

Detection of Differential Item Functioning for Polytomous Items

Ya-Hui Su

National Chung Chang University; Department of Psychology, National Chung Chang University, Chia-Yi, 621, Taiwan; E-Mail: d9132001@ccu.edu.tw

Wen-Chung Wang

National Chung Chang University; Department of Psychology, National Chung Chang University, Chia-Yi, 621, Taiwan; E-Mail:psywew@ccu.edu.tw

Keywords: item response theory, Mantel, generalized Mantel-Haenszel, graded response model, partial credit model.

Abstract

This study compares the performances in DIF detections via the Mantel method and those via the generalized Mantel-Haenszel method (GMH) on tests with polytomous items. The dependent variables were (a) DIF detection method, (b) item response model, (c) ability difference between groups, (d) test length, (e) DIF type, (f) effect size of DIF, (g) DIF percentage, and (h) DIF purification. The dependent variables were empirical Type I error rate and statistical power. One thousand replications were made under each condition. The results showed that both methods performed very well under the partial credit model but less satisfactorily under the graded response model. The power of the Mantel method was higher than that of the GMH method under all but the balanced-within-item conditions. The larger the DIF effect, the larger the power was. When the mean ability difference between groups was as large as one standard deviation and the test length was short, both methods yielded inflated Type I error rates, even when the test did not have any DIF item. When the mean ability difference was large, the purification procedure for the two methods did not function well. If the mean ability difference was small and the test condition was constant-item/balanced-test, both methods yielded reasonable DIF detection results, even if the percentage of DIF items in the test was as high as 20% or 50%, as long as the mean item difficulty difference between groups approached zero. Under the balanced-within-item condition with large DIF effects, both methods yielded inflated Type I error rates. Except under the balanced-within-item condition, the mean item difficulty difference could predict the Type I error rate very well for both methods, especially under the partial credit model, as long as the mean ability difference was not very large. For conditions where the mean ability difference is as large as one standard deviation, DIF analysts may consider using the methods directly based on item response theory.