Common test items play an important role in equating multiple test forms under the common-item nonequivalent groups design. Inconsistent item parameter estimates among common items can lead to large bias in equated scores for IRT true score equating. Current methods extensively focus on detection and elimination of outlying common items, which usually leads to enlarged random equating error and inadequate content representation of common items.

New robust scale transformation methods based on robust regression, the robust Deming regression method, the robust Haebara method, and the least absolute values (LAV) method, were proposed. In simulation studies, performances of the proposed methods were compared to the Stocking-Lord method which yields the least equating errors among the traditional method and to outlier removal methods. The results indicate: 1) the robust Haebara method and the LAV method usually outperform the robust Deming regression method, 2) the robust Haebara method and the LAV method perform as well as the Stocking-Lord method under the condition of No outlier, 3) the robust Haebara method and the LAV method perform better than the Stocking-Lord method when a single outlying common item is simulated, 4) the LAV method and the robust Haebara method are better than, or at least comparable to, the existing outlier removal methods in the presence of a single outlying common item, and 5) the LAV method and the robust Haebara method have smaller equated scores than the Stocking-Lord method using the CBASE data of English and Mathematics.

Keywords: item response theory, test equating, outlier, robust regression, Deming regression, common item