

# Impact of Non-representative Anchor Items on Scale Stability

Hua Wei

Pearson

Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, May 1-3, 2010



*Using assessment  
and research to  
promote learning*

## **Impact of Non-representative Anchor Items on Scale Stability**

### **Abstract**

Equating is a procedure designed to ensure that test scores obtained on different forms across multiple test administrations can be used interchangeably. By putting scores on the same scale, it is a mechanism to maintain scale stability. However, equating errors, both random and systematic, could result in fluctuations of the scale. Under the common-item nonequivalent groups design, a non-representative common item set is a main factor that could lead to scale drift. The current study investigates the effects of non-representative anchor items on scale stability by using simulated item response data administered across years. Results show that violation of the assumption of representativeness of anchor items results in scale drift.

Keywords: scale drift, anchor items, common-item nonequivalent groups

## Introduction

Common-item nonequivalent groups design is the data collection design most frequently used in large-scale state testing programs. A number of common items are built into the test forms across test administrations to disentangle group differences from form differences. A requirement for the use of this design is that the common items should be representative of the overall test in terms of content and statistical characteristics so that they can reflect group differences adequately (Kolen & Brennan, 2004). Failure to meet this requirement will introduce systematic equating errors and threaten the validity of the equating results.

Equating is a procedure designed to ensure that test scores obtained on different forms across multiple test administrations can be used interchangeably. By putting scores on the same scale, it is a mechanism to maintain scale stability. However, equating errors, both random and systematic, could result in fluctuations of the scale. Haberman & Dorans (2009) summarized the conditions in equating settings that could be the sources of scale drift. Under the common-item nonequivalent groups design, a non-representative common item set is a main factor that could lead to scale drift.

The need to “conduct periodic checks of the stability of the scale on which scores are reported”, as specified in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999, p. 59), has prompted researchers to monitor scale stability in many large-scale assessments. Petersen, Cook, and Stocking (1983) investigated scale drift for the verbal and mathematical portions of the Scholastic Aptitude Test (SAT) by using linear, equipercentile, and item response theory (IRT) equating methods. They defined scale drift as the condition when equating the new form

to the base form directly yields a different equating function from equating the new form to the base form through an equating chain. The amount of scale drift was used to gauge the effectiveness of the three types of equating methods. In a more recent research effort, Liu, Curley, and Low (2009) assessed the stability of the SAT scale from 1994 to 2001 and discovered that both the verbal and math sections exhibited a similar degree of scale drift.

All the existing research (for example, Puhan, 2009; Guo & Wang, 2005) evaluated the extent of scale stability by using empirical data. No simulation study has been carried out to investigate how conditions in the equating design affect the stability of the scale. This study attempts to fill this gap by simulating item response data over multiple administrations under the common-item nonequivalent groups design and examining the effects of non-representative anchor items on scale stability.

## **Method**

### Description of test forms

In this study, four test forms, Form X, Form Y, Form Z, and Form Q are simulated. Form X is administered in Year 1, Form Y in Year 2, Form Z in Year 3, and Form Q in Year 4. Form X is considered as the base form and Form Q as the new form to be equated. Each form consists of 60 unique items and 18 anchor items, which do not contribute to the total score on the test. The 18 anchor items are common across the four forms. Each test form is simulated to cover two content areas: Content Area 1 and Content Area 2. The two content areas are equally represented in the unique item set on

each form. In other words, of the 60 unique items on each form, 30 items assess Content Area 1 and the other 30 assess Content Area 2. All the items are multiple-choice items.

Factors of interest

In this study, the content and statistical characteristics of the anchor items are manipulated to investigate how the degree of representativeness of the anchor items impacts scale stability. The ability distributions of the examinee groups are manipulated to assess whether differences in group ability distributions may lead to scale drift. The correlation between the two content areas is also a factor being studied. The four factors that are of interest in this study are:

a. Correlation between the two content areas

This factor is manipulated at two levels:

Level 1 (High correlation): The correlation between the two content factors is set at .9, which represents a high correlation.

Level 2 (Medium correlation): The correlation between the two content factors is set at .7, which represents a medium correlation.

b. Content representativeness of the anchor items

This factor is manipulated at three levels:

*Level 1* (representative): As in each unique item set, the two content areas are equally represented in the anchor item set. Each content area accounts for half of the items in the anchor item set. That is to say, among the 18 anchor items, 9 items are related to Content Area 1 and the other 9 to Content Area 2.

*Level 2* (partially under-representative): In the anchor item set, one content area is partially under-represented. Content Area 1 accounts for one third of the anchor items and Content Area 2 accounts for the other two thirds. In terms of number of items, among the 18 anchor items, 6 come from Content Area 1 and the other 12 from Content Area 2.

*Level 3* (completely under-representative): In the anchor item set, one content area is not represented at all and the other content area accounts for all the items. All the 18 anchor items assess Content Area 2 exclusively, and no item assesses Content Area 1.

c. Statistical representativeness of the anchor items

This factor is manipulated at two levels:

*Level 1* (representative): The average difficulty of the anchor items is the same as that of the unique items on each test form.

*Level 2* (non-representative): The average difficulty of the anchor items is greater than that of the unique items by .5 on each test form.

d. Group ability distributions

In this study, four examinee groups are involved. Form X is administered to Group 1, Form Y to Group 2, Form Z to Group 3, and Form Q to Group 4. Two conditions are simulated for this factor:

*Condition 1* (equivalent groups): The ability distributions for the four groups are all set as a bivariate normal distribution with a mean vector of  $[0, 0]$  and a variance-covariance matrix that corresponds to the specified correlation between the two dimensions.

*Condition 2* (non-equivalent groups): The ability distributions of the four groups are not equivalent. The ability distributions for the four groups are set to have the same variance-covariance matrix, but different population mean vectors. Specifically, Group 1 has a population mean vector of [0, 0], Group 2 has a mean vector of [.1, .1], Group 3 has a mean of [.2, .2], and Group 4 has a mean of [.3, .3]. A moderate increase of .1 in the mean proficiency between adjacent years is simulated because it mimics the magnitude of increase observed in many large-scale testing programs (e.g., Campbell, Voelkl, & Donahue, 1997).

The four factors are completely crossed to yield a total of 24 conditions. Each condition is replicated 100 times.

#### Generating response data

##### *Model*

As mentioned earlier, the test covers two content areas but each item assesses only one of them. A multidimensional between-item model (Adams, Wilson, & Wang, 1997) is used in generating the response data. It is a Rasch-type model, and it takes on the following form:

$$P(X_{ij} = 1 | \mathbf{\hat{e}}_i) = \frac{\exp(\mathbf{r}'_j \mathbf{\hat{e}}_i + !_j)}{1 + \exp(\mathbf{r}'_j \mathbf{\hat{e}}_i + !_j)}$$

where  $!_j$  is the difficulty of item  $j$ ,  $\mathbf{\hat{e}}_i$  is the multivariate vector of person  $i$ , indicating person  $i$ 's positions on the multiple latent continuous scales, and

$\mathbf{r}_j = (r_{j1}, r_{j2}, r_{j3}, \dots, r_{jm}, \dots, r_{jM})'$  where

$$r_{jm} = \begin{cases} 1, & \text{if item } j \text{ measures to the } m\text{th dimension} \\ 0, & \text{otherwise.} \end{cases}$$

*Ability parameters*

Under each simulated condition, a multivariate normal distribution with a specified mean vector and a variance-covariance matrix is used to generate two sets of  $\theta$ -parameters for each of the four groups. The number of examinees taking each form is 2000. The mean vectors and variance-covariance matrices are provided in Appendix A.

*Item parameters*

Five item sets, including four unique item sets for Form X, Form Y, Form Z, Form Q, and an anchor item set, are generated separately. The  $b$ -parameters for each unique item set are sampled from  $N(0,1)$  in the range between -2.0 to 2.0. When the condition of statistical representativeness is met for the anchor items, the  $b$ -parameters for the anchor item set are sampled from the same distribution with the same constraint. When the condition of statistical representativeness is not met, the  $b$ -parameters for the anchor item set are sampled from  $N(0.5,1)$  in the range from -2.0 to 2.0.

The generated  $b$ -parameters for the anchor items are randomly assigned to one of the two content areas. If the condition of content representativeness is met, the same proportion of items are assigned to the two content areas in the anchor set. If not, different proportions of items from the two content areas are assigned to the anchor item set. For the unique items in each form, the same proportion of items are randomly assigned to the two content areas.

*Generating response data*

Under each simulated condition, by using the multidimensional Rasch model as described earlier, four sets of dichotomous item response data are generated

independently, one for each form. This procedure is replicated 100 times under each condition by refreshing the ability parameters at each replication.

### Parameter estimation

The item and ability parameters for the four forms are estimated separately by using WINSTEPS, Version 3.60.0 (Linacre, 2006). The freely calibrated item parameters for Form X are considered to be on the base scale and the item parameters of all the other three forms need to be equated to the base scale through the anchor items. The item parameter estimates for the anchor items obtained from the separate runs are used to transform the scale of Form Q onto that of Form X directly, and to transform the scale of Form Q onto that of Form X through three intervening transformations, that is, Y to X, Z to Y, and Q to Z. Therefore, for form Q, there are two sets of scale conversion results. The following equation is used in deriving scale scores from the ability estimates:

$$SS = 35 * ! + 600 .$$

Evaluation of scale drift is based on a comparison of the two sets of conversions. All the scale transformations are done by using the Mean/Mean method.

### Evaluation criteria

Under each simulated condition, the amount of scale drift is evaluated by comparing the scale scores obtained by equating Form Q through the equating chain against the scale scores obtained by equating Form Q to Form X directly when the assumption of representativeness of anchor items is satisfied. The root mean squared error (RMSE), as a summary index of scale drift, is calculated through the following formula:

$$RMSE = \sqrt{\frac{1}{2000} \sum f_z [S_i(z) - S_d(z)]^2},$$

where  $S_i(z)$  is the scale score conversion at score point  $z$  obtained by equating Form Q to Form X through the intervening Forms Y and Z,  $S_d(z)$  is the scale score conversion at each score point  $z$  obtained by equating Form Q to Form X directly at the baseline condition, 2000 is the number of examinees taking Form Q, and  $f_z$  is the frequency at score point  $z$ . The average of the RMSEs is taken over the 100 replications.

Bias is another index to detect the direction of scale drift. It is calculated as:

$$BIAS = \frac{1}{2000} \sum f_z [S_i(z) - S_d(z)].$$

The average of the BIASs is taken over the 100 replications for each condition.

### Results

Table 1 presents the average RMSE and BIAS over the 100 replications under each simulated condition when the correlation between the two content areas is equal to .9. As shown in the table, the absolute values of RMSE and BIAS start to increase when the anchor items become under-representative or non-representative. In addition, there seems to be a strong interaction effect between content representativeness and statistical representativeness on both RMSE and BIAS. Similar patterns are observed in Table 2, which presents the average RMSE and BIAS under each simulated condition when the correlation between the two content areas is equal to .7.

Two ANOVA tests are conducted to investigate the main effects and interaction effects of the four factors on RMSE and BIAS. Table 3 provides the results of the ANOVA test on RMSE. As shown in the table, content representativeness, statistical representativeness, and group ability distribution have statistically significant effects on RMSE. Besides, the two-way and three-way interactions among the three factors are all

significant. For the statistically significant main effects, the effect sizes are computed. The computed  $\eta^2$  is equal to .0761 for content representativeness, .0144 for statistical representativeness, and .0029 for group ability distribution. According to Cohen (1988)'s rule of thumb, the cutoff point is set to .0099 for a small effect size, .0588 for a medium effect size, and .1379 for a large effect size. By applying Cohen's rule, the effect size of content representativeness is considered as at the medium level, which means it has a practically meaningful effect on RMSE.

Table 4 provides the results of the ANOVA test on BIAS. As shown in the table, content representativeness, statistical representativeness, and correlation between the two content factors have statistically significant effects on BIAS. The combination of content representativeness and statistical representativeness has a significant interaction effect on BIAS. For the statistically significant main effects, the effect sizes are computed. The computed  $\eta^2$  is equal to .1732 for content representativeness, .0463 for statistical representativeness, and .0024 for correlation between content factors. By applying Cohen's rule, the effect size of content representativeness is considered as large, which means it has a practically meaningful effect on BIAS.

Table 1: Evaluation criteria when the correlation between the two content factors is equal to 0.9

Content Representativeness	Statistical Representativeness	Group Ability Distributions	RMSE	BIAS
representative	representative	equivalent groups	0.011	0.000
		non-equivalent groups	.900	-.144
	non-representative	equivalent groups	3.743	-3.719
		non-equivalent groups	3.858	-3.833
partially under-representative	representative	equivalent groups	3.229	-3.200
		non-equivalent groups	3.183	-3.152
	non-representative	equivalent groups	.867	-.097
		non-equivalent groups	.885	-.038
completely under-representative	representative	equivalent groups	1.179	.791
		non-equivalent groups	1.198	.886
	non-representative	equivalent groups	1.247	-.963
		non-equivalent groups	1.280	-1.006

Table 2: Evaluation criteria when the correlation between the two content factors is equal to 0.7

Content Representativeness	Statistical Representativeness	Group Ability Distributions	RMSE	BIAS
representative	representative	equivalent groups	.005	-.000
		non-equivalent groups	.894	.108
	non-representative	equivalent groups	3.667	-3.643
		non-equivalent groups	3.687	-3.664
partially under-representative	representative	equivalent groups	3.119	-3.087
		non-equivalent groups	3.054	-3.022
	non-representative	equivalent groups	.875	-.012
		non-equivalent groups	.833	.099
completely under-representative	representative	equivalent groups	1.261	.919
		non-equivalent groups	1.241	.864
	non-representative	equivalent groups	1.185	-.943
		non-equivalent groups	1.226	-.973

Table 3: Four-way ANOVA - Effects on RMSE

## Tests of Between-Subjects Effects

Dependent Variable: RMSE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3648.987(a)	23	158.652	332.959	.000	.763
Intercept	7571.950	1	7571.950	15891.095	.000	.870
content	364.781	2	182.391	382.780	.000	.244
statistical	69.312	1	69.312	145.464	.000	.058
group_ability	14.267	1	14.267	29.942	.000	.012
correlation	1.188	1	1.188	2.493	.115	.001
content * statistical	3130.954	2	1565.477	3285.434	.000	.734
content * group_ability	31.667	2	15.833	33.229	.000	.027
statistical * group_ability	9.130	1	9.130	19.161	.000	.008
content * statistical * group_ability	24.772	2	12.386	25.995	.000	.021
content * correlation	.659	2	.330	.692	.501	.001
statistical * correlation	.329	1	.329	.691	.406	.000
content * statistical * correlation	1.566	2	.783	1.643	.194	.001
group_ability * correlation	.176	1	.176	.370	.543	.000
content * group_ability * correlation	.028	2	.014	.029	.971	.000
statistical * group_ability * correlation	.033	1	.033	.069	.793	.000
content * statistical * group_ability * correlation	.124	2	.062	.130	.878	.000
Error	1132.141	2376	.476			
Total	12353.078	2400				
Corrected Total	4781.128	2399				

a R Squared = .763 (Adjusted R Squared = .761)

Table 4: Four-way ANOVA - Effects on BIAS

## Tests of Between-Subjects Effects

Dependent Variable: BIAS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6837.568(a)	23	297.286	373.924	.000	.784
Intercept	3252.144	1	3252.144	4090.530	.000	.633
content	1513.370	2	756.685	951.755	.000	.445
statistical	405.161	1	405.161	509.610	.000	.177
group_ability	.147	1	.147	.185	.667	.000
correlation	4.249	1	4.249	5.345	.021	.002
content * statistical	4907.878	2	2453.939	3086.552	.000	.722
content * group_ability	2.150	2	1.075	1.352	.259	.001
statistical * group_ability	.009	1	.009	.012	.914	.000
content * statistical * group_ability	.608	2	.304	.382	.682	.000
content * correlation	.722	2	.361	.454	.635	.000
statistical * correlation	.145	1	.145	.183	.669	.000
content * statistical * correlation	.096	2	.048	.060	.942	.000
group_ability * correlation	.630	1	.630	.793	.373	.000
content * group_ability * correlation	1.493	2	.747	.939	.391	.001
statistical * group_ability * correlation	.096	1	.096	.120	.729	.000
content * statistical * group_ability * correlation	.812	2	.406	.511	.600	.000
Error	1889.020	2376	.795			
Total	11978.732	2400				
Corrected Total	8726.588	2399				

a R Squared = .784 (Adjusted R Squared = .781)

## Conclusions

This study is intended to investigate how violations of representativeness of anchor items affect the equating results and, consequently, the stability of scale scores across years in the common-item nonequivalent groups design. Stability of scale scores is gauged by the amount of discrepancies in scale score conversions between equating the new form to the base form directly at the baseline condition (when the assumption of representativeness of anchor items is satisfied) and equating the new form to the base form through two intervening forms. Results from the simulations show that the amount of scale drift is greater under the conditions in which the anchor items are only partially representative or not representative at all of the whole test in terms of both content and statistical characteristics. Results of ANOVA tests also show that all the four studied factors have significant independent effects on either or both of the two evaluation indices of scale drift.

Results of the study show that violation of the assumption of representativeness of anchor items has a significant impact on scale drift. Of the two types of representativeness, content representativeness seems to have a stronger effect on scale drift than statistical representativeness, based on the finding that it has medium to large effect sizes on both RMSE and BIAS. Furthermore, by looking at the  $\eta^2$  values provided at Tables 3 and 4, it seems that the interaction between the two types of representativeness accounts for more than 70% of the variance in RMSE and BIAS across all the simulation conditions. These findings all support the conclusion that to maintain scale stability across years it is very important to balance content and adjust for

item difficulty of the anchor set so that it are representative of the whole test in terms of both content and statistical characteristics.

As with any simulation study, caution needs to be applied when generating results from this study to more general scenarios. Findings from this study are limited to the equating design – the common-item nonequivalent groups equating design with an external anchor item set, and other specific conditions, like the large to middle sized correlations between the two dimensions. When the correlation becomes smaller, it is not evident to what degree the content and statistical representativeness factors may affect the equating results and the raw to scale score conversions across years.

In this study, the Rasch model is used in calibrating item response data. This model ignores the discriminating property of items and the guessing behavior of examinees when they respond to items. In a future study, the three-parameter logistic model can be used instead to calibrate items. In addition, the anchor items are considered as external in the study and do not change across years. However, in operational equating practices the internal anchor item design is more frequently used, and the anchor item set is refreshed every year. Additional research on the internal anchor design under violations of the assumptions of anchor items should prove enlightening. Also, as a research effort, performance of the multidimensional model can be studied to allow for a comparison with the unidimensional model under conditions of assumption violation.

### References

- Adams, R. J., Wilson, M., & Wang, W-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Béguin, A. A. (2002). *Robustness of IRT test equating to violations of the representativeness of the common items in a nonequivalent groups design*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Campbell, J. R., Voelkl, K. E., & Donahue, P. L. (1997). *NAEP 1996 trends in academic progress*. Washington, DC: National Center for Educational Statistics.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Revised Edition). Hillsdale, NJ: Erlbaum.
- Guo, F., & Wang, L. (2005). *Evaluating scale stability of a computer adaptive testing system* (GMAC RR-05-12). McLean, VA: Graduate Management Admission Council.
- Haberman, S., & Dorans, N. J. (2009). *Scale consistency, drift, stability: Definitions, distinctions and principles*. Paper presented at the National Council on Measurement in Education, San Diego, CA.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2<sup>nd</sup> ed.). New York, NY: Springer-Verlag.
- Linacre, J. M. (2006). WINSTEPS version 3.60 [Computer Software]. Chicago, IL: Author.
- Liu, J., Curley, E., & Low, A. (2009). *A scale drift study*. Paper presented at the National Council on Measurement in Education, San Diego, CA.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, 8(2), 137-156.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, 22, 79-103.

## Anchor Items and Scale Stability

### Appendix A: Mean Vectors and Variance-Covariance Matrices of the Ability Distributions

$r = 0.9$

	Mean Vector				Variance-Covariance Matrix			
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Equivalent Groups	$\mu = [0,0]$	$\mu = [0,0]$	$\mu = [0,0]$	$\mu = [0,0]$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$
Non-equivalent Groups	$\mu = [0,0]$	$\mu = [0.1,0.1]$	$\mu = [0.2,0.2]$	$\mu = [0.3,0.3]$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$

$r = 0.7$

	Mean Vector				Variance-Covariance Matrix			
	Group 1	Group 2	Group 3	Group 4	Group 1	Group 2	Group 3	Group 4
Equivalent Groups	$\mu = [0,0]$	$\mu = [0,0]$	$\mu = [0,0]$	$\mu = [0,0]$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$
Non-equivalent Groups	$\mu = [0,0]$	$\mu = [0.1,0.1]$	$\mu = [0.2,0.2]$	$\mu = [0.3,0.3]$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$	$\begin{matrix} 1 & \# \\ \rho & 1 \end{matrix}$