Differential Item Functioning Analysis of High-Stakes Test in Terms of Gender: A Rasch Model Approach

[1] University of Tehran, Tehran, Iran Email: Mohammed.Alavi@gmail.com
[2] PhD candidate of TEFL, Tehran University, Tehran, Iran Email: ut.sbordbar@yahoo.com

Seyed Mohammad Alavi [1], Soodeh Bordbar [2]

ABSTRACT

Differential Item Functioning (DIF) analysis is a key element in evaluating educational test fairness and validity. One of the frequently cited sources of construct-irrelevant variance is gender which has an important role in the university entrance exam; therefore, it causes bias and consequently undermines test validity. The present study aims at investigating the presence of DIF in terms of gender in a high stakes language proficiency test in Iran, the National University Entrance Exam for Foreign Languages (NUEFL). The participants’ responses (N = 5000) were selected randomly from a pool of examinees who had taken the NUEFL in 2015. The results displayed DIF between male and female test takers. Hence, on the basis of the findings, it is concluded that the NUEFL test scores are not free of construct-irrelevant variance and the overall fairness of the test is not confirmed. Also, both Rasch assumptions (i.e., unidimensionality and local independence) are hold in the present research.

Keywords: Differential Item Functioning, Dimensionality, Rasch Model
INTRODUCTION

In language testing and educational measurement the discussions about test use and the consequences of tests have increased. Since the National University Entrance Exam for Foreign Languages (NUEEFL) is administered annually to a large number of test takers country-wide in Iran, the consequences of failure on the test are serious. It could result in spending one or more years for test preparation and two-year military service (for males).

Therefore, it is essential to examine the extent to which the instrument assesses what it is intended to measure (validity) as well as the test consistency (reliability) (Pae, 2011) in measuring the English ability in the high-stakes test, such as NUEEFL. Nonetheless, despite the heated nature of the debates, there has been little empirical evidence for the validity of the NUEEFL test and its fairness. Specifically, there is no ample evidence of test fairness among male and female test takers. In the absence of such evidence, any talk of the fairness of the selection policy would be doomed to fail.

The present study aims at investigating the validity of a high-stakes test in general and to considering the role of gender as a source of bias in the NUEEFL, in particular. Regardless of the content of the debates over the gender issue, it appears that there is no evidence on the effect of gender on the performance on the NUEEFL. If gender asserts a large influence, then it would be a case of bias and will undermine validity of the test. This is because gender is not part of the construct measured by the test and any significant impact by gender is a case of construct-irrelevant variance. As a part of standard process, Differential Item Functioning (DIF) analysis is conducted on the test items, as a main factor in the evaluation of the fairness, and validity of educational tests.
(Doctoral dissertation, University of Northern Colorado).

