AN OBJECTIVE PROCEDURE FOR DETERMINING IF ITEMS HAVE GOOD DISCRIMINATION

Wen-Chung Wang, Lai-Fa Hung

Department of Psychology, National Chung Cheng University

Item difficulty and discrimination analyses are the two most important item analyses. Several indices such as the index of discrimination, the point-biserial correlation, or the biserial correlation, and Fleiss's odds ratio have been proposed to depict item discrimination power. Although we could test if these conventional discrimination indices are significantly different from zero, there is no objective criterion to determine how large they should be for an item to have good discrimination. Practically, test analysts usually use 0.3 or 0.4 as a cut-point. If the index of discrimination or the point-biserial correlation exceeds the cut-point, the item is flagged as exhibiting good discrimination power. In addition to the drawback of no objective criterion available, these indices depend on sample characteristics and item difficulty, for example, they will yield higher values for item difficulty (i.e., passing rate) close to 0.5 than for it at the extremes of difficulty.

Although the cut-point may be a useful guideline, a statistical procedure is preferred. This study attempts to establish an objective statistical procedure for determining if an item has good discrimination or not. To do so, good discrimination is first defined. An item is said to have good discrimination if *it discriminates every score point equally well for the target population.* We find this definition appropriate because every score point is considered equally important. We use a linear relationship between the probability of passing an item and the test score to depict how the regression should look like when an item has good discrimination. For binary outcome variables, the logistic distribution can better depict the relationship between probabilities of passing an item and test scores. In order to hold the equal discrimination power assumption, we then derive a logistic regression curve that is closest to the ideal discrimination line. Once this "theoretical" logistic regression curve is derived, the observed logistic regression curve, derived from test data, could be compared to the theoretical logistic regression curve. If the observed logistic regression curve is statistically different from the theoretical one, the item is said not to have good discrimination.

A simulation study was conducted to compare the detection of item discrimination with the linear regression model and the logistic regression model when the underlying test score distributions follow the normal, uniform, or chi-square distribution, and the sample sizes are 40, 100, 500, 2000, or 5000. When the test scores follow the normal or chi-square distribution, the linear model and the logistic regression model yield almost identical results. Only when the sample sizes are extremely large, say up to 5000, would these two models yield different results. A real data set with 50 multiple-choice items and 500 examinees was analyzed to illustrate the similarity and difference between the proposed method of logistic regression model and the conventional item discrimination indices. Five items were arbitrarily chosen and analyzed. The item difficulties (percentage of correct responses) of these five items are between 0.50 and 0.81. Only one item is flagged as not exhibiting good discrimination with a p value of 0.000. Basically, the three conventional discrimination indices lead to almost identical results. This is expected because all of these procedures are invented to depict item discrimination, however, only the proposed objective procedure is statistically sound.

Keywords: Item discrimination, Logistic regression model, Likelihood ratio test, Pearson chi-squared test