Measuring the Power of Learning.™

Research Memorandum
ETS RM–16-05

The Impact of Population Shift on Equating and Differential Item Functioning of Anchor Items: An Empirical Study

Yanxuan Qu

Youhua Wei

Rick Morgan

September 2016

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

The Impact of Population Shift on Equating and Differential Item Functioning of Anchor Items: An Empirical Study

Yanxuan Qu, Youhua Wei, and Rick Morgan
Educational Testing Service, Princeton, New Jersey

September 2016

Corresponding author: Yanxuan Qu, E-mail: yqu@ets.org

Suggested citation: Qu, Y., Wei, Y., & Morgan, R. (2016). T*he impact of population shift on equating and differential item functioning of anchor items: An empirical study* (Research Memorandum No. RM-16-05). Princeton, NJ: Educational Testing Service.

Action Editor: Gautam Puhan

Reviewer: Hongwen Guo

Abstract

This study examined the sensitivity of linking and anchor item differential item functioning (DIF) analyses to varying degrees of population shifts when population invariance did not hold. Five levels of population shift were controlled. The results of linking and DIF under different levels of population shift were compared against the baseline results when no population shift was present. The analyses of this study revealed that the impact of population shift on linking and anchor DIF was small when the degree of population shift was low. However, as the degree of population shift reached beyond a certain level, linking results became substantially different from the baseline and more anchor items were flagged with DIF. The results indicated that linking can be inadequate and actions are needed to improve the quality of linking when population shift is severe.

Key words: population shift, population invariance, differential item functioning, linking, equating, anchor item

An ideal equating relationship has to meet five requirements (Dorans, Pommerich, & Holland, 2007): same construct, equal reliability, symmetry of equating relationship, equity, and population invariance. When population invariance does not hold approximately, rigorous equating or interchangeable scores are not achieved, but weaker forms of linking or comparable scores may still be achieved if the two tests measure the same (or similar) construct with the same (or similar) reliability (Dorans, 2004). In order to ensure similarity of test construct and reliability, the equating sample of the new and reference forms should be as similar as possible and both samples should represent the total test-taking population (Kolen & Brennan, 2004).

Population shift occurs when the equating samples of the new and reference forms differ in subgroup compositions. Population shift may happen to a mature test due to a sudden change in the test user's policy or a sudden natural disaster. It also occurs during the first few years after a new test is launched. In the cases when test takers in one state cannot take a scheduled test like those in other states due to a policy change, an earthquake, or an outbreak of disease, the test-taking population of the new form will be different from the regular test-taking population. When population shift happens, the requirement of having similar samples between new and reference forms for equating is violated. Therefore, for every testing program that conducts equating, it is important to understand how much these shifts in population composition will affect equating so the right practice for equating can be selected when population shift happens.

Few studies can be found in the literature related to the impact of population shift on equating or linking. Liang, Dorans, and Sinharay (2009) and Sinharay, Dorans, and Liang (2011) evaluated the impact of population shift on equating and subgroup differential item functioning (DIF) analysis results. In their study, the degree of population shift was manipulated by the proportion difference of non-English as first language (NEFL) students in new and reference forms. Under the condition with the largest degree of population shift, the percentage of NEFL students was 50% in the new form and 10% in the reference form. In other conditions, the percentage differences for NEFL students between the new and reference forms were smaller than or equal to 5%. Liang et al. and Sinharay et al. used the root mean square deviation (RMSD) index (Dorans & Liu, 2009) to compare the equating results based on subgroups (English as first language [EFL] or NEFL) against the equating results based on total group (EFL and NEFL combined). This method is similar to what is usually done in a population invariance study. When the equating was conducted based on each subgroup (EFL or NEFL), there was no

population shift between new and reference forms. However, when the equating was conducted based on the combined group, population shift was evident between the new and reference forms because the proportion of NEFL students was different between the new and reference forms. Liang et al. and Sinharay et al. found that equating results based on subgroups and total groups were similar and therefore the impact of population shift in the proportion of NEFL students was negligible on equating. Thus, it did not matter whether equating was done based on the combined total group or based on EFL-only test takers. They also compared subgroup DIF results, such as gender and ethnicity DIF, based on total group and EFL-only group results, respectively. They found that DIF analysis was relatively more sensitive to population shift than equating.

The purpose of this study is to examine the potential impact of different degrees of population shift on both linking and anchor DIF results. Anchor DIF in this context refers to examining differential performance for the anchor items (or common items) in new and reference form groups using DIF methods. In a previous study, Qu, Costanzo, and Stuart (2012) compared the linking results based on subgroups (Group A and Group B with different ethnicity and geographic background) and the linking results based on total group (Qu et al., 2012). They found that the conversion lines based on each subgroup were different from the conversion line based on total group. Therefore population invariance does not hold between Group A and Group B even when the new and reference groups have similar compositions (original data in Table 1). This current study used the same data that were used in Qu et al. and manipulated the degree of population shift in order to investigate how much population shift may call for changes to the existing linking practice.

Because population invariance did not hold approximately in our data, the word "linking," instead of equating, is used throughout this paper when referring to the current study.

Methods

Data

Real data from three randomly selected test forms were used for this empirical study. Each test form has two sections with different contents measuring different constructs. Each section has its own linking using the nonequivalent group anchor test (NEAT) design. Above chance level, Section 1 has a raw score range from 65 to 100, and Section 2 has a raw score range from 63 to 100. Each form was linked back to only one reference form with about 32% common items. The test-taking population consists of examinees from two major subgroups

(Group A and Group B) with different ethnicity and geographic background. Operationally, the Group A and Group B combined group in a new form is linked back to the combined group in the corresponding reference form. The top panel in Table 1 presents the sample sizes of Group A, Group B, and total group, taking each of the new and reference forms in operation. New forms (NF) are denoted as NF1, NF2, and NF3, and their corresponding reference forms (RF) are denoted as RF1, RF2, and RF3. The degree of population shift was small in this original set of data: The proportions of Group A were 18% in NF1 and 14% in RF1, 13% in NF2 and 16% in RF2, and 31% in NF3 and 39% in RF3.

Table 1. Sample Sizes of New and Reference Form Groups Under Each Condition of Population Shift

New form	Total	Group A	Group B	% Group A	Reference form	Total	Group A	Group B	% Group A
Original data									
NF1	4264	751	3513	18	RF1	7353	988	6365	13
NF2	5660	817	4843	14	RF2	6338	1011	5327	16
NF3	3325	1034	2291	31	RF3	3160	1245	1915	39
Condition 1									
NF1	4264	853	3411	20	RF1	7353	1471	5882	20
NF2	5660	1132	4528	20	RF2	6338	1268	5070	20
NF3	3325	665	2660	20	RF3	3160	632	2528	20
Condition 2									
NF1	4264	1279	2985	30	RF1	7353	1471	5882	20
NF2	5660	1698	3962	30	RF2	6338	1268	5070	20
NF3	3325	998	2327	30	RF3	3160	632	2528	20
Condition 3									
NF1	4264	1706	2558	40	RF1	7353	1471	5882	20
NF2	5660	2264	3396	40	RF2	6338	1268	5070	20
NF3	3325	1330	1995	40	RF3	3160	632	2528	20
Condition 4									
NF1	4264	2132	2132	50	RF1	7353	1471	5882	20
NF2	5660	2830	2830	50	RF2	6338	1268	5070	20
NF3	3325	1663	1662	50	RF3	3160	632	2528	20
Condition 5									
NF1	4264	0	4264	0	RF1	7353	7353	0	100
NF2	5660	0	5660	0	RF2	6338	6338	0	100
NF3	3325	0	1662	0	RF3	3160	3160	0	100

Note. NF = new form; RF = reference form. NF1, NF2, and NF3 are three new forms. Their corresponding reference forms are denoted as RF1, RF2, and RF3, respectively.

Study Design

To study the impact of population shift on linking, the differences in the proportion of Group A test takers between the NF and RF groups were controlled at five levels (see Condition 1 through Condition 5 in Table 1):

1. Both NF and RF groups have 20% of Group A test takers; that is, the proportion difference is 0%.
2. NF group has 30% of Group A, and RF group has 20% of Group A; that is, the proportion difference is 10%.
3. NF group has 40% of Group A, and RF group has 20% of Group A; that is, the proportion difference is 20%.
4. NF group has 50% of Group A, and RF group has 20% of Group A; that is, the proportion difference is 30%.
5. NF group has 0% of Group A, and RF group has 100% of Group A; that is, the proportion difference is 100%.

Condition 1 had no population shift, as the proportion of Group A test takers was the same between new and reference groups. Condition 2 had the minimum level of population shift; that is, the proportion of Group A examinees increased from 20% in the reference administration to 30% in the new administration. Condition 3 had a population shift from 20% to 40%. Condition 4 had a population shift from 20% to 50%. Condition 5 was designed to have the largest population shift (100% difference in the proportion of Group A examines between new and reference groups).

In each linking, the sample size of the total group was set to be as same as that in the original data (Table 1) in order to roughly control the random equating errors related to sampling. The numbers of test takers in Group A and Group B were then determined by the proportions in each of the five conditions. As seen in Tables 1 to 3, the sample size of each subgroup can be larger than that in the original data. For example, the number of test takers from Group A in NF1 under condition one was 853, larger than the number of examinees from Group A (N = 751) in the original data. This larger sample of Group A was created from the original data set by random sampling with replacement.

Table 2. Basic Statistics for New and Reference Forms: Section 1 Original Data

		New form						Reference Form					
		Total			Anchor			Total			Anchor		
Form	Statistic	A + B	A	B	A + B	A	B	A + B	A	B	A + B	A	B
	N	4264	751	3513	4264	751	3513	7353	988	6365	7353	988	6365
1	Mean	37.58	38.01	37.48	11.94	11.73	11.98	34.81	36.06	34.61	11.70	11.60	11.71
	SD	6.98	6.58	7.06	2.86	2.86	2.86	8.12	7.47	8.20	3.01	2.97	3.01
	N	5660	817	4843	5660	817	4843	6338	1011	5327	6338	1011	5327
2	Mean	36.87	37.12	36.83	12.42	12.51	12.41	39.51	38.61	39.68	12.26	12.35	12.24
	SD	6.94	6.08	7.08	2.51	2.13	2.57	5.94	4.94	6.10	2.44	1.99	2.52
	N	3325	1034	2291	3325	1034	2291	3160	1245	1915	3162	1245	1917
3	Mean	35.04	32.95	35.99	10.12	9.00	10.62	34.30	33.41	34.88	9.58	8.94	10.00
	SD	7.97	7.11	8.16	2.86	2.76	2.76	6.66	6.27	6.85	2.82	2.70	2.82

Table 3. Basic Statistics for New and Reference Forms: Section 2 Original Data

		New form						Reference form					
		Total			Anchor			Total			Anchor		
Form	Statistic	A + B	A	B	A + B	A	B	A + B	A	B	A + B	A	B
	N	4264	751	3513	4264	751	3513	7353	988	6365	7353	988	6365
1	Mean	29.56	34.27	28.56	9.45	11.00	9.12	27.98	33.34	27.14	9.15	10.76	8.90
	SD	7.80	7.65	7.46	3.23	2.97	3.19	9.15	8.66	8.94	3.32	3.00	3.30
	N	5660	817	4843	5660	817	4843	6338	1011	5327	6338	1011	5327
2	Mean	26.47	30.70	25.76	9.01	10.28	8.80	27.42	31.74	26.60	8.94	10.45	8.66
	SD	8.03	7.59	7.88	3.02	2.64	3.03	8.41	7.56	8.32	3.02	2.42	3.04
	N	3325	1034	2291	3325	1034	2291	3160	1245	1915	3162	1245	1917
3	Mean	27.69	28.75	27.21	8.99	9.35	8.82	29.91	32.01	28.54	8.69	9.31	8.29
	SD	8.42	7.88	8.61	2.62	2.53	2.64	7.74	7.41	7.64	2.72	2.65	2.68

Equating Methods

Two observed score nonlinear linking methods commonly used by testing programs were implemented in this study: post-stratification equipercentile (PSE) and chained equipercentile (CE) methods. Loglinear presmoothing with five moments reserved for the two marginal distributions and one cross-moment reserved was conducted for both new form and old form data before each equating. The five-moment loglinear model is reckoned as a sufficient model given the data we have.

Evaluation Criterion

The raw-to-scale linking results from Condition 1 (i.e., 20% of Group A examinees in both NF and RF groups) were used as a baseline for comparing linking results. Because population invariance does not hold (Qu et al., 2012), there is no true equating or ideal equating here. However, the quality of linking based on Condition 1 is considered relatively better than the other conditions. When the composition of the NF and RF groups is similar in terms of Group A and Group B, it is more likely to have similar linking samples between new and reference forms, and more likely to have similar construct and reliability. Linking differences larger than one difference that matters (DTM) in absolute values were considered as substantially large (Dorans & Feigenbaum, 1994). Because the scaled scores for this test increased in two-point units, one DTM equals one scale score point.

Differential Item Functioning Methods

MH D-DIF (Holland & Thayer, 1985) and STD P-DIF (Dorans & Holland, 1992) analyses were run on anchor items between NF and RF linking samples. The matching criterion for both DIF statistics is the sum of all the anchor item scores. Given that population invariance does not hold well enough between Group A and Group B, when population shift appears between NF and RF groups, anchor items may not perform the same in NF and RF groups due to the dissimilar group compositions. Our speculation is that more anchor items will tend to have MH D-DIF larger than or equal to 1.5, or STD P-DIF larger than or equal to 0.1 in the DIF analyses on anchor items between NF and RF linking samples.

Holland and Thayer (1985) defined MH D-DIF as,

$$MH-D-DIF=-2.35\ln\left(\frac{\sum_s\left(\frac{C_{Rs}I_{Fs}}{N_{Total,s}}\right)}{\sum_s\left(\frac{I_{Rs}C_{Fs}}{N_{Total,s}}\right)}\right)$$

where C_{Rs} is the number of reference group test takers at score level *s* who answered the studied item correctly. I_{Fs} is the number of focal group test takers at score level *s* who answered the studied item incorrectly. $N_{Total,s}$ is the number of total group test takers at score level *s*.

Dorans and Holland (1992) defined STD P-DIF as a weighted sum of conditional proportion correct differences:

$$STD\ P-DIF=\sum_{s}\left(\frac{N_{Ts}}{\sum_{s}N_{Ts}}\right)\left(\frac{R_{Fs}}{N_{Fs}}-\frac{R_{Rs}}{N_{Rs}}\right)$$

where, R_{Fs} is the number of focal group test takers at score level *s* who answered the studied item right. N_{Fs} is the total number of focal group test takers at score level *s*. R_{Rs} is the number of reference group test takers at score level *s* who answered the studied item right. N_{Rs} is the total number of reference group test takers at score level *s*. N_{Ts} is the number of total group test takers at score level *s*.

Results

Linking Results

Tables 4 and 5 present standardized group ability differences and variance ratios calculated based on anchor test statistics under each condition. Except for Condition 5, group differences were relatively small and variance ratios were close to 1.

Table 4. Standardized Group Mean Difference and Variance Ratio Based on Anchor Items: Section 1

Form	Original data	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
NF1-RF1	0.08 (0.90)	0.08 (0.91)	0.08 (0.91)	0.06 (0.90)	0.05 (0.92)	0.15 (0.92)
NF2–RF2	0.07 (1.06)	0.07 (1.04)	0.08 (1.03)	0.09 (0.98)	0.07 (0.94)	0.03 (1.61)
NF3-RF3	0.19 (1.03)	0.16 (1.03)	0.1 (1.04)	0.05 (1.06)	-0.02 (1.07)	0.62 (1.07)

Note. NF = new form; RF = reference form. Variance ratios are noted in parenthesis.

Table 5. Standardized Group Mean Difference and Variance Ratio Based on Anchor Items: Section 2

Form	Original data	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
NF1-RF1	0.09 (0.95)	0.07 (0.94)	0.14 (0.95)	0.18 (0.95)	0.24 (0.94)	-0.52 (1.18)
NF2–RF2	0.02 (1)	0.03 (1)	0.08 (0.98)	0.13 (0.96)	0.15 (0.96)	-0.6 (1.54)
NF3-RF3	0.11 (0.93)	0.16 (0.95)	0.18 (0.95)	0.21 (0.93)	0.21 (0.95)	-0.17 (1.00)

Note. NF = new form; RF = reference form. Variance ratios are noted in parenthesis.

Tables 6 and 7 show the anchor total correlations in the new and reference forms for each condition in Sections 1 and 2.

Table 6. Anchor-Total Correlations in New Forms (NF) and Reference Forms (RF) in Each Condition: Section 1

Condition	NF1	RF1	NF2	RF2	NF3	RF3
Original	0.91	0.9	0.87	0.87	0.89	0.87
Condition 1	0.91	0.9	0.87	0.87	0.89	0.87
Condition 2	0.91	0.9	0.86	0.87	0.89	0.87
Condition 3	0.90	0.9	0.86	0.87	0.88	0.87
Condition 4	0.90	0.9	0.86	0.87	0.88	0.87
Condition 5	0.91	0.9	0.88	0.81	0.90	0.86

Table 7. Anchor-Total Correlations in New Forms (NF) and Reference Forms (RF) in Each Condition: Section 2

Condition	NF1	RF1	NF2	RF2	NF3	RF3
Original	0.9	0.91	0.88	0.87	0.85	0.86
Condition 1	0.90	0.91	0.88	0.87	0.85	0.86
Condition 2	0.90	0.91	0.87	0.87	0.85	0.86
Condition 3	0.90	0.91	0.88	0.87	0.85	0.86
Condition 4	0.90	0.91	0.87	0.87	0.85	0.86
Condition 5	0.90	0.90	0.88	0.81	0.86	0.85

Figures 1 through 6 present the raw-to-scale linking results based on the PSE method. Each figure depicts the linking difference of each linking conditioned at each raw score point.

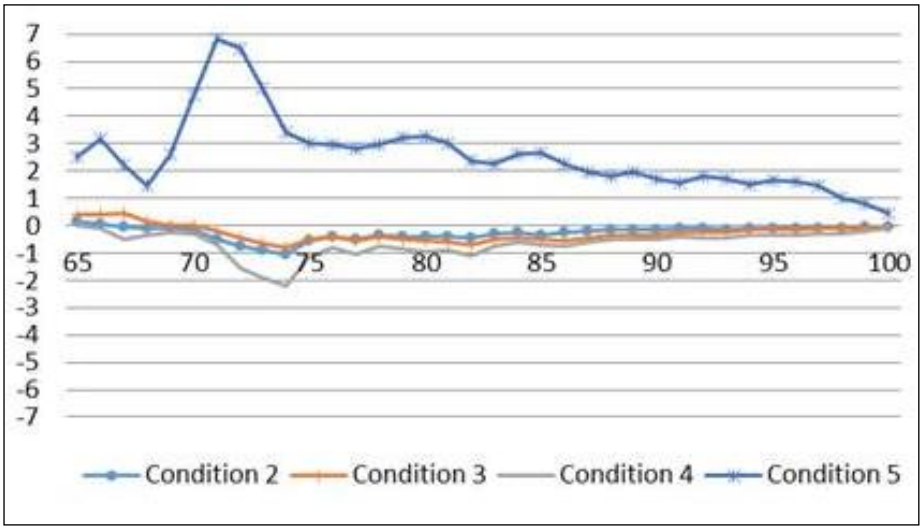

Figure 1. Above chance level PSE linking differences for Form 1, Section 1.

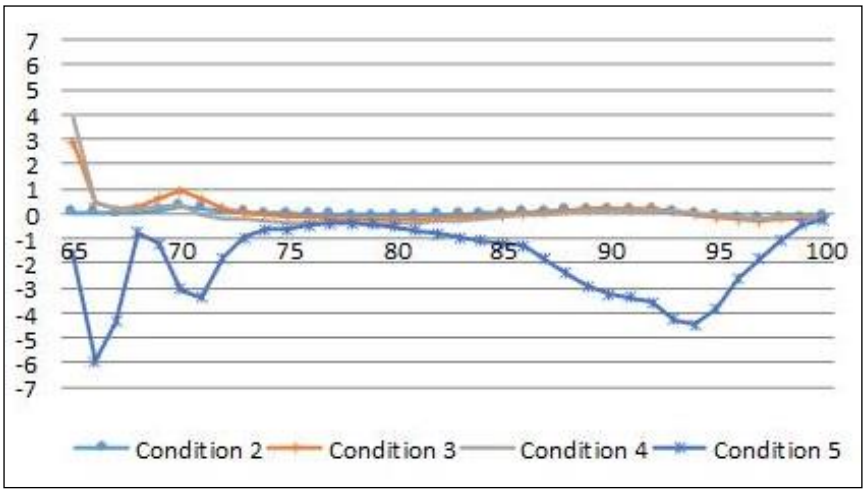

Figure 2. Above chance level PSE linking differences for Form 2, Section 1.

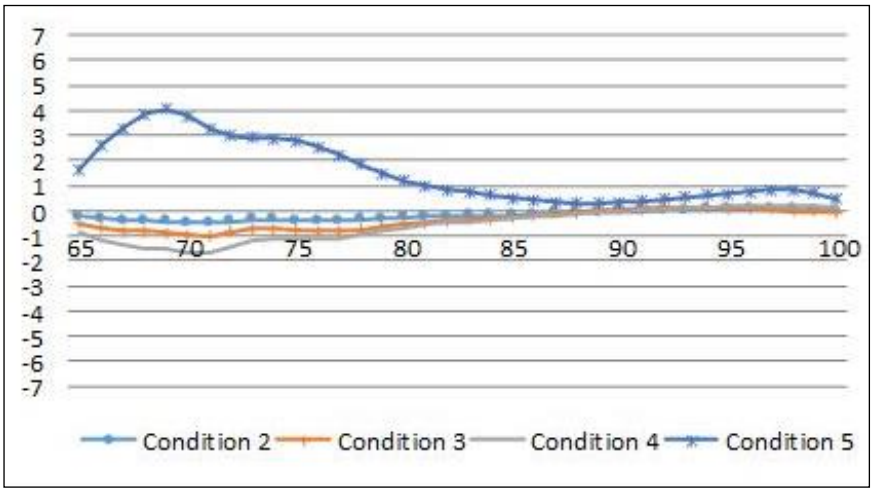

Figure 3. Above chance level PSE linking differences for Form 3, Section 1.

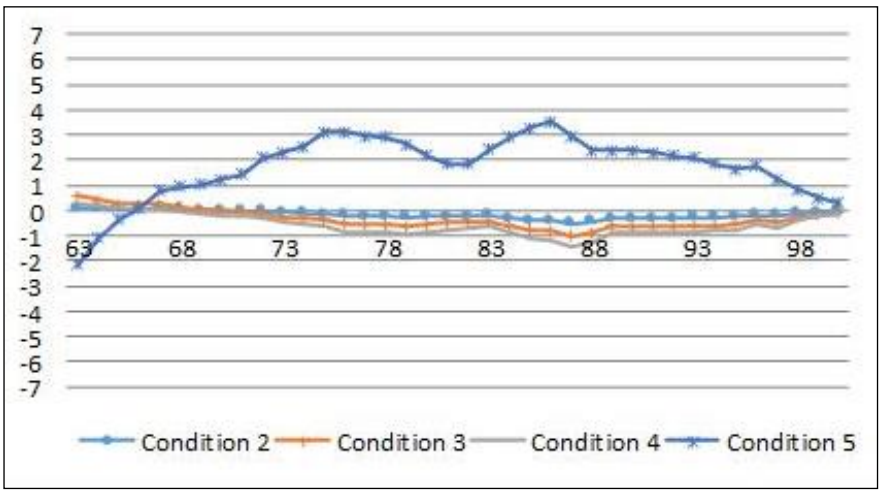

Figure 4. Above chance level PSE linking differences for Form 1, Section 2.

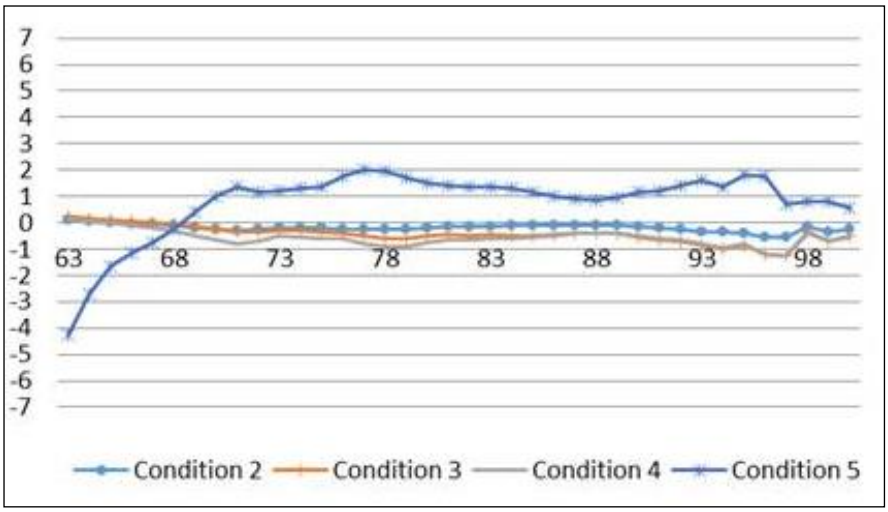

Figure 5. Above chance level PSE linking differences for Form 2, Section 2.

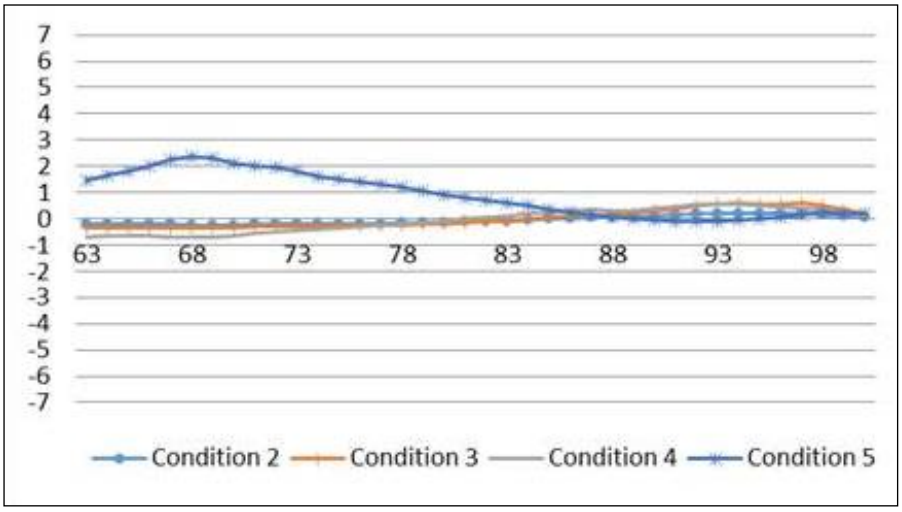

Figure 6. Above chance level PSE linking differences for Form 3, Section 2.

Tables 8 and 9 present the summary statistics of the conditional linking differences. As the new and reference linking samples became more different, the linking differences got larger. This pattern was observed in both test sections. Condition 5, with all examinees in the reference group from Group A, had the largest shift in population composition and correspondingly the largest linking difference. In Condition 5, almost all of the score ranges above chance level (from 65 to 100 for Section 1 and from 63 to 100 for Section 2) had linking differences larger than one DTM. Condition 4 (i.e., the new group had 30% more Group A examinees than the reference group) had linking differences larger than one DTM at a few raw score points above chance level (see Figures 1, 3, 4, and 5).

Table 8. Summary Statistics of Poststratification Equipercentile Linking Differences for Score Points Above Chance: Section 1

Form	Statistic	Condition 2	Condition 3	Condition 4	Condition 5
NF1	Avg	-0.27	-0.27	-0.66	2.59
	Min	-1.03	-0.83	-2.18	0.45
	Max	0.15	0.46	0.01	6.85
NF2	Avg	0.02	0.12	0.03	-1.91
	Min	-0.20	-0.29	-0.39	-5.96
	Max	0.29	2.87	3.84	-0.21
NF3	Avg	-0.20	-0.38	-0.57	1.53
	Min	-0.48	-0.98	-1.69	0.29
	Max	0.06	0.14	0.23	4.07

Note. NF = new form.

Table 9. Summary Statistics of Poststratification Equipercentile Linking Differences for Score Points Above Chance: Section 2

Form	Statistic	Condition 2	Condition 3	Condition 4	Condition 5
NF1	Avg	-0.18	-0.39	-0.62	1.94
	Min	-0.50	-0.99	-1.43	-0.34
	Max	0.08	0.28	0.10	3.52
NF2	Avg	-0.21	-0.48	-0.61	1.01
	Min	-0.57	-1.27	-1.28	-1.66
	Max	0.02	0.12	-0.01	1.99
NF3	Avg	-0.05	0.02	-0.03	0.87
	Min	-0.24	-0.37	-0.71	-0.09
	Max	0.21	0.60	0.55	2.37

Note. NF = new form.

Condition 2 and Condition 3 (i.e., the new group had at most 20% more Group A examinees than the reference group) had very similar amounts of linking differences, with almost all of the score ranges above chance level, having linking differences smaller than one DTM. Similar findings are observed based on the results from the CE method (Tables 10 and 11, and Figures 7 through 12).

Table 10. Summary Statistics of Chained Equipercentile Linking Differences for Score Points Above Chance: Section 1

Form	Statistic	Condition 2	Condition 3	Condition 4	Condition 5
NF1	Avg	-0.30	-0.28	-0.71	2.59
	Min	-1.03	-0.69	-2.06	0.24
	Max	0.01	0.61	-0.08	6.53
NF2	Avg	0.02	0.13	0.02	-2.00
	Min	-0.25	-0.36	-0.43	-6.53
	Max	0.29	2.61	4.12	-0.17
NF3	Avg	-0.24	-0.50	-0.71	1.84
	Min	-0.50	-1.16	-1.74	0.48
	Max	0.03	0.00	0.14	4.05

Note. NF = new form.

Table 11. Summary Statistics of Chained Equipercentile Linking Differences for Score Points Above Chance: Section 2

Form	Statistic	Condition 2	Condition 3	Condition 4	Condition 5
NF1	Avg	-0.09	-0.23	-0.42	1.00
	Min	-0.37	-0.85	-1.19	-1.88
	Max	0.15	0.56	0.04	2.70
NF2	Avg	-0.14	-0.35	-0.40	0.06
	Min	-0.55	-1.28	-1.22	-4.91
	Max	0.15	0.30	0.27	1.94
NF3	Avg	-0.04	0.02	0.04	0.51
	Min	-0.27	-0.53	-0.70	-0.24
	Max	0.23	0.59	0.63	1.88

Note. NF = new form.

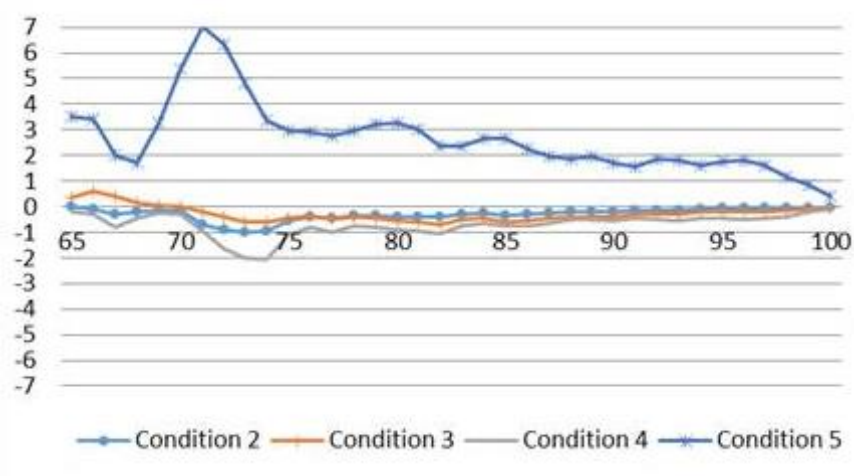

Figure 7. Above chance level CE linking differences for Form 1, Section 1.

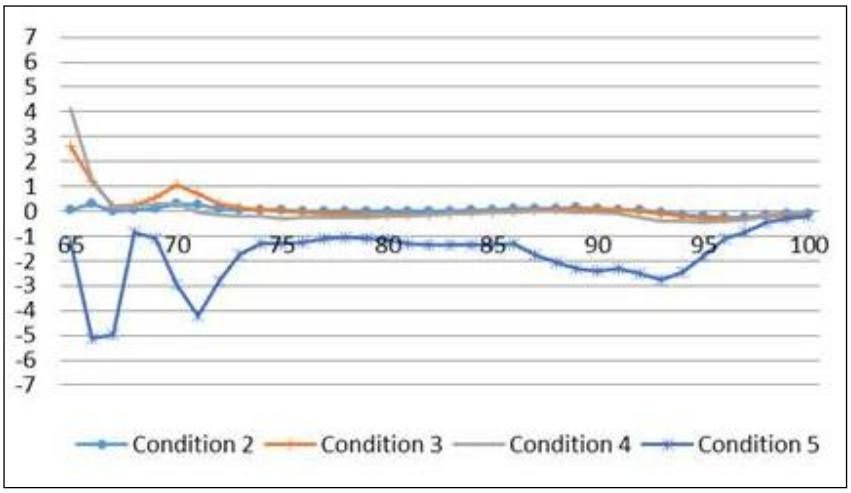

Figure 8. Above chance level CE linking differences for Form 2, Section 1.

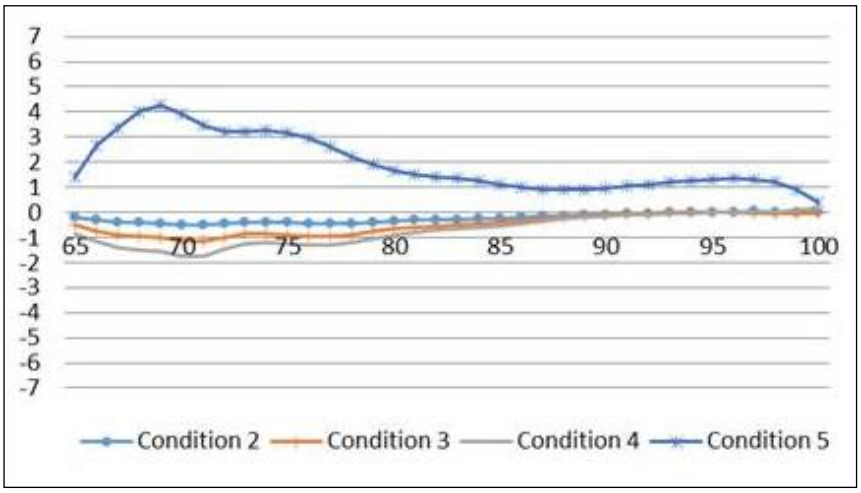

Figure 9. Above chance level CE linking differences for Form 3, Section 1.

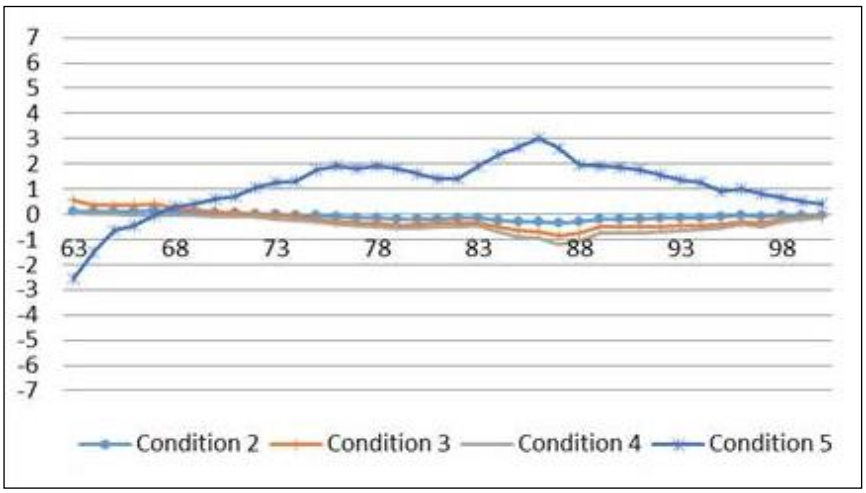

Figure 10. Above chance level CE linking differences for Form 1, Section 2.

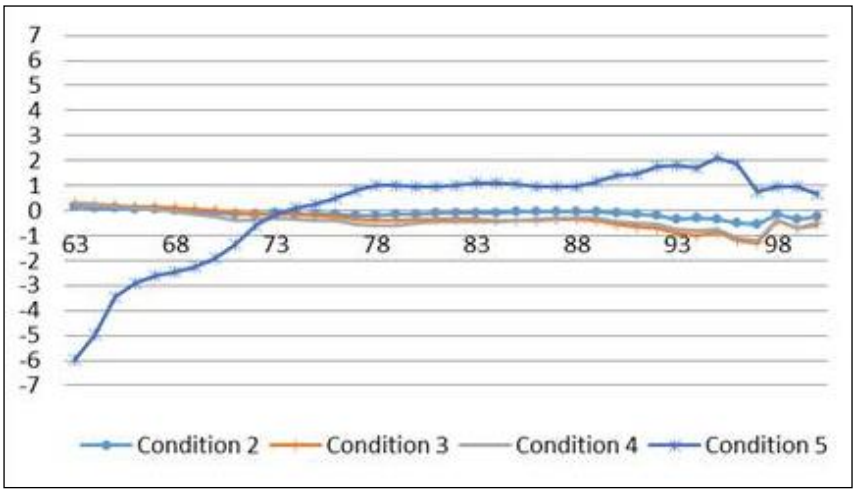

Figure 11. Above chance level CE linking differences for Form 2, Section 2.

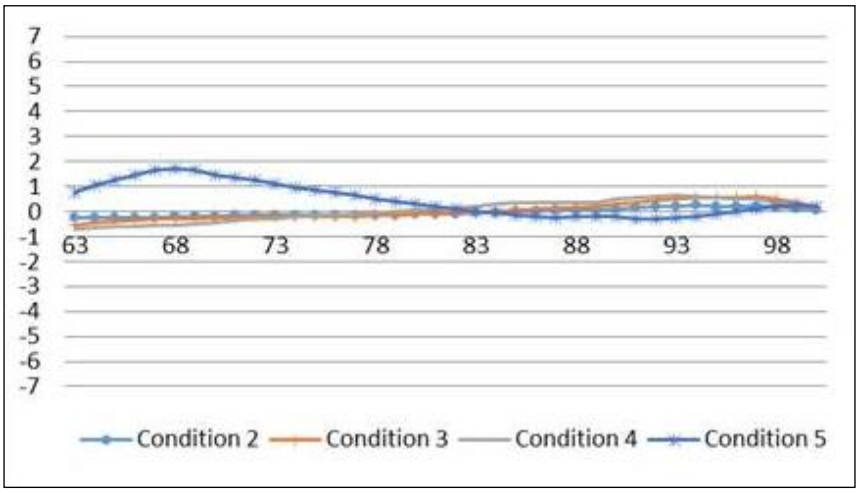

Figure 12. Above chance level CE linking differences for Form 3, Section 2.

DIF Results

Tables 12 and 13 present the number of anchor items with MH D-DIF greater than or equal to 1.5 or STD P-DIF larger than 0.1 under each condition of resampled data for each linking. Conditions 4 and 5 (i.e., when the degree of population shift was larger than or equal to 30%) tend to have more DIF flags than other conditions when population shift is less severe. The STD P-DIF did not flag any items when the degree of population shift was smaller than 30%, with only one exception in Condition 1 of Section 1 for NF3.

Table 12. Number of Anchor Items With Differential Item Functioning (DIF) Flags in Each Equating Condition: Section 1

Form	Conditions	Number of anchor items with MH D-DIF ≥ 1.5	Number of anchor items with STD P-DIF > 0.1
NF1 (total # of anchor items = 16)	1	1	0
	2	1	0
	3	2	0
	4	3	1
	5	8	6
NF2 (total # of anchor items = 16)	1	0	0
	2	0	0
	3	0	0
	4	0	1
	5	4	2
NF3 (total # of anchor items = 15)	1	0	1
	2	0	0
	3	0	0
	4	0	0
	5	5	5

Note. NF = new form.

Table 13. Number of Anchor Items With Differential Item Functioning (DIF) Flags in Each Equating Condition: Section 2

Form	Conditions	Number of anchor items with MH D-DIF ≥1.5	Number of anchor items with STD P-DIF > 0.1
NF1 (total # of anchor items = 16)	1	0	0
	2	0	0
	3	0	0
	4	0	1
	5	4	4
NF2 (total # of anchor items = 15)	1	0	0
	2	0	0
	3	0	0
	4	0	1
	5	5	5
NF3 (total # of anchor items = 14)	1	0	0
	2	0	0
	3	0	0
	4	0	0
	5	3	5

Note. NF = new form.

Discussions and Conclusions

Similar to Liang et al. (2009) and Sinharay et al. (2011), the term population shift was used in this study to describe any differences in subgroup compositions, not the differences in group abilities, between NFs and RFs. As seen in Tables 4 and 5, when the degree of population shift increased from Condition 1 to Condition 5, ability differences between NF and RF groups did not necessarily increase. Population shift and population invariance are two independent concepts. In the case of Sinharay et al. (2011), population invariance was observed no matter how big population shift was. But in the case of Qu et al. (2012), population invariance was not observed even when population shift was very small.

Different from Liang et al. (2009) and Sinharay et al. (2011), our study examined the sensitivity of linking and DIF analyses to varying degrees of population shifts when population invariance did not hold. In addition, our design and criterion for evaluating linking were different from theirs. The way they evaluated equating was like a population invariance study, where conversion lines based on subgroups were compared with the conversion line based on total group. Our study compared the linking results from different degrees of population shifts with the linking result from no population shift. Lastly, our study examined DIF of anchor items instead of DIF by gender or ethnicity groups for total or EFL-only test takers.

Based on our results, we can see that population shift does have an impact on the results of linking and DIF. The more the new and reference groups differ in population composition, the bigger the linking differences are. When group proportion difference reaches 30% or more (i.e., Condition 4 or Condition 5), not only do the linking differences become larger than one DTM, but also the anchor items are more likely to have MH D-DIF larger than or equal to 1.5 or STD P-DIF larger than 0.1. This pattern is more distinct with the STD P-DIF than the MH D-DIF. Items with significant DIF values suggest that they have different performance (or different difficulty levels) between new and reference groups. When population invariance does not hold and the sample compositions differ between new and reference forms, some unknown group differences (that are not intended to be measured by the test) may have affected anchor item performance in the two groups differently. We know that in equating or linking with a NEAT design, the performance of anchor items plays an important role. When anchor items do not perform similarly between new and reference groups, equating or linking becomes susceptible.

This explains why DIF and linking results become worrisome around the same degree of population shift (Condition 4 and Condition 5).

When the quality of linking is deemed susceptible because of large population shift, new linking samples can be constructed to have similar group compositions. There could be different ways to construct linking samples with similar group composition. One way is to apply random sampling with replacement to either the RF group or the NF group. If there is reason to believe that the NF group will be more likely representing the future test-taking population, then sampling can be done to the RF group to match the NF group proportions. Otherwise, if there is reason to believe that the future population will be closer to the RF group, then sampling can be done to the NF group to match the group composition in the reference form group. Another way to adjust similarity in group composition could be to apply a weighting scheme to individual test takers in the RF (or the NF) so that the group taking the RF has the same group composition as the group taking the NF (Qian, von Davier, & Jiang, 2013).

Aside from dealing with the linking samples, people may also work with test developers to avoid items with differential performance between different major subgroups. People may just remove anchor items with STD P-DIF before conducting equating or linking if there are many anchor items between the NFs and RFs.

However, the focus of this study was to raise the awareness of the impact of population shift on the results of linking and DIF analyses, not to recommend a specific course of action. Even though Sinharay et al. (2011) did not find impact of population shift on equating, the results of this study indicate that population shift may still affect equating substantially even if the two subgroups are similar in terms of ability. In our data, Group A and Group B perform similarly on Section 1, especially on NF1 and NF2 (Tables 2 and 3). Even for these two forms, the impact of population shift is still pronounced.

Finally, this study has its own limitations. Most apparently, the finding of this study may not be generalized to other programs. This study found that a population shift of 30% or more made a difference in linking and DIF. Another study using data from a different testing program may find different results and come up with different conclusions. For example, Sinharay et al. (2011, p. 27, description of Synthetic Subsample 9) did not find significant influence of population shift on equating even when the difference in the NEFL proportions was 40% between new and reference equating samples. On the other hand, this lack of generalizability

tells the need for each testing program to check the impact of population shift using its own data and to decide when and how to protect linking when population shift happens based on its own data characteristics and practical limitations.

References

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum No. RM-94-10). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01994.x

Dorans, N. J., & Holland, P. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (Research Report No. RR-92-10). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dorans, N. J., & Liu, J. (2009). *Score equity assessment: Development of a prototype analysis using SAT mathematics test data across several administrations* (Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02165.x

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales.* New York, NY: Springer-Verlag.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. RR-85-43). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00128

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Liang, L., Dorans, N., & Sinharay, S. (2009). *First language of examinees and its relationship to equating* (Research Report No. RR-09-08). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02162.x

Qian, J., von Davier, A. A., & Jiang, Y. (2013). Achieving a stable scale for an assessment with multiple forms: Weighting test samples in IRT linking and equating. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Springer proceedings in mathematics & statistics* (pp. 171–185). New York, NY: Springer.

Qu, Y., Costanzo, K., & Stuart, J. (2012). *Score equity assessment: Invariance of TOEIC Bridge equating across different examinee subgroups* (Statistical Report No. SR-2012-034). Princeton, NJ: Educational Testing Service.

Sinharay, S., Dorans, N., & Liang, L. (2011). First language of test takers and fairness assessment procedures. *Educational Measurement: Issues and Practice*, *30*(2), 25–35.