World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:12, No:04, 2018

Comparing Test Equating by Item Response Theory and Raw Score Methods with Small Sample Sizes on a Study of the ARTé: Mecenas Learning Game

Authors : Steven W. Carruthers

Abstract : The purpose of the present research is to equate two test forms as part of a study to evaluate the educational effectiveness of the ARTé: Mecenas art history learning game. The researcher applied Item Response Theory (IRT) procedures to calculate item, test, and mean-sigma equating parameters. With the sample size n=134, test parameters indicated “good” model fit but low Test Information Functions and more acute than expected equating parameters. Therefore, the researcher applied equipercentile equating and linear equating to raw scores and compared the equated form parameters and effect sizes from each method. Item scaling in IRT enables the researcher to select a subset of well-discriminating items. The mean-sigma step produces a mean-slope adjustment from the anchor items, which was used to scale the score on the new form (Form R) to the reference form (Form Q) scale. In equipercentile equating, scores are adjusted to align the proportion of scores in each quintile segment. Linear equating produces a mean-slope adjustment, which was applied to all core items on the new form. The study followed a quasi-experimental design with purposeful sampling of students enrolled in a college level art history course (n=134) and counterbalancing design to distribute both forms on the pre- and posttests. The Experimental Group (n=82) was asked to play ARTé: Mecenas online and complete Level 4 of the game within a two-week period; 37 participants completed Level 4. Over the same period, the Control Group (n=52) did not play the game. The researcher examined between group differences from post-test scores on test Form Q and Form R by full-factorial Two-Way ANOVA. The raw score analysis indicated a 1.29% direct effect of form, which was statistically non-significant but may be practically significant. The researcher repeated the between group differences analysis with all three equating methods. For the IRT mean-sigma adjusted scores, form had a direct effect of 8.39%. Mean-sigma equating with a small sample may have resulted in inaccurate equating parameters. Equipercentile equating aligned test means and standard deviations, but resultant skewness and kurtosis worsened compared to raw score parameters. Form had a 3.18% direct effect. Linear equating produced the lowest Form effect, approaching 0%. Using linearly equated scores, the researcher conducted an ANCOVA to examine the effect size in terms of prior knowledge. The between group effect size for the Control Group versus Experimental Group participants who completed the game was 14.39% with a 4.77% effect size attributed to pre-test score. Playing and completing the game increased art history knowledge, and individuals with low prior knowledge tended to gain more from pre- to post test. Ultimately, researchers should approach test equating based on their theoretical stance on Classical Test Theory and IRT and the respective assumptions. Regardless of the approach or method, test equating requires a representative sample of sufficient size. With small sample sizes, the application of a range of equating approaches can expose item and test features for review, inform interpretation, and identify paths for improving instruments for future study.

Keywords : effectiveness, equipercentile equating, IRT, learning games, linear equating, mean-sigma equating
Conference Title : ICGBLSG 2018 : International Conference on Game-Based Learning and Serious Games
Conference Location : Boston, United States
Conference Dates : April 23-24, 2018