

From Questions to Scores

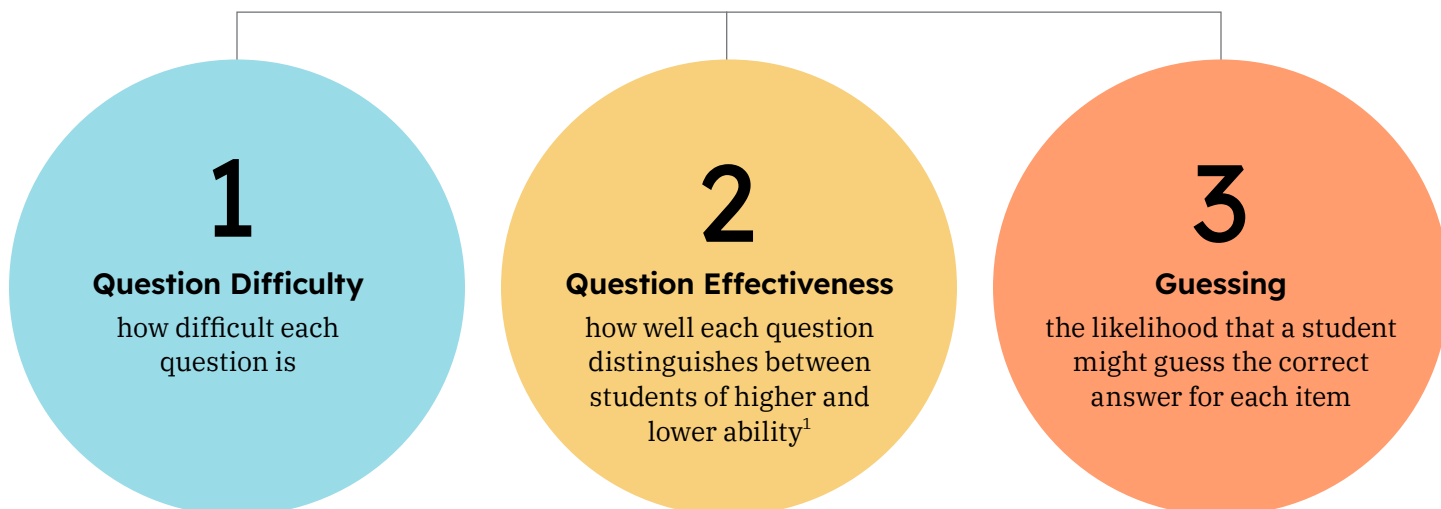
How Item Response Theory Enhances EQAO Assessments

Like many large-scale assessment developers around the world, EQAO uses statistical methods informed by a set of ideas called **item response theory (IRT)** to measure how much students know and can do in three key subjects: reading, writing and mathematics. While the name may sound complex, IRT's role is fairly simple: it ensures that assessments are as fair and accurate as possible for every student. Here is an overview of what IRT is and how it improves EQAO assessments.



What Is Item Response Theory (IRT)?

IRT is a sophisticated statistical method used to design assessments and evaluate student performance. It allows for better test design and interpretation of results by taking into account the three following characteristics of questions:



By analyzing these characteristics of each question, IRT measures student ability on a common and comparable scale, even when students answer different sets of questions.

¹ The term “effectiveness” here is more commonly referred to as “discrimination” in psychometric literature.



Selecting High-Quality Questions Using IRT

Understanding the characteristics of each question is essential for selecting the best questions for EQAO assessments. To achieve this understanding, questions are field tested with students throughout the province, and the IRT-based statistical data collected undergoes a rigorous review process.

This ensures that questions are

- **appropriately challenging:** neither too easy nor too difficult for the intended student population and
- **effective at distinguishing ability:** clearly differentiating between students at different ability levels.

By carefully selecting and refining questions based on field-test data, EQAO enhances the overall quality of assessments and, in turn, ensures that student scores accurately reflect their abilities.

IRT also strengthens the ability to detect and eliminate potential bias in test questions. EQAO uses a process called differential item functioning (DIF) analysis to identify questions that may unfairly favour or disadvantage certain student groups.

For example, a question may unintentionally be easier for certain groups of students due to differences in gender or linguistic background. DIF analysis flags such questions, allowing EQAO to revise or replace them, so that all students are assessed fairly with the best questions possible. By leveraging IRT, DIF analysis can be conducted with higher accuracy, leading to fairer and more equitable assessments.



IRT also strengthens the ability to detect and eliminate potential bias in test questions.

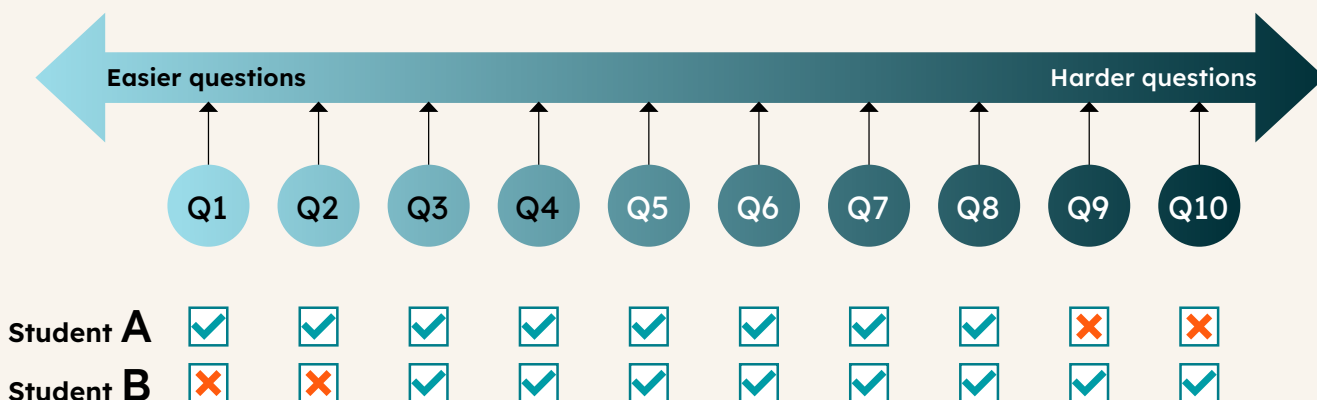
How IRT Determines Student Scores

IRT goes beyond simply counting the number of correct answers by examining the characteristics of each question a student receives, such as its difficulty and effectiveness in measuring student ability. This approach provides a more precise way of measuring student performance, compared to traditional methods.

For example, imagine a test with 10 questions:

Student A answers all but the two hardest questions correctly.

Student B answers all but the two easiest questions correctly.



With traditional scoring, both students would receive the same result, for example, 80%. However, with IRT, their scores would differ, because IRT accounts for the characteristics of the questions.

In IRT, a student's score is determined through statistical models that estimate the student's ability based on the characteristics of the questions answered, such as difficulty, effectiveness and guessing factor, as we saw above.

Question Difficulty

If a student answers hard questions correctly, there is stronger evidence that the student has a higher ability than if the student answered only easy questions correctly.

Question Effectiveness

IRT gives different weights to questions depending on how well they can distinguish between students with higher and lower ability.

Guessing

IRT accounts for the chance of guessing to reduce the likelihood of scores being overestimated just because a student guessed some multiple-choice questions correctly.

This depth of analysis enables IRT to produce a more nuanced and reliable measure of student performance, rather than relying simply on the number of correct answers.

Creating Different Test Versions

In EQAO assessments, students may receive different sets of questions for several important reasons. One key reason is test security. Because school administrators and teachers can schedule assessments at different times, using multiple test versions helps prevent students from sharing test content.

Even though students may get different questions, IRT ensures fairness in several key ways:



Analyzing each question carefully to understand its difficulty and effectiveness



Placing all questions on a common scale, so that scores reflect student ability, regardless of the specific set of questions students received



Creating multiple test versions that are comparable in overall difficulty, effectiveness and content coverage

This process ensures that scores are comparable across students, no matter which version of the assessment they take.



Comparing Student Performance Over Time

The same process of using a common scale applies when comparing student performance across different years. When new assessment questions are introduced, IRT ensures that these questions are placed on the same scale as questions from previous years. This means that the scores remain on the same scale consistently from year to year. Through the application of the same rigorous methods each year, scores can be compared across different years, enabling the tracking of student performance trends over time.

Key Takeaways

1

IRT helps **select high-quality assessment questions** by analyzing their characteristics (question difficulty, effectiveness and guessing). It also strengthens EQAO's ability to identify and replace biased questions, which in turn promotes fairness.

2

Unlike traditional scoring, which simply counts correct answers, IRT estimates student ability based on the characteristics of the questions students received. This results in **more accurate scores** that are less influenced by lucky guesses or the specific set of questions received.

3

IRT allows EQAO to **create multiple test versions that are comparable** in overall difficulty, effectiveness and content coverage. This ensures fairness and maintains test security while keeping scores comparable across different versions.

4

IRT ensures that scores remain on a common scale, making it possible to **compare results across different years** and track student performance trends reliably.

