# DETECTING LEARNING AND FATIGUE EFFECTS BY INSPECTION OF PERSON-ITEM RESIDUALS

Gunnar R. Bergersen (gunnab@simula.no)
Jo E. Hannay

**Conference track:** Psychometrics

**Type of paper:** Theoretical / methodological

**Short summary:** When items are administered in random order, learning or fatigue effects can be detected by inspection of the averaged person-item residuals. The presence of practice and fatigue effects may also be detected by reduced reliability or reversed item thresholds; however, this can easily be misinterpreted as random instead of systematic error.

Background: A basic assumption in tests of achievement (or ability) is that the latent variable under investigation can be regarded as stable throughout the administration of a test. However in practice, individuals may sometimes display a systematic increase or decrease in their success-rate on items throughout a test (item difficulty accounted for). When individuals show a systematic increase in achievement throughout a test, this is called a practice effect. Oppositely, if individuals show a systematic decrease in achievement throughout a test, this is called a fatigue effect.

Systematic practice or fatigue effects cause several concerns. Assume that a practice effect is present and that all persons are presented with the same items for the purpose of calibrating item difficulty and estimating person ability. Then, from the perspective of person ability estimates, a threat to validity may be present if substantive theory dictates that the latent variable under investigation is supposed to be stable. Further, the practice effect will bias item difficulty parameters; for example, if an item is consistently presented as the first item of a test, item difficulty will be biased upwards (the item will appear more difficult than it really is). This may, further, limit the feasibility of computer adaptive testing, as item parameters are dependent on the order of when they appeared during the calibration of the test.

Aims: We propose a method for inspecting whether there is a systematic bias present in person ability and item difficulty parameters due to practice/fatigue effects by the following: Both person ability and item difficulty parameters are in the Rasch model represented as constants which determine item responses according to a probabilistic function. The person-item residual is the discrepancy between model expectations, inferred from an individual's total score, and a single observed response. A negative person-item residual implies that the individual received a lower item score than expected, while a positive person-item residual implies that an individual received a higher item score than expected.

Solution: In our proposed solution, items are first presented in randomized order for each individual. This implies that potential learning or fatigue effects in the calibration of item difficulty parameters are evenly spread over the duration of the test. Second, we average the standardized item-person residuals conditional on item order and display the averaged value depending on item order. If, for example, a practice effect is present, this can then be detected by an increasing trend of the averaged person-item residual depending on item order.

We demonstrate the proposed method using three simulated data sets that contain (1) no learning/fatigue effect, (2) a moderate learning effect and (3) a strong fatigue effect. Further,

we show that the proposed method can clearly distinguish between the three datasets. Moreover, we contrast these results with real data from an achievement test on programming skills. In our own data, there seems to be a slight presence of a learning effect for the first three items presented to an individual.

Discussion: Conventional Rasch analysis does not account for item order. By using a conventional analysis of the simulated data, we did, however, indentify that if a learning or fatigue effect is present, it will mainly be detected as reduced reliability or an increased presence of reversed item thresholds. The paired t-test for unidimensionality was, however, seemingly unaffected by learning/fatigue effects. This implies that learning/fatigue effects may go undetected in a conventional Rasch analysis.

Further, we note that both a practice effect and a fatigue effect may be present in the same data, effectively cancelling each other out. We also discuss the utility of the proposed method when the goal is to maximise the amount of information about an individual's ability, given a fixed time limit for the test. Further, we argue that if small systematic learning/fatigue effects are present, items in a test should be calibrated using randomized item order.

Conclusion: We believe that the proposed method that employs random item order (both in administration and analysis) is useful for the empirical investigation of potential learning/fatigue effects. It is acknowledged that the proposed procedure may encounter problems where bias (differential item functioning) is already present in the data due to other sources.