

Computer Adaptive Assessments

Item Response Theory vs. Curriculum Based Theory

A White Paper April 2011

“For a long time educational testing has focused mainly on paper-and-pencil tests and performance assessments” (van der Linden and Glas, 2000). Due to the growing popularity of personal computers in education in the late 1980s, computer-based delivery of tests has become common place. Not only does computer-based testing allow for greater flexibility for teachers and students, according to van der Linden and Pashley, statistical accuracy of test scores is increased by using computer adaptive testing (CAT). “Instead of giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees. After each response, the examinee’s ability estimate is updated and the subsequent item is selected to have optimal properties at the new estimate” (van der Linden and Glas, 2000). For many decades, linear tests were

Item Response and Curriculum Based Theory: two different theories for two different uses. The first theory is meant to calculate a very accurate scale score to determine a student's level on a given scale; the later theory is used to identify specific learning objectives students missed in their current and/or previous grades.

The Item Response Theory

“The development of item response theory (IRT) in the middle of the last century has provided a sound psychometric footing for CAT.”

IRT is based on a three (3)-parameter model that selects the student's next question based on his response to the preceding question. This "brackets" the student's knowledge by starting with a question on or near the student's current grade level. If the student answers the question incorrectly, the test will drop down a specific portion of the scale (based on the formula below) and display another question that is at a lower level than the question he answered incorrectly. If the student gets that question wrong, the test will continue this downward progression until the student answers a question correctly. Once he answers a question correctly, the bottom bracket has been assigned and the next question will go up a specific portion of the scale (usually one grade level) and display another question. If the student answers that question incorrectly, the test engine will go in down mode. This bracketing continues until the standard error of measurement reaches a preset limit. The difficulty indices are calculated at the item or question level. Each question has its own difficulty index. This theory uses each question on its own merit to determine the movement of the test. The following is the formula for the three parameter logistic model (3-PL),

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}}$$

Where:

$P(\theta)$ is the probability of someone with proficiency θ responding correctly to the question.

a represents the *item discrimination* and characterizes the slope of the item characteristic curve.

b represents the difficulty of the question.

c represents a binomial floor on the probability of getting an item correct. On a multiple-choice test with four answers, $c = .25$

The Curriculum Based Theory is dependent on the database structure and on algorithms that use a back-mapping strategy for each learning objective. Each learning objective has defined prerequisite learning objectives that have their own prerequisite objectives and so on. Each learning objective also has its own “going forward” learning objective, similar to the prerequisite learning objectives; they are the skills that directly follow the learning objective. The algorithms for this model create virtual silos of each learning strand that can be bracketed. This model ties the difficulty index to the learning objective *instead* of the question. Bracketing occurs at the learning objective level and includes an initial leveling of the program. To assure validity when measuring a student’s skill level, each learning objective should have multiple questions that are used to determine whether a student has mastered a specific skill or whether he should receive remediation. This model uses the dynamic bracketing approach of the Item Response Theory along with the traditional formative assessment model of testing each learning objective with multiple questions to ensure mastery. The computation is vastly different from the Item Response Theory in that this theory is highly dependent on the database structure of the curriculum. The difficulty indices are tracked at the skill level and verified through multiple exposures of test takers. This theory uses the vertical assessment approach of the traditional computer adaptive assessments to find the floor or a student’s base level, and then automatically employs traditional formative or benchmark assessment routines to verify mastery of each learning objective. This type of test is usually administered at the strand level of a subject because of the interwoven intricacies of the virtual back mapping. One obvious advantage is that the results are black and white, either a student mastered the skill or he didn’t; instead, the results will also indicate the depth of a student’s skill gap. This can make a *huge* difference in where the teacher begins remediation. Is the student missing only the two prerequisite

skills leading up to the current learning objective, or is the gap more severe? Using the vertical assessment approach, a teacher can drill all the way down to the highest prerequisite skill that the student has mastered. In some cases, an eighth grade student may have skill gaps that go all the way back to third grade.

Consider the following, and very common, scenario: On the day that a seventh grade pre-algebra teacher introduces a unit about solving equations, she administers a diagnostic assessment test to her students that came with the math textbook. There are twenty-five multiple-choice questions with four possible answers.

When the teacher evaluates the class results at the end of the day, she arranges the students into three groups, a high group, an average group, and a low group. By simply looking at the test results alone, she has placed two students in the low group because they received failing grades on the assessment. The problem is that one student had a very minor skill gap. He didn't remember one prerequisite skill that caused him to answer the majority of questions incorrectly. The skill was using the order of operations to solve equations. Most likely, if he receives remediation for this one skill, he will be successful in learning to solve more advanced equations.

The second student that the teacher has placed in the low group also scored poorly on the diagnostic assessment. However, it will take *much* more than one remediation session to bring this student back to grade level. His deficiencies go all the way back to third grade when he failed to master multiplication. If he didn't master multiplication, he certainly can't perform division. It would also be almost impossible for him to be successful with problems involving exponents or square roots. If he had been given an assessment that drilled all the way down to the level where he did demonstrate mastery, the teacher would have quickly seen his enormous skill gap. And while the first student might have success solving equations after only one remediation session, the second student may need several weeks or even months of intensive remediation because he didn't have a chance to master any skill that required multiplication in third grade.

While there is a place for both of these types of computer adaptive assessment, it is important that educators have access to the appropriate type of diagnostic assessment when evaluating their students' current performance level.

Works Cited

The Basics of Item Response Theory. Baker, Frank. ERIC Clearinghouse on Assessment and Evaluation. 2001

Computerized Adapted Testing: Theory and Practice. Wim J. van der Linden and Cees A.W. Glas., Editors. 2000. Kluwer Academic Publishers: The Netherlands.

Lin, C.-J. (2008). Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel Test Forms. *Journal of Technology, Learning, and Assessment*, 6(8). Retrieved 11-4-2010 from <http://www.jtla.org>.