

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Office of State Assessment
Albany, New York 12234

The Use of Item Response Theory (IRT)
on New York State Fourth and Eighth Grade
ELA and Mathematics Tests

The probabilities that a student would correctly answer the questions on the New York State Assessment tests in *fourth and eighth* grade were estimated using item response theory (IRT) models. Item response theory is a statistical procedure that takes into account the fact that not all test questions are alike and that all questions need not be given equal weight to determine how much students really know and can do. Teachers often give differing values (weights) to questions on tests they devise, depending on how much they think each question contributes to a student's understanding of the student's *knowledge* of the *subject* being tested. Computer programs that implement IRT models for weighting questions use actual students' data to estimate the characteristics of the questions on a test - called "parameters." The parameter estimation process is also called "item calibration."

For the New York State tests, three parameters are estimated: the discrimination parameter, the difficulty parameter(s), and, for multiple-choice questions, the guessing parameter. The discrimination parameter measures how well related answering a question correctly is to overall performance on the test as a whole. A question that lowperforming students cannot answer correctly, but high-performing students can will have a high discrimination value. The difficulty parameter is an index of how easy or difficult a question is. The higher the difficulty parameter, the harder the question. The guessing parameter is the probability that a student with very low ability will answer the question correctly.

The estimated parameters are then used to determine weights for the questions in computing student scale scores. The scale score (SS) is the basic score for the New York State tests. It is used to derive other scores that describe test performance, such as the four performance levels and the standard-based index scores (SPIs). For interpretations of various scores reported on the New York State Assessment Score Reports, see "Guide to New York State Testing Program Score Reports." Scale scores can be obtained by one of two scoring methods: IRT item-pattern scoring, or number-correct scoring.