COMBINING TIME AND CORRECTNESS IN THE SCORING OF PERFORMANCE ON ITEMS

Gunnar R. Bergersen

Conference track: Psychometrics

Type of paper: Theoretical / methodological (differential item functioning)

Short abstract: Performance on items that vary simultaneously on time and correctness dimensions can be scored as Guttman-structured data, a requirement for use of the polytomous Rasch model. We show how differential item functioning can be used to detect whether overfitting has occurred during construction of the scoring rules.

Background: The polytomous Rasch model is an extension of the dichotomous Rasch model where successive ordinal integers are used as scores. The scoring structure is required to follow a probabilistic Guttman structure that imposes specific constrains on how items can be scored. In general, an individual must pass all thresholds below, and fail thresholds above, the level that is reflected in the item score.

An individual's performance on an item with a constructed response format often operates in a time versus correctness trade-off. Usually time and correctness are analysed separately, although these variables are closely related. For example, an improved (i.e., more correct) solution can usually be obtained if more time is made available. Further, if two solutions are deemed correct, a higher score could be given to the solution with the least time used, as this would indicate "better performance" on an item.

Aim: The aim in the present study is to show how time and correctness can be integrated in the scoring of performance on items. We present data from an instrument for assessing programming skill as an example (N = 65, items = 17) using the partial credit Rasch model. In the instrument, the level of programming skill is inferred from programming performance on programming tasks (items) varying in both time and correctness. The instrument is based on a previous confirmatory factor analysis of performance data were it was found that performance (i.e., time and correctness) on items could meaningfully be explained by a single latent variable.

Solution: The instrument we present uses the following general scoring structure for items: Starting from incorrect solutions (or solutions submitted too late), score points are first awarded to increases in solution correctness until the solution is deemed correct. Second, additional scores are awarded to correct solutions based on the time to obtain a solution; less time implies even higher scores. Further, calibration of the instrument was conducted in two steps. First, scoring rules for all items were constructed based on data from two-thirds of the subjects, randomly selected (*model building sample*). Second, we imported the remaining individuals (*model evaluation sample*) and checked for differential item functioning (DIF)

between the two data samples. Five out of seventeen items were found to be candidates for removal through this process.

Discussion: During the first step of the process described above, it is hypothesized that the Rasch model is true for making inferences about individual programming skill. In the second step, we attempt to falsify this hypothesis by importing new data that may not fit the model equally well as the original data.

During the adjustment of scoring rules, we regularly made adjustments to both the number of available (collapsed) scores for items and the item difficulty (threshold) locations. We found that the maximum number of score categories (without reversed thresholds occurring) varied from item to item. We acknowledge that item threshold locations are somewhat arbitrarily placed in the proposed solution. However, we do not see this as a problem, because we believe we are in congruence with the required probabilistic Guttman structure in the scoring of the items. Further, we argue that person ability estimates should not be impacted (in a major way) if the difficulty of an item threshold is, for example, increased through the readjustment of a specific scoring rule. However, because the readjustment of scoring rules for items is a data-driven process, model overfitting is likely to occur. This was, however, accounted for in the two-step data splitting procedure.

Conclusion: Results indicate that the polytomous Rasch model may be used to score performance data in a unidimensional manner, even when the variables used to infer each individual's position on the latent variable vary along both time and correctness dimensions. We further believe that the results may have applicability in other domains, such as achievement testing where constructed response formats are employed. Finally, we argue that splitting data into model building and model evaluation samples is a sensible way to validate the scoring rules from the perspective of DIF.