Using Item Response Mixed (IRM) Models to Improve the Comparability of Educational Assessment Scores

Tufi M. Soares*
Federal University of Juiz de Fora, Juiz de Fora, Brazil Tufi@caed.ufjf.br

In educational assessment, Differential Item Functioning (DIF) occurs when students with the same proficiency levels have different probabilities of giving the same answer to some of the test items. DIF can be a hindrance to the quality of the comparability of measures produced by large-scale evaluations. The problem can be particularly serious in evaluations that aim at making international comparisons of education quality, such as for instance the PISA program, which deals with a considerable amount of cultural, linguistic and curricular differences among its participating countries. Traditional methods that analyze the problem have a generally good performance when DIF is an exception, restricted to a few items of the test. However, this is not the case when, for instance, one compares the functioning of the items in a test administered in Japan and Brazil. Recently new methods for DIF detection have been proposed, based on IRT procedures. Among them, it is worth mentioning those that employ Item Response Mixed (IRM) models for DIF analysis, which allows them to treat the problem even under unfavorable circumstances, as in the cases when there is a considerable amount of DIF in the test items. This article makes a review or these models and proposes new DIF analysis methods that are alternatives which are safe enough to guarantee the quality of the comparability of the results. Simulations and applications using results of PISA and some educational tests in Brazil are presented in order to show the efficacy of the methods.

Key words: Differential Item Functioning, Item Response Theory, Comparability of International Assessment Scores