In 1997, based on the amendments to Individuals with Disabilities Education Act (IDEA), all states were faced with a statutory requirement to develop and implement alternate assessments for students with disabilities unable to participate in the statewide large-scale assessment. States were given the challenge of creating, implementing, and executing these tests within a short period of three years. In most cases, innovative assessments were developed by tweaking, combining and modifying traditional assessments already in place for non-disabled students. When changes were made to the traditional assessment program, the psychometric properties associated with the test and scores are likely to differ due to the unique characteristics of the student population and testing program. However, few changes had been made to the traditional analyses of scores of students with disabilities.

This study examines the impact of changes to three psychometric elements - design, examinee population and analysis issues on ability and item estimates from the assessment computed using Rasch measurement model. Using data from a statewide selected response alternate assessment designed for students with disabilities in southern United States, the impact of test design and analysis decisions related to partial credit scoring, analyses within homogeneous groups and treatment of lead questions were examined to determine the effect on test outcomes.

We examined estimates of item difficulty and ability as well as a select sample of items and students’ scores to compare them to estimates from “original” analyses. The results indicated
considerable variability of student scores and differential impacts on item measures based on the choices or a combination of choices made during analyses. Clear patterns of change in scores when the lead questions were treated as correct seem to suggest a systemic change in item measures. High correlations between student scores from the homogeneous and the heterogeneous group analyses indicated little difference between the two. The findings in this study provided evidence for building validity evidence for an assessment, decisions made in computing psychometric measures need to be explicitly based on theory and aligned to the use and the interpretations of scores.