1	1	**8**	40	40	30	**110**	36.67
Riley	1	1	1	1	1	1	0	0	1	**7**	40	40	20	**100**	33.33
Sydney	0	1	1	1	1	1	1	1	1	**8**	20	40	30	**90**	30.00
Tesha	0	1	0	0	1	1	1	0	1	**5**	10	10	20	**40**	13.33
Thomas	0	1	1	1	1	0	1	1	1	**7**	40	40	20	**100**	33.33
Walker	0	0	1	1	1	1	1	1	0	**6**	0	10	0	**10**	3.33

Appendix 3

Responses to Feedback from Pearson

based on Draft 1

Response to Pearson feedback – February 4, 2013

Question 1
Is there inconsistency in the results shown in Figure 3, Table 5 and Table 6? The example given in the Pearson feedback shows a charted comparison of items A2 and C2.

Response:
All figures and tables were verified to be accurate. Figure 3 and Table 6 are related. The bar graph in Figure 3 is essentially a visual representation of the 'Mean' column from Table 6. For example, Table 6 shows that the mean of item A2 was .84 which equates directly to the bar graph for A2 (Assessment A, dark colour). Likewise, the mean of item C2 in Table 6 is .53, and is represented in the graph by the C2 bar (Assessment C, lighter colour). The comparisons of means are likely the most important way for Pearson to make decisions based on the sample size.

In the visual representation given in Figure 3, there is a large difference in the results for this item pairing. The t-test was completed to consider whether this difference in mean scores for the two items was numerically significant and couldn't simply be attributed to chance. The t-test compares the means of two groups and helps us to answer the question, "Is this difference big enough to be worried about?" In this case, the t-test confirmed the visual interpretation that there was a large difference in the percentage of correct responses and that it is worth further investigation.

Reliability is the fact that a scale should consistently reflect the construct it is measuring. Another way to think of reliability is that a person should get the same score on a test if they complete it a two different points in time (if no learning has occurred in between).

There is discussion in the research literature about what constitutes an "acceptable" alpha level. You'll often see in books or journal articles that a value or 0.7 – 0.8 is acceptable and that a value below 0.5 is unacceptable, but these books and papers refer to standardized or large scale norm-referenced tests such as intelligence tests that have many items measuring the same construct. In the case of this study, only two items were compared and reported on for each alpha level, which gives a greater range of flexibility in the scaling. The alpha value of .42 for items A2 and C2 is on the lowest end of what is considered acceptable and therefore it makes sense to take this information into consideration in the context of other information provided in the report.

Question 2
In the comparison of the % results in Figures 2 and 3, is it safe to assume that Q1 represents A1=B3 and A1=C5?

Response:
Yes, that is correct.

Question 3
What does t(115)=-.41 mean?

Response:
This is a standard method of reporting t-test results. The t is for t-test and the number in brackets represents the degrees of freedom of the test. In this case, the number is 115,

which represents the number of values in the calculation. You can see that for each item pairing, there may be a slightly different number in the brackets, based on the number of students who answered the particular question. The number after the equals sign represents the computed t-test value.

The formula to compute a t-test requires the computation of the difference between the means of two questions or groups. This is the numerator. The bottom part of the equation is the standard error of the difference, which is the variance for each question or group divided by the number of people (or responses).

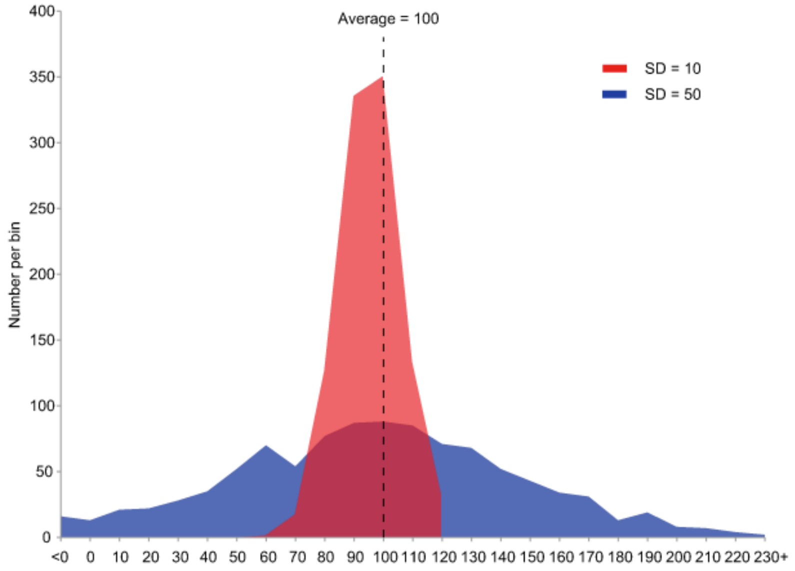
The t test value will be positive if the first mean is larger than the second and negative if the first mean is smaller than the second.

T-tests don’t have an acceptable value, rather they are classical hypothesis tests. In our example, we hypothesized that the items Pearson created were equal in difficulty to the EQAO released test item because they were written to match the released items. The t-test calculates the probability that the hypothesis is true and the premise behind a p value indicates whether or not the observed results would be likely under the given hypothesis.

Standard deviation is a measure of dispersion and is often reported along with a mean (or average). Another way to think about standard deviation is how spread out values are in a data set.

For more information on the definitions of t-tests, p values and standard deviation, a good source is to simply Google these terms.
For example this visual of the meaning of standard deviation is available on Wikipedia: http://en.wikipedia.org/wiki/Standard_deviation

A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.



Question 4
What happens to these scores when students don't complete a question?

Answer
Blank data is not included in the calculation.

Question 5
We're noticing that in table 6 the higher the mean, the higher the t-value.

Response:
See explanation in Question 1 response.

Question 6
How do you reconcile the differences between the p value and the alpha value?

Response:
See explanation in Question 1.
The p value from the t-test tells us whether the students did equally well on parallel items. A p value of less than 0.5 is weak and tells us that the items are significantly different in difficulty. It is a comparison of mean responses for each question pair.

The alpha value is a measure of how consistently these two questions measure the same construct (e.g., topic such as extending a growing pattern).

The comparison of group means, or t-tests, is more suited to this study because the study did not include multiple questions measuring the same construct.

Question 7
Open response question are missing. Is this an oversight?

Response:
Please see final report.

Question 8
a. Does the fact that the order of matching questions was changed from Assessment A to Assessment B or C have a bearing on the end results?

Response:
There is a body of research investigating item positioning in large-scale assessment, such as national and international work. Depending on the study and the year, there may be some observations about item fatigue whereby students do worse on items placed near the end of the test, but these tests tend to be longer assessments (e.g. > 1 hour in length) and therefore not comparable to eight multiple choice questions.

b. In general, do students tend to perform more poorly on the later multiple-choice questions?

Response:
Not necessarily. Reports of this phenomenon have to be interpreted with cautious consideration for the length of the assessment and the assessment conditions. For example, this effect is magnified if the assessment is a timed assessment and students who don't finish the entire assessment leave blank answers near the end of the test. The Pearson difficulty study tests were not timed.

c. Why were the questions in different orders?

Response:
The items were rotated slightly, mostly for fit on the page to keep the assessment structure to two pages.

Question 9
a. Why are there such large number of students who did not complete the assessments?

Response:
This is always the case when administering two tests in real classroom contexts. Student absence, teacher absence, school schedule interruptions, and lack of attention to detail are all possible reasons why both tests were not completed.

b. If a student did not complete one or both tests, were responses included?
Response:
We only used data for students who had completed both assessments. If students did not complete at least one question on both tests, their data was not included. There may have been some unanswered questions for some of these students but the majority of questions on both forms were completed for each student.