Population Invariance of Equating Functions

Zhaohui Sheng, Yanyan Sheng
Western Illinois University, Macomb, IL 61455
Southern Illinois University, Carbondale, IL 62901

Abstract

The study investigates equating invariance in situations where some equating samples are considerably small. Using an equivalent groups equating design, the study compares equating functions developed from the linear equating procedure, the unsmoothed equipercentile equating procedure, and the three-parameter IRT true-score equating procedure on examinee samples that were formed based on gender and ethnicity as well as on the total examinee sample. Multiple evaluative measures were used to evaluate equating differences. Findings illustrates that sample sizes have an impact on equating invariance. With samples unequal in size, equating functions developed from the total examinee sample tend to be influenced by the dominating examinee group, whereas equating functions obtained from small examinee samples are subject to larger sampling error. Linear equating has better equating precision for equating samples in size less than 1500 per form based on examination of equating precision.

Key Words: equating, population invariance, resampling method, sample size

1. Introduction

Population invariance requires that equating functions be approximately the same regardless of the examinee subpopulations from which they are developed. It is one of the equating requirements that are regarded as theoretical guidelines for all test equating (Dorans & Holland, 2000; Holland & Dorans, 2006; Peterson, Kolen, & Hoover, 1989). Among the equating requirements, population invariance is considered as the most important and practically useful one (Dorans & Holland, 2000; Holland & Dorans, 2006) to evaluate the equating relationship for score interchangeability. Moreover, gathering evidence that equating relationship is consistent across examinee subpopulations ensures fair assessment and has implications for construct validation (Dorans, 2004).

Population invariance of equating has generally been established in prior studies for large volume tests where it is possible to obtain a sufficient number of examinees even with examinee subpopulations (Angoff & Cowell, 1988; Davier & Wilson, 2004; Dorans & Holland, 2000; Harris & Kolen, 1986; Yang, 2004 etc.). However, for many testing programs, it may not be feasible to obtain sufficiently large equating samples for desired equating precisions especially with some examinee subpopulations. As a result, the equating samples from different examinee subpopulations can differ considerably in size. The purpose of the study is to investigate equating invariance in situations where examinee subpopulation samples vary considerably in size and attempts to explore the impact of examinee subpopulations on the population invariance of equating.

2. Method

Data

Equivalent groups equating design was used to equate Form LQ and Form LO for the College BASE English subject test. As typical with College BASE examinee subpopulations, for both forms, the female examinees contribute to over 77% of the total examinee sample and the white examinees account for 90% of the total examinee sample. Descriptive statistics on the English subtest, including sample size, mean, standard deviation, skewness, kurtosis, Cronbach alpha, and effect size between the two forms were given in Table 1 for the total examinee group as well as for each gender or ethnicity subgroup.

Table 1

Descriptive Statistics for Form LQ and Form LO by Examinee Group

	Examinee Group				
	Total	Male	Female	Nonwhite	White
# of Items	41	41	41	41	41
Form LQ					
Sample Size	6363	1371	4982	574	5741
Mean	24.98	24.89	25.00	22.08	25.26
Standard Deviation	6.43	6.43	6.43	6.48	6.35
Skew	-0.06	-0.07	-0.06	0.27	-0.08
Kurtosis	-0.42	-0.33	-0.44	-0.16	-0.40
α	0.80	0.80	0.80	0.79	0.80
Form LO					
Sample Size	6800	1564	5227	621	6133
Mean	25.39	24.83	25.56	21.23	25.83
Standard Deviation	6.39	6.48	6.35	6.54	6.21
Skew	-0.17	-0.06	-0.20	0.14	-0.17
Kurtosis	-0.26	-0.36	-0.21	-0.05	-0.27
α	0.81	0.81	0.81	0.80	0.80
Effect Size	0.06	0.01	0.09	0.13	0.09

Procedures

Linear, unsmoothed equipercentile, and IRT true-score equating were performed on the total examinee population as well as on the examinee subpopulations defined by gender and ethnicity. Score equivalents resulting from each equating method were compared between the total examinee population and each examinee subpopulation as well as between the examinee subpopulations. Score equivalents were compared across equating methods as well. Multiple evaluative measures were used to evaluate equating differences, including summary indices (Mean Absolute

Difference and Mean Signed Difference), equating difference plots, and empirical standard errors of equating. Investigation of equating sample sizes on equating invariance was evaluated with resampling methods.

3. Results

The summary indices (Table 2) and standardized difference plots (Figure 1) illustrate that substantial equating differences exist between the nonwhite and the white examinees or between the nonwhite and the total examinee samples, that equating differences between the gender groups were close to the one-tenth standard deviation unit – a criterion recommended by Kolen & Brennan (1996, 2004), and that very small equating differences existed between the female and the total examine sample or between the white and the total examinee sample. These results indicate that equating sample sizes do influence invariance of equating relationship. When examinee subpopulations differ considerably in size, equating functions developed from the total examinee sample tend to be influenced by the dominating examinee subpopulation.

Table 2: Weighted and Unweighted Indices for Group Comparisons by Equating Methods

	Linear					
	Gender Groups	Ethnic Groups	Comparison with the Total Group			
	M vs. F	NW vs. W	M vs. T	F vs. T	NW vs. T	W vs. T
Weighted						
MAD$_w$	0.61	1.39	0.46	0.15	1.23	0.16
MSD$_w$	-0.61	-1.39	-0.46	0.15	-1.23	0.16
Unweighted						
MAD$_u$	0.69	1.5	0.52	0.17	1.27	0.47
MSD$_u$	-0.69	-1.5	-0.52	0.17	-1.27	0.23
	Equipercentile					
Weighted						
MAD$_w$	0.64	1.42	0.5	0.15	1.27	0.15
MSD$_w$	-0.6	-1.42	-0.47	0.13	-1.27	0.15
Unweighted						
MAD$_u$	1.11	1.5	0.79	0.33	1.1	0.4
MSD$_u$	0.32	-1.5	0.16	-0.16	-1.1	0.4
	IRT True Score					
Weighted						
MAD$_w$	0.57	1.35	0.44	0.13	1.2	0.15
MSD$_w$	-0.55	-1.35	-0.43	0.13	-1.2	0.15
Unweighted						
MAD$_u$	0.45	0.97	0.36	0.09	0.79	0.18
MSD$_u$	-0.22	-0.97	-0.15	0.06	-0.79	0.18

Note. M=Male, F=Female, NW=Nonwhite, W=White, and T=Total;
MAD = Mean Absolute Difference, and MSD = Mean Signed Difference.

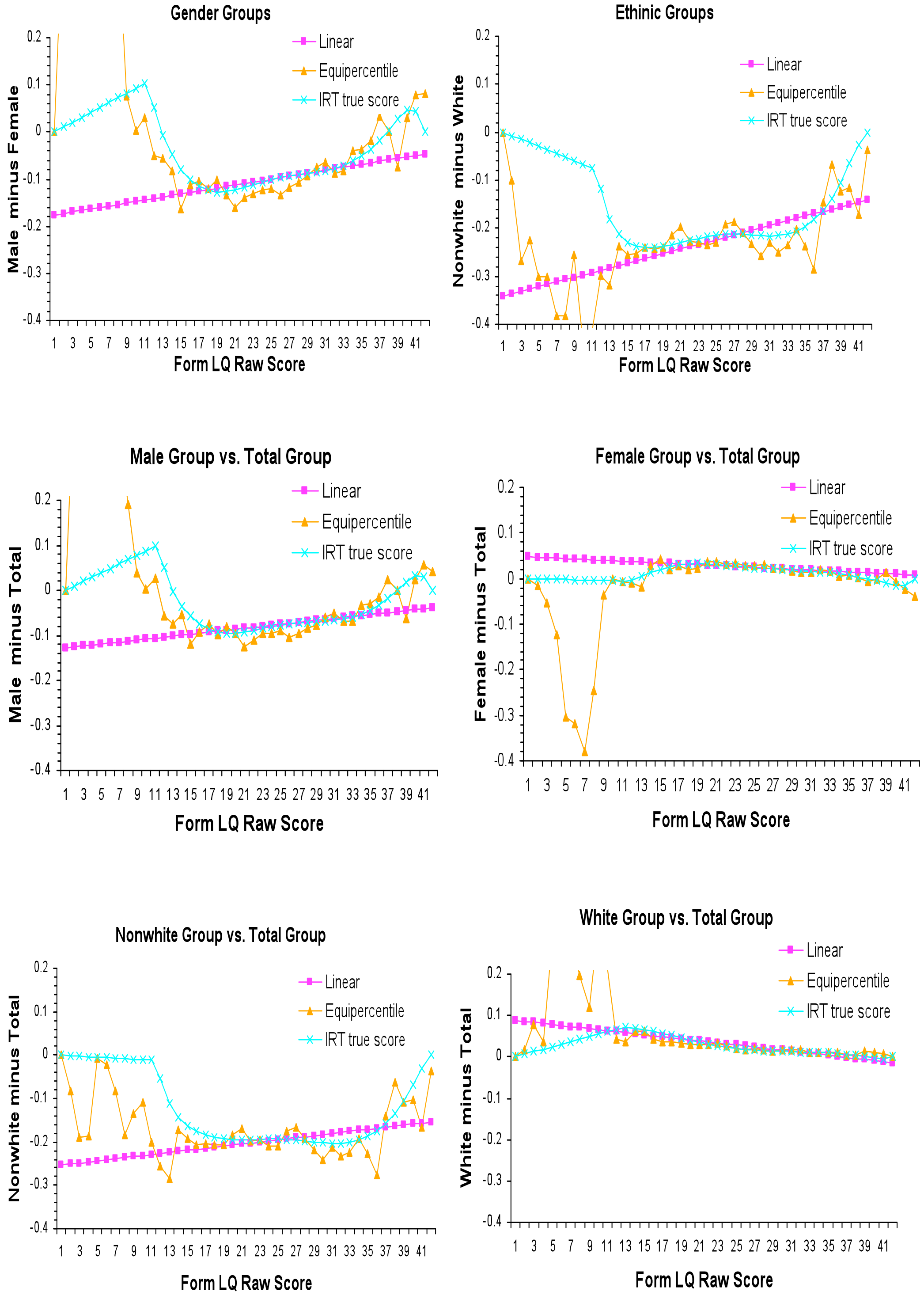


Figure 1: Standardized difference between comparison groups

Further investigations were conducted to assess if large equating differences between equating samples that vary considerably in size were due to sampling variations. Evaluations of equating differences with standard errors of equating indicate that equating functions developed from the nonwhite examinees were substantially different from equating functions developed from the white or from the total examinee sample (Figures 2a & 2b).

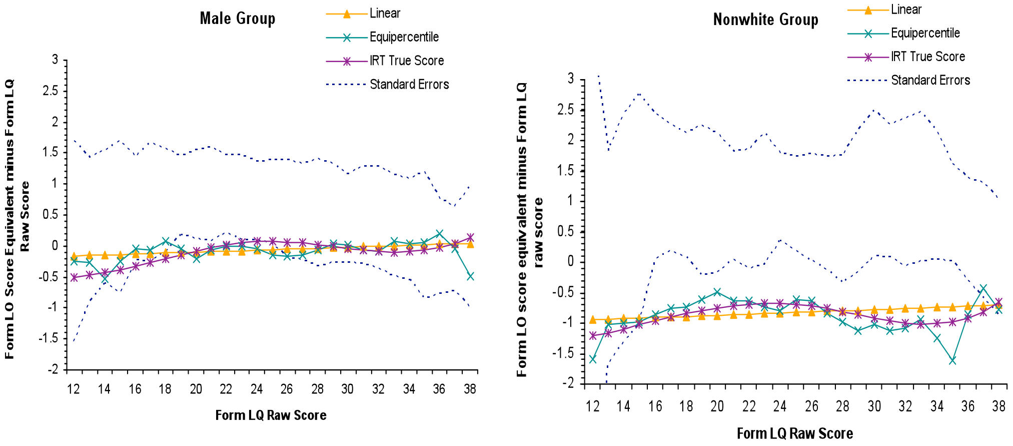


Figure 2a: Standard errors of equating based on examinee subgroup

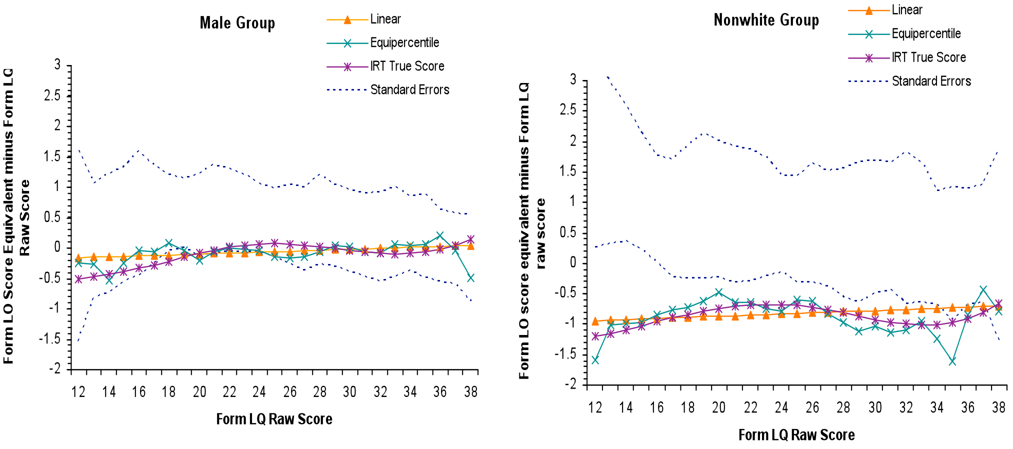


Figure 2b: Standard errors of equating based on the total examinee group

The impact of equating sample sizes on equating invariance was evaluated with resampling methods. The weighted mean absolute difference (MADw) summary statistic, which evaluates the overall equating difference, was examined with replicated samples in size of 100, 300, 500, and 1000 drawn from the total examinee sample. Figure 3 shows that equating functions developed from the examinee subpopulations that are considerably smaller in size are prone to large sampling variations. For equating samples that are similar to the total examinee data used in the study, equating samples in size of at least 500 per form are desired for the overall equating difference in linear equating to be within the one-tenth standard deviation unit.

The Influence of equating sample sizes on equating precision was investigated with bootstrap standard error of equating. Figures 4 illustrates that for equating samples that are sufficiently large, linear and equipercentile equating give similar equating precision along most of the score scale. With smaller equating samples, larger sample sizes are needed for equipercentile equating than for linear equating along all the score scale.

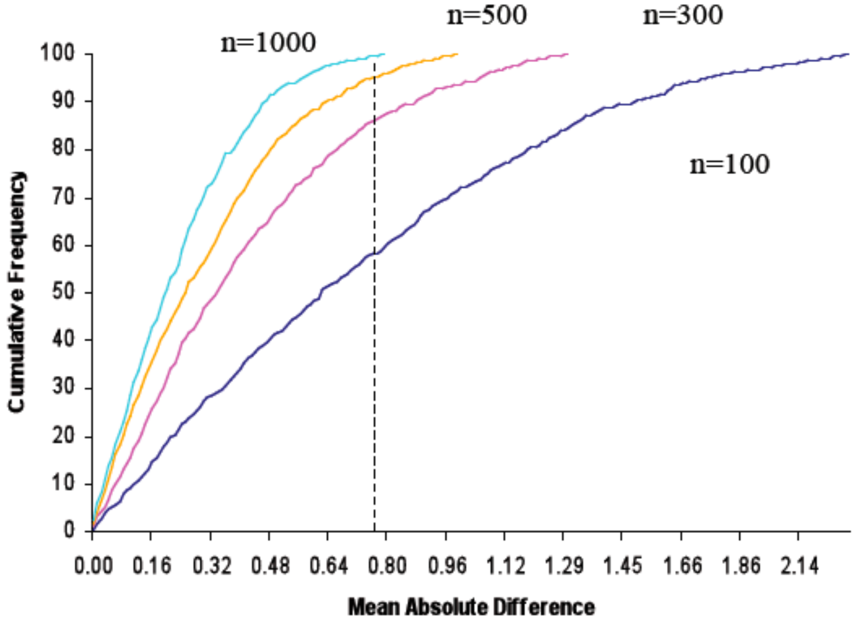
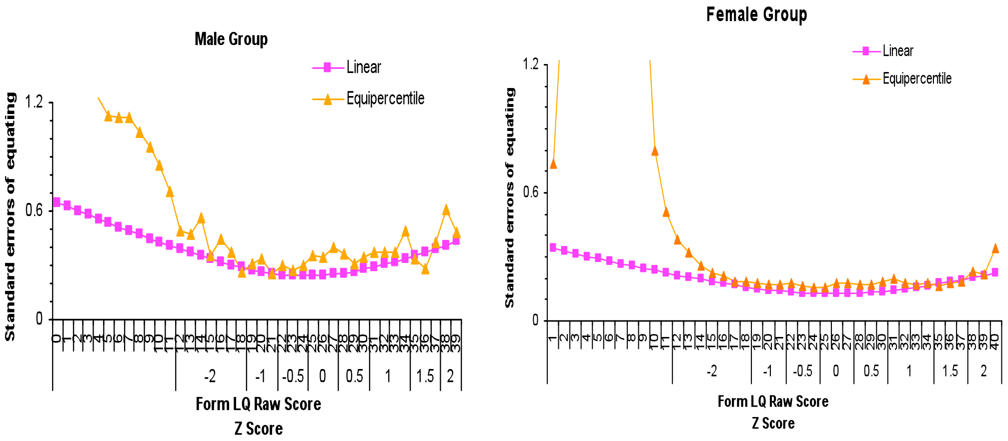


Figure 3: Cumulative frequency of mean absolute difference (MAD$_w$) for four sample sizes



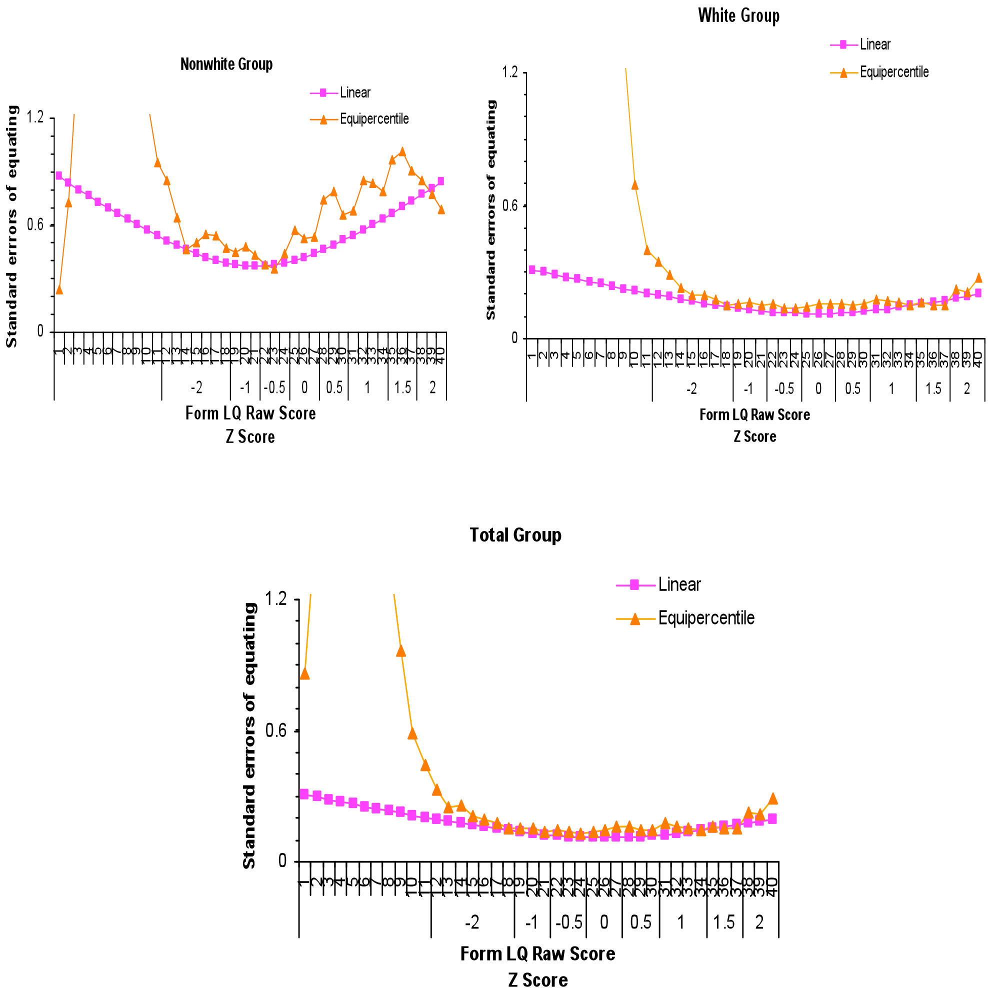


Figure 4: Bootstrap standard errors of equating by equating method and by equating group

No equating procedure is clearly superior to the other equating procedures for assessing equating invariance (Table 2, Figures 2a & 2b). Though IRT true-score equating is developed because of its capability to address large form or group differences, the study findings do not support that IRT true-score equating outperform observed-score equating procedures for test forms similar to those considered in this study.

4. Conclusions

The study illustrates that sample sizes have an impact on equating invariance. With samples unequal in size, equating functions developed from the total examinee sample tend to be influenced by the dominating examinee group, whereas equating functions obtained from small examinee samples are subject to larger sampling error. Linear equating has better equating precision for equating samples in size less than 1500 per form based on examination of equating precision. Equating samples in size of at least 500 per form are desired to achieve equating invariance in linear equating.

The study examined linear, equipercentile, and IRT true-score equating within an equivalent groups equating design. Future studies are encouraged to replicate the study with single group or common-item nonequivalent groups design and with other equating methods, especially the recently developed kernal method of equating, to compare findings.

References

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23,* 327-345.

von Davier, A. A., & Wilson, C. (2004, April). *Population invariance of IRT equating for Advanced Placement (AP®) program exams.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement,41,* 43-68.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement,37,* 281-306.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10,* 35-43.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practices, 7,* 29-36.

Kolen, M. J. (2004a). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41, 3-14.*

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillam.

Yang, W. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement,41,* 33-41.