DRAFT: What lies behind the dots and checkmarks?

Student performance on the assessment is reported as a Performance Level (from 1 to 4) in Reading, Writing, and Mathematics.

We also include a visual representation of student performance in Mathematics and Reading in the form of a dot indicating the student's placement in their performance level. **New:** Writing will also be represented by a dot in the 2022-2023 school year.

For example, a dot to the left of level 3 means the student performed closer to the bottom of level 3, near level 2.

Reading Performance	Not enough evidence too few questions attempted	Level 1 below the expectation	Level 2 approaching the expectation	Level 3 at the expectation	Level 4 above the expectation
					

A dot to the right of level 3 means the student performed closer to the top of level 3, near level 4.

Reading Performance	Not enough evidence too few questions attempted	Level 1 below the expectation	Level 2 approaching the expectation	Level 3 at the expectation	Level 4 above the expectation
					

This document provides a brief description of the assessment scoring process, then describes the generation of scores for Reading and Math and finally Writing.

Scoring the assessments and examinations

After the assessments or examinations are completed, schools box them up and ship them back to the Department of Education and Early Childhood Development. Here at the Department we count and sort thousands of booklets. For virtual scoring sessions, we then cut and scan the booklets (described later). For in-person sessions, we shuffle and count the booklets before sending them to a central scoring session.

At the scoring sessions, teachers score constructed response items (two writing tasks for literacy assessments, several items for Mathematics 10). After comprehensive training on the scoring rubrics and exemplars, teachers assign scores. At least two teachers score each writing task. If the scores are not adjacent (e.g. if a booklet receives a 2 and a 4 in Ideas for task 1), a third teacher provides a judgement.

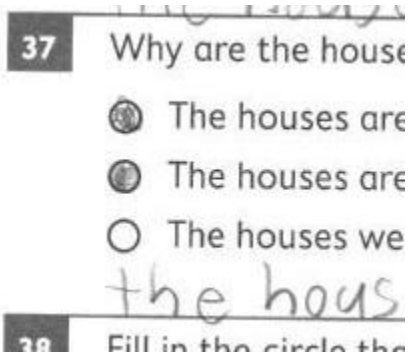
In the scanning process, we cut the spines off all the booklets and run them through high-volume scanners.

The scanning software looks at each question the student has bubbled in and decides whether it is choice A, B, C or D. The software also looks at the scores teachers filled in at in-person scoring sessions.

If a question is blank or looks like it has multiple responses, the software sends an image of that question for a human to check out. The person can see if the student filled in the question very lightly, used a checkmark, or crossed out an answer. The person will then update the answer to reflect what they see on the image.

There is only one correct answer for each question. If the question is genuinely blank or if the student bubbled in multiple answers with no indication as to which one the student preferred, we count that question as “incorrect” for the student.

If a student leaves too many questions blank, we consider that we cannot make a fair judgment of the student performance and we assign that student “Not Enough Evidence.”



This student bubbled in two answers but also wrote one of the answers below.

A human reviewed this question during the scanning process and indicated the written answer as being the intended answer.

Quality check: Looking at the assessment items

After the assessments and examinations are scanned and scored, we look at how the questions performed through the process of item analysis. Questions perform well when the question is clear, the right answer is clear, when students who perform well on the test overall perform well on the question, and when students who perform less well on the test perform less well on the question. If the question does not perform well (for example, if it seems there are two correct responses or the question was confusing), it is removed from the assessment.

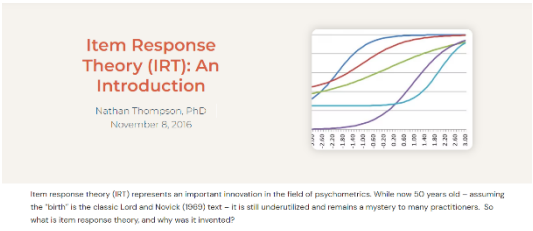
Questions that are very easy and questions that are very difficult can both give great information about student performance. Very easy questions "discriminate" (give information about) students at the lower end of ability levels and very difficult questions "discriminate" (give information about) students at the higher end. We try to include a wide range of questions for students of different abilities.

We calculate the student score based only on the questions that meet our performance guidelines.

The student score in Reading and in Mathematics is calculated using *3-parameter Item Response Theory (IRT).* The score represents the ability of the student as measured by the assessment. The score is calculated in a very different way compared to the "percentage correct" score students receive on traditional tests.

In IRT the score depends not only on number of questions the student got correct but also on the *characteristics of each question* (for example, the difficulty of each question).

While the scoring method is very different with IRT, the score the student receives in the end usually looks a lot like the score a student would receive with number correct. That's because both number correct and IRT scores are good methods of measuring ability.



This article provides a good introduction to IRT
Item Response Theory (IRT): An Introduction
Nathan Thompson, PhD (2016)
https://assess.com/item-response-theory/

Reading and Mathematics: Establishing cut points for each performance level

Under Item Response Theory, not only does each STUDENT get a score, but each ITEM gets a score. This score represents the difficulty of the item but also represents a student ability score[1].

After all the items are examined for quality and then analyzed using IRT, we create an Ordered Item Booklet. This booklet ranks each item in the assessment from least difficult to most difficult.

We then invite teachers and other education professionals to an expectation setting session. There, the participants carefully examine each item in the booklet, looking at the difficulty of the item and the thinking processes a student would have to draw on to answer the item correctly (*cognitive demand*).

Although difficulty and cognitive demand are related, they are not the same. For example, a question that asks a student to pick out a word from a reading text draws on a simple cognitive process (retrieval). But if the word is placed in a long or complicated text, the question can still be quite difficult.

We also provide the participants with detailed performance level descriptions. These descriptions showewhat kinds of knowledge, skills, and abilities students would have to draw on to answer questions at each level. For example, a student at level 1 in reading should be able to retrieve information directly from the text. At student at level 4 should be able to make complicated connections between different pieces of information in the text.

Using these descriptions and the ordered item booklet, each participant places a bookmark where they think level 1 ends, another where they think level 2 ends and a third where they think level 3 ends. In the end, the bookmarks that the teachers and education professionals provide are the foundation for the cutpoint for each level.

Because cutpoints are based on ability scores, and not the number of correct items, we can easily compare items, assessments, and student performance across years.

Assigning performance Levels

Once we have cutpoints telling us what the ability score at the top of each level is, it is relatively easy to assign a level to each student. For example, if the top of level 3 is ability score “1.3”, and a student has an ability score of “1.5”, that student is in level 4.

However, it is also clear that there is a range of ability in each performance level. Some students in level 3 are closer to a level 2, and some students are closer to a level 4.

[1] The difficulty of an item in IRT is usually defined as the student ability score at which the student has a 50% chance of getting the item right.

Reading and Mathematics: Drawing the dot for each student

Different performance levels have different ranges of ability. For example, there are relatively few students in performance level 4 but they have a wide range of ability. There are many more students in level 3, but the range of ability represented is much smaller.

Think of height – most people are within a few cm of average height, but the range of human heights is very wide at the extremes.

To draw the dot for each student, we took the range of ability in each performance level and placed the student's ability score where it fell in that range. If the student's ability score fell halfway in the range, the dot will be in the middle of the level. If the student's ability level falls at the top of the range, the dot will be close to the next level.

There are more than a hundred places the dot can be drawn in each level (there are thousands of individual ability scores in the province). If the dot is touching the top edge of level 4 or the bottom edge of level 1, that represents the limits of the graph. The student may have an ability level beyond the graph but we felt that the range we chose to represent is a good indication of the variety of student ability[2].

Sun Mingming, world's tallest professional basketball player at 2.36 m.

Reading and Mathematics: Component scores

For subcomponents (e.g. genres, strands), there are only two places the checkmark can be: to the left ("May require further attention") or to the right ("At or above the expectation").

This is because there are many fewer items feeding into the checkmarks (subcomponents of the assessment) than to the overall score.

We set a curriculum expectation for each component using multiple sources of feedback. These include the judgments of ability for the overall score from the bookmark session, professional judgment, careful reference to curriculum documents, and consultation.

We calculate a student score for each subdomain and compare it to the cutpoint. If it is at or above the cutpoint, the checkmark is "At or above". If it is below, the checkmark is "May require further attention."

may require further attention	at/above the expectation
	✓
✓	

[2] We chose a range of ability of three standard deviations above or below the mean ability. Anything beyond +3 or -3 is rounded to +3 or -3.

Writing: Scores for each component

There are two writing tasks in our Reading and Writing assessments. In order to receive a score for writing, a student must attempt both tasks, and provide enough writing to be scored for each task. Otherwise the student received a score of "Not enough evidence" in writing.

Both tasks are scored by teachers using the same Analytic Rubric (available on our website at http://plans.ednet.ns.ca). For assessments, the rubric is based on four categories: Ideas, Organization, Language Use, and Conventions. For each of these categories, the student receives a performance level score from 1 to 4.

Each of the two tasks is scored separately by two teachers. The second teacher does not look at the scores given by the first teacher, so the judgments are independent. In order to ensure fairness and reliability, if the two writing scores are very different for a component (more than a level apart), a third teacher provides a score, and the two scores that agree most closely are kept.

Because at least two teachers look at each of two tasks, there are four separate judgments of student ability in each category. In order to provide one score for each category, we take the median of the scores. If the median falls in between two levels, we round up to the next level, since we know the student had at least two scores above the median.

Here is an example of a fictional student's writing performance and how the score is calculated for each component.

Category of writing	Writing Task 1		Writing Task 2		Median	Final score
	First scorer	Second Scorer	First Scorer	Second Scorer		
Ideas	2	3	3	4	3	Level 3
Organization	3	4	4	4	4	Level 4
Language Use	1	2	3	3	2.5	Level 3
Conventions	3	4	4	4	4	Level 4

Below is an example of how those scores have been reported up until 2021-2022 school year. Each individual criterion was reported separately with a level.

Writing Performance	Not enough evidence too little writing provided	Level 1 below the expectation	Level 2 approaching the expectation	Level 3 at the expectation	Level 4 above the expectation
Ideas				✓	
Organization					✓
Language use				✓	
Conventions					✓

New in 2022-2023: Student writing will be reported as an overall score (with a dot similar to reading and mathematics) and component scores will be reported as May Require Further Attention (levels 1 and 2), At/above the Expectation (levels 3 and 4).

The scoring of writing will continue as before, but after the scoring session, we will use item response theory to generate a scaled score for each student in writing. We will also conduct expectation setting sessions to set cut scores in writing. The advantages of this new system include:

- Simpler interpretation of writing scores over time.
- Emphasise writing as a holistic process while still providing detailed information on the four components of writing to regions and schools.
- More reliable results, as more scores go into making the overall score.
- Ability to equate writing over time.

Proposed new student report (comparable to reading and mathematics)

Writing Results

The student demonstrated a writing performance of Level 3 on this assessment.
Students writing text at performance Level 3 can generally communicate clearly in writing. They produce a clear main idea with relevant details and organize their writing with effective use of transitions. The language they use contributes to clear and fluid writing, and they generally demonstrate correct use of a variety of conventions.

	Not enough evidence	Level 1 below the expectation	Level 2 approaching the expectation	Level 3 at the expectation	Level 4 above the expectation
Writing Performance				●	

The table below shows how the student performed on the components of writing on this assessment. This information may guide teachers in looking further into the student's achievement on various components of reading.

Writing Components	not enough evidence	may require further attention	at/above the expectation
Ideas			✓
Organization		✓	
Language use			✓
Conventions		✓	

For more information

Please visit our website (http://plans.ednet.ns.ca) or contact Sarah Curry (sarah.curry@novascotia.ca).

NOVA SCOTIA