

TEST BIAS AND DIFFERENTIAL ITEM FUNCTIONING

Randall E. Schumacker, Ph.D.

Introduction

Test score validity is of primary importance in a certification program because it pertains to the inference we can make from a persons' test score. If a test favors one group of examinees over another, the test is considered biased and violates the principle of test fairness. One potential threat to test score validity is item bias, which occurs when a test item unfairly favors one group of examinees over another, meaning that one group more often correctly answers the item. A biased item will exhibit differential item functioning (DIF). DIF occurs when examinees from different groups with equal knowledge exhibit different probabilities of success on an item. It is important to note that differential item functioning, in and of itself, is not evidence of item bias. A difference in item responses would be expected when the examinee groups differed in knowledge. Consequently, a difference in item performance obtained from groups of examinees with different ability levels does not represent item bias, but rather item impact.

Statistical procedures

While a variety of DIF screening procedures have been developed, a relatively few number of methods have been recommended. The Mantel-Haenszel method is popular, but a logistic regression approach is currently supported. The logistic regression approach compares attribute variables (gender, ethnicity, age) and/or item parameters associated with two groups: a group designated as the focal group and a comparison group designated as the reference group. In the absence of DIF, item characteristic curves (ICC) for the two groups will be the same. In the simplest case of DIF, items may differ across groups solely in terms of difficulty. In more elaborate analyses numerous attribute variables and/or item parameters (discrimination, difficulty, and guessing) may differ. The Mantel-Haenszel and logistic regression approaches offer distinct advantages. Thus, it is important that the selection of the approach used reflect the unique conditions of the measurement study. For example, the Mantel-Haenszel method can be used with smaller sample sizes, while logistic regression, which can be conceptualized as a link between the contingency table method (Mantel-Haenszel) and IRT method, offers a more robust solution under both uniform and non-uniform DIF conditions.

Policy decisions

While the specifics of DIF detection methodology are guided by technical decisions, policy decisions govern the process. The decision to implement a DIF study is, in itself, a policy decision. Other decisions would include when to implement a DIF study, selection of the focal group, and what is to be done with items that display DIF.

Because DIF is a necessary, but not sufficient condition for item bias, a very important decision is whether to consider items exhibiting DIF as biased items until proven valid or unbiased items until proven biased. As with any test validation study, DIF analysis is a process of collecting evidence. Our trained personnel are available to assist professional associations in weighing, interpreting, and making sound policy decisions regarding efficacy of training and associated measurement and assessment procedures.