EXAMINING CONTENT VALIDITY OF THE INDUCTIVE REASONING TEST BASED ON RATER ASSESSMENT USING MANY-FACET RASCH MEASUREMENT

Soeharto Soeharto

University of Szeged, Doctoral School of Education

Keywords: inductive reasoning; rater assessment; content validity

This study aimed to confirm the validity of the inductive reasoning test based on rater assessment; to investigate raters' severity/leniency in assessing the inductive reasoning test; and to identify bias interaction. Many studies have revealed that inductive reasoning is of importance in various contexts. However, previous studies have validated the inductive reasoning test using external factors only, that is, by assessing student answers. Therefore, this study attempts to offer another perspective in assessing the instrument's validity based on rater assessment with many-facet Rasch measurement (MFRM). This study used quantitative analysis with a cross-sectional method. The data were collected using an online form. The inductive reasoning test consisted of 10 figural series (FS) tasks, 10 figural analogies (FA) tasks, 10 number series (NS) tasks, and 10 number analogies (NS) tasks. Seven raters were chosen to rate 40 tasks in the inductive reasoning test using seven-item criteria with three scale categories from 1 (need improvement) to 3 (excellent). Background information of raters like gender, university major, and work experience were collected to check bias interaction. A training session was conducted to converge raters' perceptions in evaluating the inductive reasoning test. The MFRM with FACETS Version 3.83.5 was used to analyze reliability, fit validity, rater measurement, and bias interaction.

For reliability, the results were .97 and .76, respectively, where the inter-rater agreement achieved 57.1% for the exact score and 63.4% for the expected score. For fit validity, all raters achieved with the infit and outfit MNSQ for all items ranging from 0.63 logits to 1.50 logits (Infit and Outfit MNSQ acceptable criteria around 0.5-1.5). Raters had different severity levels ranging from -0.75 logits (rater 7) to 2.74 logits (rater 4) with a 5.68 separation value. Almost all tasks achieved the fit validity criteria with the infit and outfit MNSQ ranging from 0.53 logits to 1.48 logits, except task number 37, indicating that this task was difficult to understand for raters. The item criteria also indicated that the evaluation form with 7 criteria had a well-functioning scale category where the Andrich threshold improved monotonically. The pairs of bias interactions were identified based on rater gender, rater teaching experiences, and rater major. However, no significant bias interaction was detected toward tasks in the inductive reasoning test indicating fair and reliable assessment by raters. In summary, this study has confirmed that the content validity of the inductive reasoning test can be assessed from the raters' perspective, where we can identify the bias interaction that may contaminate the evaluation results. We hope this study can inspire other researchers to perform content validity using MFRM.