World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:18, No:08, 2024

National Assessment for Schools in Saudi Arabia: Score Reliability and Plausible Values

Authors : Dimiter M. Dimitrov, Abdullah Sadaawi

Abstract : The National Assessment for Schools (NAFS) in Saudi Arabia consists of standardized tests in Mathematics, Reading, and Science for school grade levels 3, 6, and 9. One main goal is to classify students into four categories of NAFS performance (minimal, basic, proficient, and advanced) by schools and the entire national sample. The NAFS scoring and equating is performed on a bounded scale (D-scale: ranging from 0 to 1) in the framework of the recently developed "D-scoring method of measurement." The specificity of the NAFS measurement framework and data complexity presented both challenges and opportunities to (a) the estimation of score reliability for schools, (b) setting cut-scores for the classification of students into categories of performance, and (c) generating plausible values for distributions of student performance on the D-scale. The estimation of score reliability at the school level was performed in the framework of generalizability theory (GT), with students "nested" within schools and test items "nested" within test forms. The GT design was executed via a multilevel modeling syntax code in R. Cut-scores (on the D-scale) for the classification of students into performance categories was derived via a recently developed method of standard setting, referred to as "Response Vector for Mastery" (RVM) method. For each school, the classification of students into categories of NAFS performance was based on distributions of plausible values for the students' scores on NAFS tests by grade level (3, 6, and 9) and subject (Mathematics, Reading, and Science). Plausible values (on the D-scale) for each individual student were generated via random selection from a statistical logit-normal distribution with parameters derived from the student's D-score and its conditional standard error, SE(D). All procedures related to D-scoring, equating, generating plausible values, and classification of students into performance levels were executed via a computer program in R developed for the purpose of NAFS data analysis.

Keywords : large-scale assessment, reliability, generalizability theory, plausible values
Conference Title : ICESEA 2024 : International Conference on Educational Statistics, Evaluation and Assessment
Conference Location : Amsterdam, Netherlands
Conference Dates : August 05-06, 2024