A SYSTEMATIC REVIEW, EVALUATION, AND DEMONSTRATION OF TESTLET RESPONSE THEORY MODELS

Testlets are clusters of items that share the same stimuli. International and national assessments are often assembled with testlets. However, the measurement of the examinee’s ability might be influenced by the presence of testlets (Yen, 1993), as responses may be dependent on each other. As a result, testlet data might violate the assumption of conditional independence for item response theory (IRT; DeMars, 2006). When test items meet this assumption, responses that are locally independent are solely explained by the examinees’ ability measured by the test. That is, no additional variables influence the performance of such a test (Lord & Novick, 1968; Yen, 1993). However, this might not be the case for tests containing testlets. When performance is impacted by an additional variable in such a test, it is named the testlet effect. In this situation, traditional IRT models are inaccurate in assessing such test’s responses. Conversely, conventional statistics (e.g., Yen’s Q3) that are proposed to detect local independence assumption (or local item dependence) might not be useful in identifying this assumption for testlet data. Innovative models are needed to improve assessing responses and detection of local item dependence for testlet data. Testlet response theory (TRT) models may be able to assist in evaluating interrelated responses, detecting and accounting for local item dependence. Nevertheless, the TRT models lack (a) a systematic review about its performance, (b) information about its performance on parameter estimation, model selection, and accuracy for detecting local conditional dependence under different conditions (i.e., number of items per testlet, local item dependence, sample size, and missing values), and (c) use in applied studies. The purposes of this work are to review systematically the TRT literature, examine the performance of a TRT model under several simulated conditions, and provide an applied example. Two studies are conducted to accomplish these purposes. Collectively, the findings can inform the field about the current state of the literature with TRT models, situations encountered in practice when a TRT model is and is not useful, and illustrate the usage of a TRT model with actual data.