Observed Score and True Score Equating for Multidimensional Item Response Theory under Nonequivalent Group Design

Ou Zhang
Pearson
M. David Miller
James Algina
University of Florida

What Is Test Equating?

- Equating is a ***statistical process*** that is used to adjust scores on different test forms so that scores on the forms are comparable (Kolen & Brennan, 2004).

Five Basic Requirements of Test Equating

- Equal Constructs
- Equal reliability
- Symmetry
- Equity
- Population Invariance

Five Basic Requirements of Test Equating

- Equal Constructs
- Equal reliability
- Symmetry
  (A → B transformation ⟷ B → A transformation)
- Equity
- Population Invariance

Multidimensional Item Response Theory

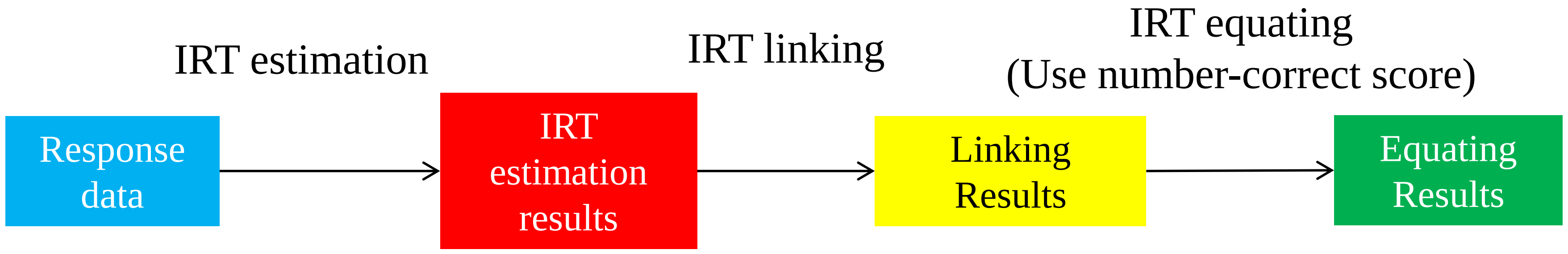
- Multidimensional Item Response Theory Model (MIRT)
  - Compensatory MIRT model (McKinley & Reckase, 1983)

$$P(x_{ij}=1|\boldsymbol{\theta}_j,\mathbf{a}_\mathbf{i},d_i)=\frac{e^{D(\mathbf{a'}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j}+d_i)}}{1+e^{D(\mathbf{a'}_\mathbf{i}\boldsymbol{\theta}_\mathbf{j}+d_i)}}$$

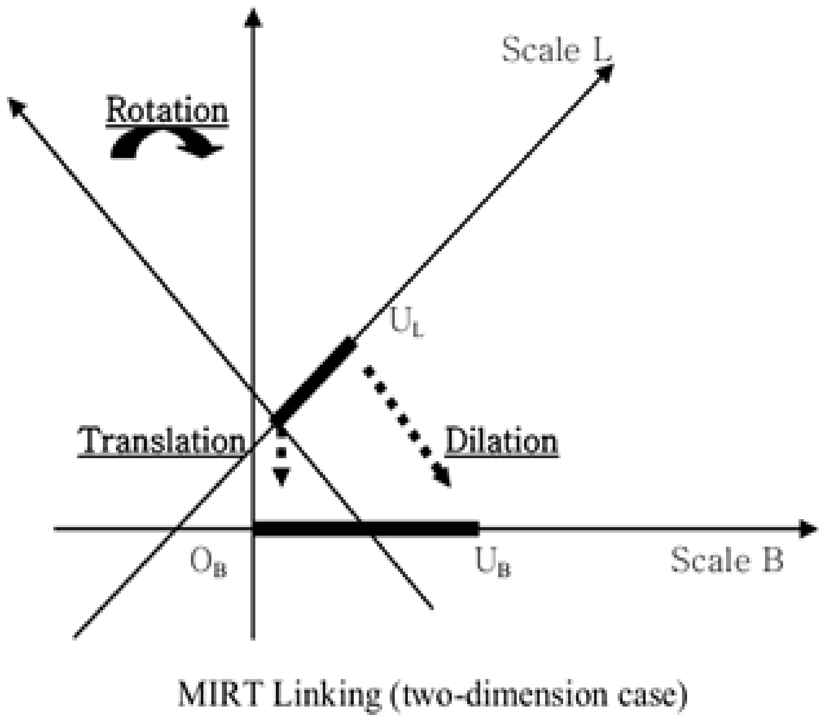
$\boldsymbol{\theta}_s$ represents **multiple** ability parameters associated with each respondent,
$\mathbf{a}_i$ represents **multiple** discrimination parameters associated with each item,
and d_i represents an item's location on an item response **surface**.

Common Procedure of Equating in IRT

- Step 1: IRT Estimation
- Step 2: IRT Linking/Scaling Aligning
- Step 3: IRT Equating (use number-correct scores, if necessary)



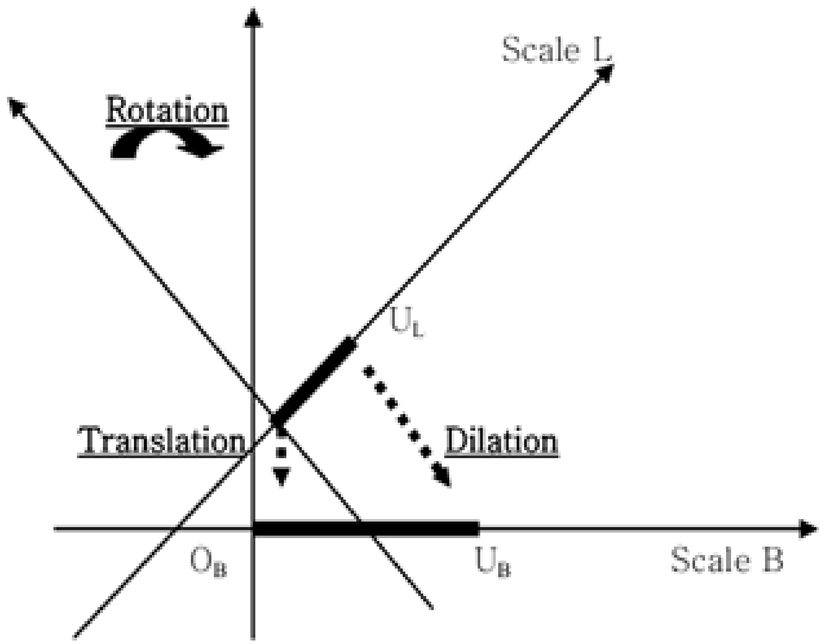
MIRT Linking/Scale Aligning



- Dilation: adjust unit

(Figure adapted from Min, 2003)

MIRT Linking/Scale Aligning

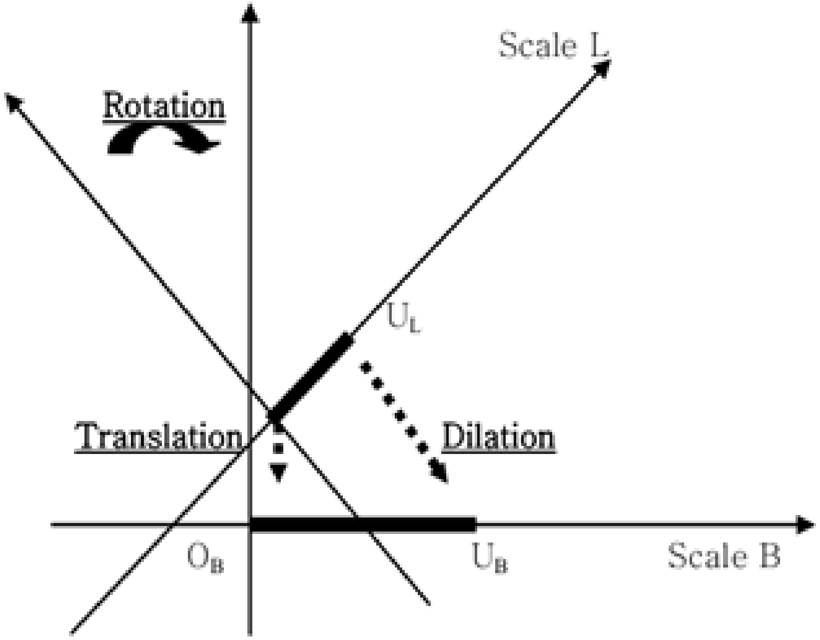


MIRT Linking (two-dimension case)

- Dilation: adjust unit
- Translation: adjust original zero point

(Figure adapted from Min, 2003)

MIRT Linking/Scale Aligning



MIRT Linking (two-dimension case)

- Dilation: adjust unit
- Translation: adjust original zero point
- Rotation: adjust the entire multidimensional axis systems so that both axis systems are in the same direction.

(Figure adapted from Min, 2003)

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?
- In the MIRT, the ability is a vector- $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?
- In the MIRT, the ability is a vector- $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$
- Demonstrating equivalence between two ability vectors from different test forms is :
  - complex

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?
- In the MIRT, the ability is a vector- $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$
- Demonstrating equivalence between two ability vectors from different test forms is :
  - complex
  - indirect

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?
- In the MIRT, the ability is a vector- $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$
- Demonstrating equivalence between two ability vectors from different test forms is :
  - complex
  - indirect

$$\theta_1 = [1.0, 1.0] \text{ in Form A}$$
$$\theta_2 = [1.0, 1.5] \text{ in Form B}$$

Equivalent or not?

Symmetry Property and Unidimensionalization

- Are We Done after MIRT Linking/Scale Aligning?
- In the MIRT, the ability is a vector- $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]$
- Demonstrating equivalence between two ability vectors from different test forms is :
  - complex
  - indirect

$$\theta_1 = [1.0, 1.0] \text{ in Form A}$$
$$\theta_2 = [1.0, 1.5] \text{ in Form B}$$

Equivalent or not?

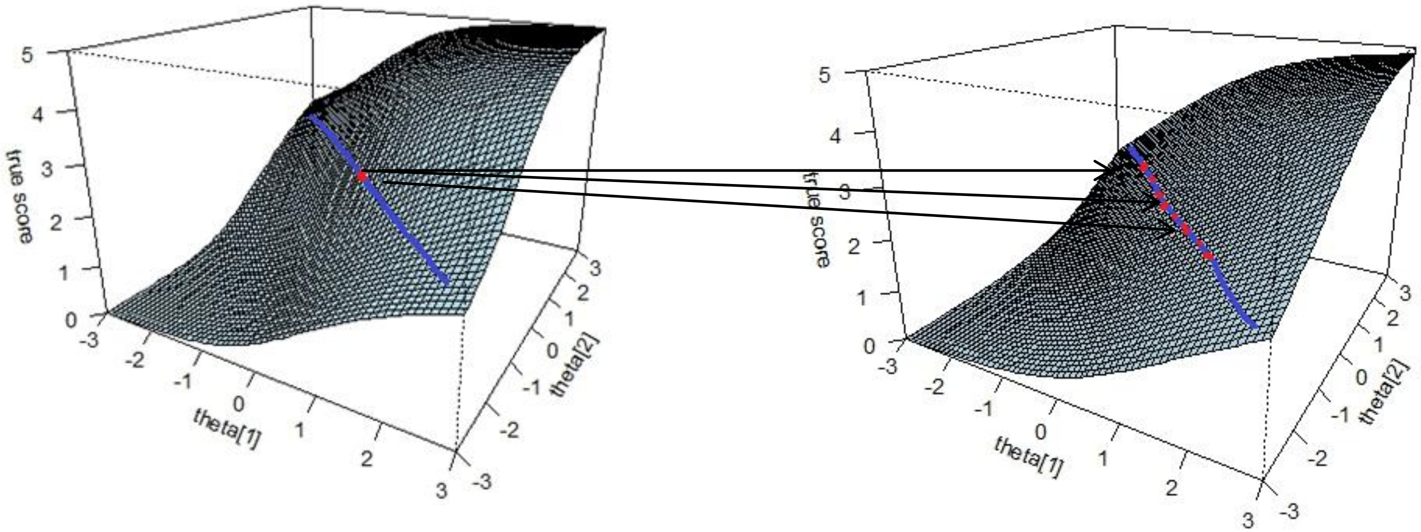
Comparability of the MIRT measure ?

Symmetry Property and Unidimensionalization (cont.)

- Possible Violation of Test Equating's Symmetry Requirement

Symmetry Property and Unidimensionalization (cont.)

- Possible Violation of Test Equating's Symmetry Requirement
  - If we use the MIRT ability estimate **vector** as a measure of ability, a particular true score ($\tau(\boldsymbol{\theta}) = \sum p(\boldsymbol{\theta})$) for one test form (i.e., test A) on the test characteristic surface (TCS), corresponds to infinite numbers of combinations of ability vectors on the other test form's TCS equiprobable contour (i.e., Test B) when both test forms are already in the same scale.



Symmetry Property and Unidimensionalization (cont.)

- One possible solution to make the MIRT equating available is to use the number-correct score or true score as the ability measure in MIRT.

Symmetry Property and Unidimensionalization (cont.)

- One possible solution to make the MIRT equating available is to use the number-correct score or true score as the ability measure in MIRT.

- When the number-correct score or true score is used as the ability measure, the MIRT ability vector is unidimensionalized.

Symmetry Property and Unidimensionalization (cont.)

- One possible solution to make the MIRT equating available is to use the number-correct score or true score as the ability measure in MIRT.

- When the number-correct score or true score is used as the ability measure, the MIRT ability vector is unidimensionalized.

- This process is a linear combination procedure and called "**unidimensionalization**" (Zhang, 2012).

Symmetry Property and Unidimensionalization (cont.)

- One possible solution to make the MIRT equating available is to use the number-correct score or scale score as the ability measure in MIRT.
- When the number-correct score or scale score is used as the ability measure, the MIRT ability vector is unidimensionalized.
- This process is a linear combination procedure and called "unidimensionalization" (Zhang, 2012).
- Unidimensionalization process devectorizes the vector or multidimensional features in the MIRT framework so that the ability measures from different test forms are comparable..

Symmetry Property and Unidimensionalization (cont.)

- One possible solution to make the MIRT equating available is to use the number-correct score or scale score as the ability measure in MIRT.
- When the number-correct score or scale score is used as the ability measure, the MIRT ability vector is unidimensionalized.
- This process is a linear combination procedure and called "unidimensionalization" (Zhang, 2012).
- Unidimensionalization process devectorizes the vector or multidimensional features in the MIRT framework so that the ability measures from different test forms are comparable..
- Most importantly, through the process of unidimensionalization, the symmetry property of equating (Lord, 1980) for two test forms under MIRT framework is satisfied.

A Methodology Foundation of Unidimensionalization

Unidimensional Approximation of MIRT (Zhang & Stout, 1999)

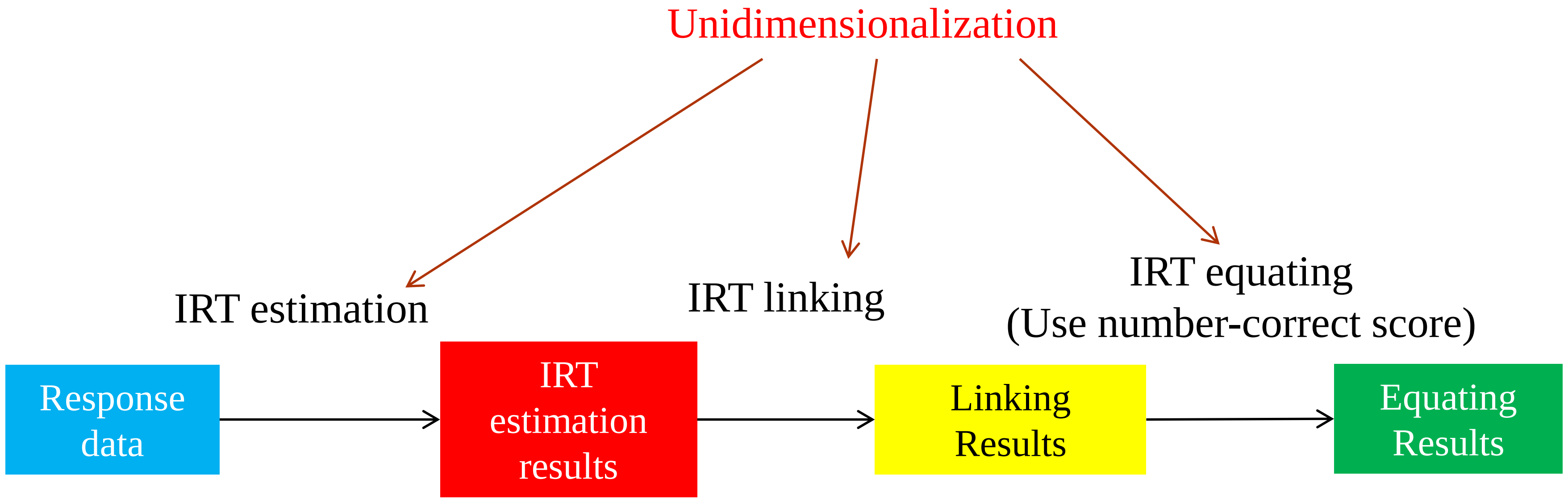
- Any set of item responses adequately modeled by a MIRT model, can be closely approximated by a unidimensional IRT model with estimated unidimensional ability composite (Θ_α) and estimated unidimensional item parameters ($\hat{a}_{\alpha j}$, $\hat{b}_{\alpha j}$, $\hat{T}_{\alpha j}$) (Zhang & Stout, 1999).

- The ability composite Θ_α of the multidimensional ability vector (i.e., $\mathbf{\Theta} = [\theta_1, \theta_2, ..., \theta_m]$) is defined as

$$\Theta_\alpha = \hat{\mathbf{a}}^{\mathbf{T}}\hat{\boldsymbol{\theta}} = \boldsymbol{\alpha}^t\mathbf{\Theta} = \sum_{j=1}^{d}\alpha_j\theta_j$$

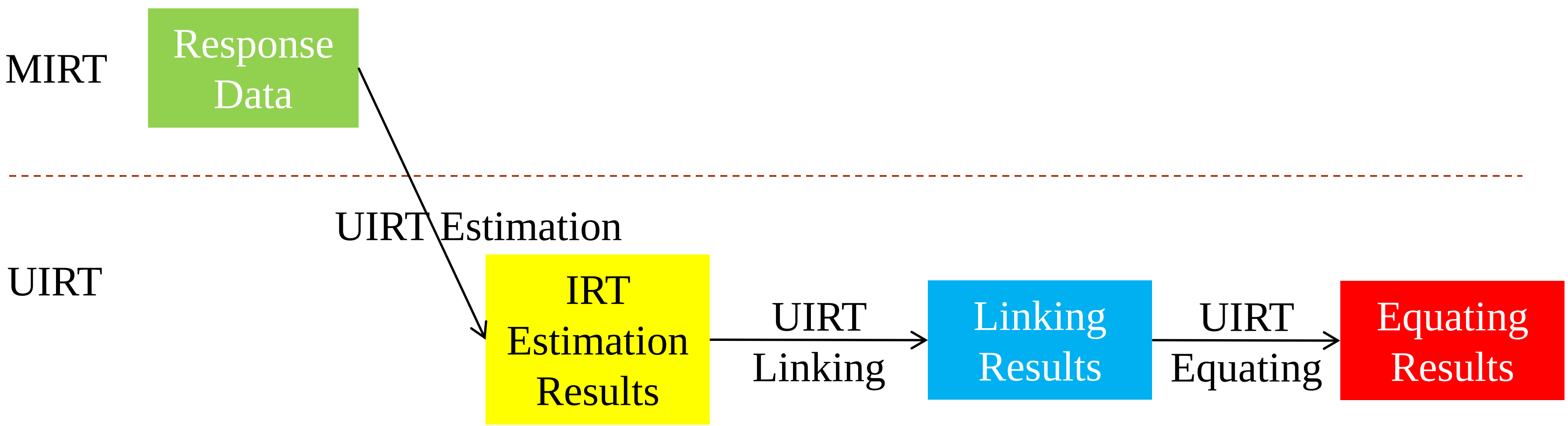
Unidimensionalization in MIRT Equating Procedures

- 4 Possible Procedures of MIRT Equating



Unidimensionalization in MIRT Equating Procedures

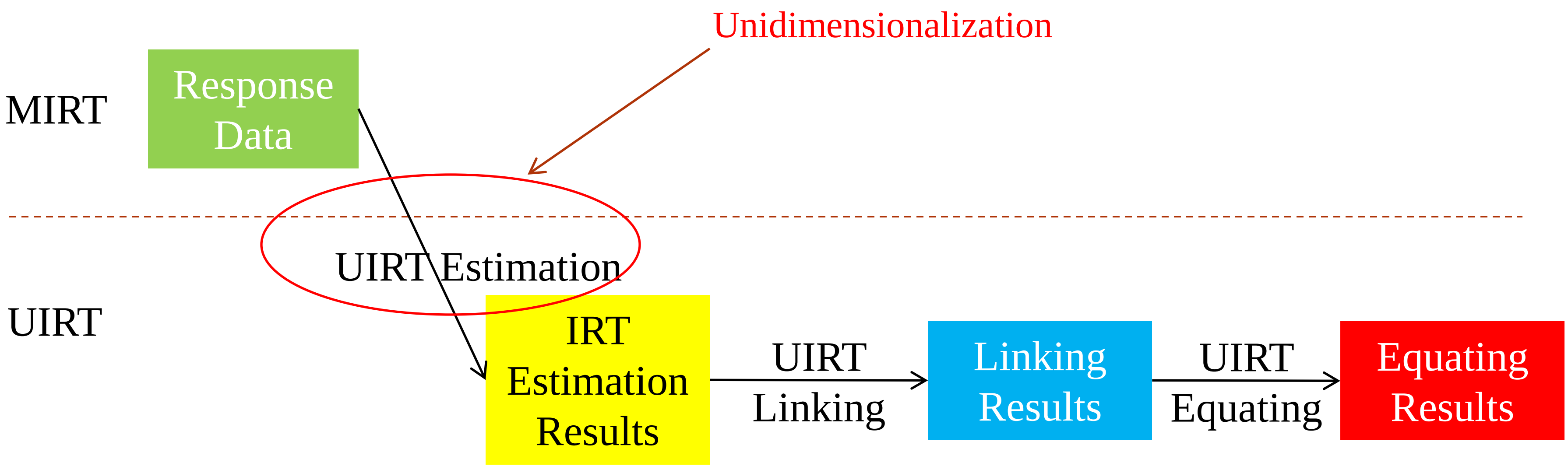
- Possible Procedure 1:
  - UIRT estimation - UIRT linking - UIRT equating



Unidimensionalization at IRT Estimation stage

Unidimensionalization in MIRT Equating Procedures

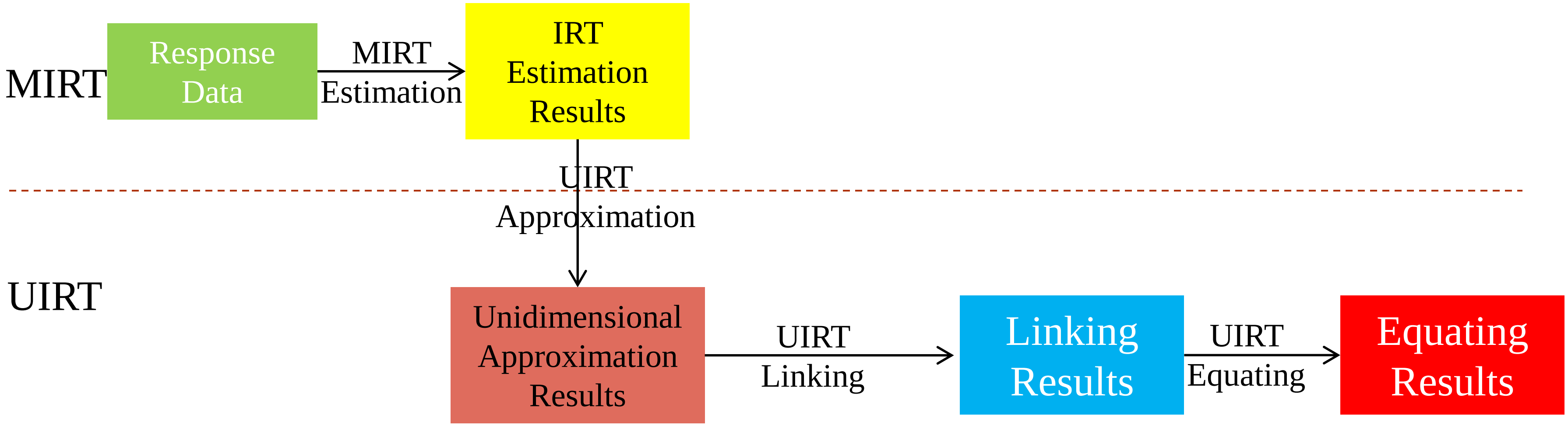
- Possible Procedure 1:
  - UIRT estimation - UIRT linking - UIRT equating



Unidimensionalization at IRT Estimation stage

Unidimensionalization in MIRT Equating Procedures

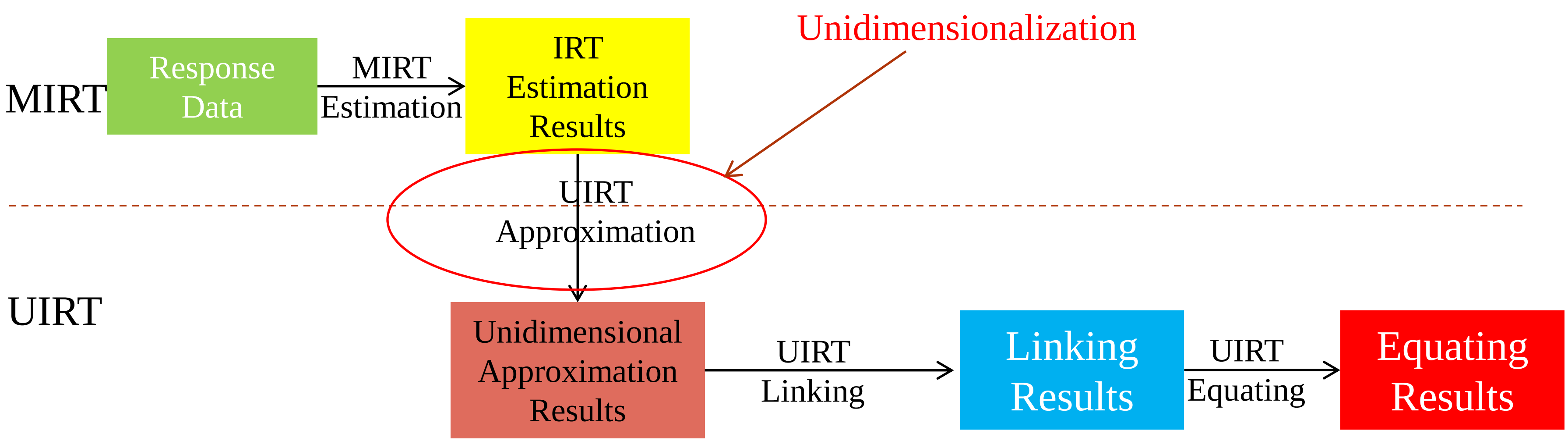
- Possible Procedure 2:
  - MIRT estimation - UIRT approximation - UIRT linking - UIRT equating



Unidimensionalization before IRT linking

Unidimensionalization in MIRT Equating Procedures

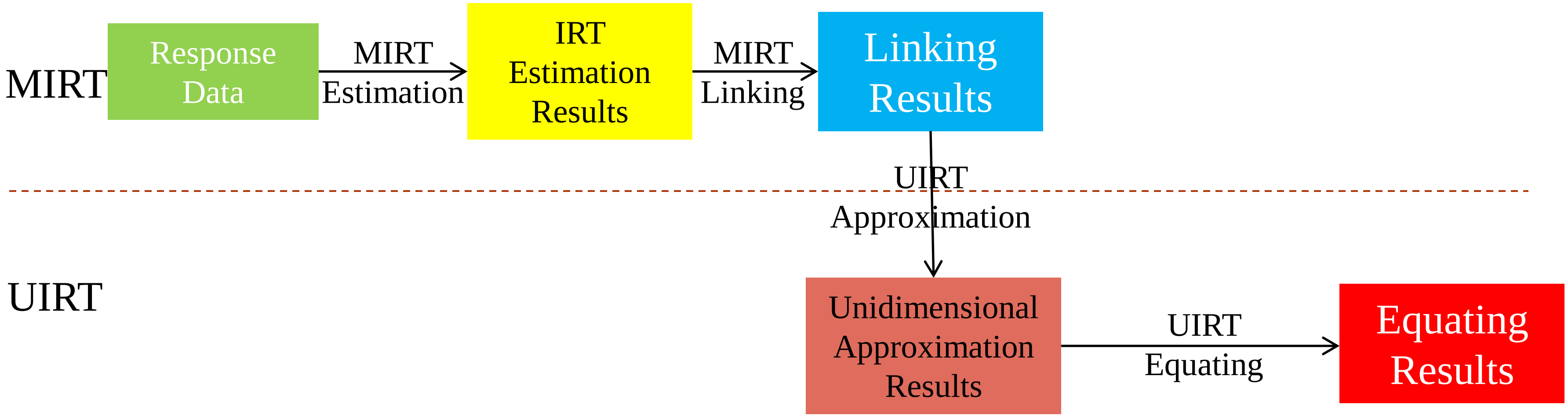
- Possible Procedure 2:
  - MIRT estimation - UIRT approximation - UIRT linking - UIRT equating



Unidimensionalization before IRT linking

Unidimensionalization in MIRT Equating Procedures

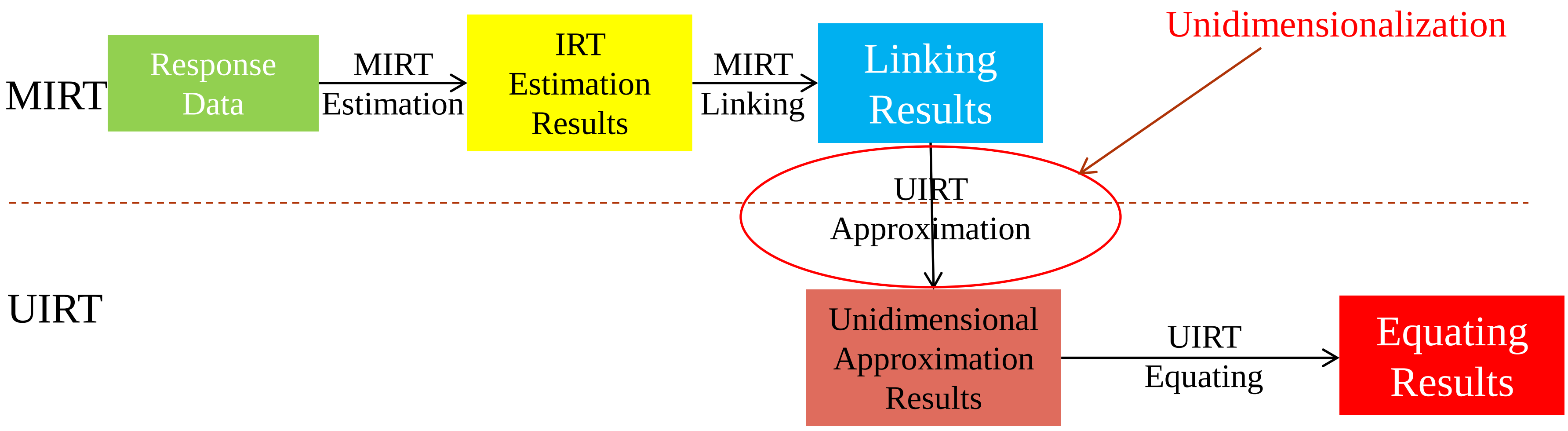
- Possible Procedure 3:
  - MIRT Estimation - MIRT Linking - UIRT Approximation - UIRT Equating



Unidimensionalization before Test Equating stage

Unidimensionalization in MIRT Equating Procedures

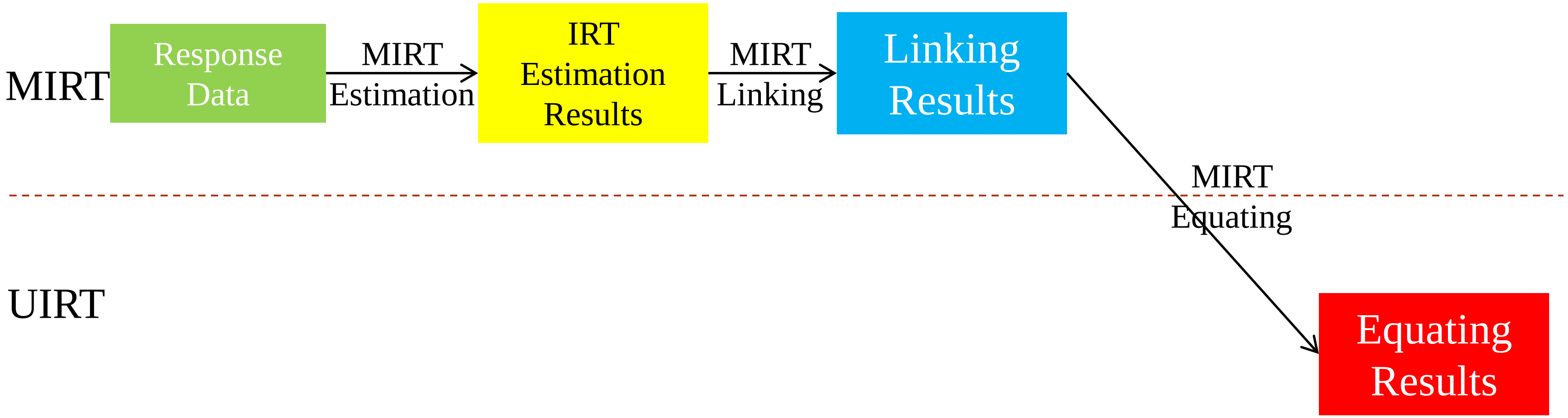
- Possible Procedure 3:
  - MIRT Estimation - MIRT Linking - UIRT Approximation - UIRT Equating



Unidimensionalization before Test Equating stage

Unidimensionalization in MIRT Equating Procedures

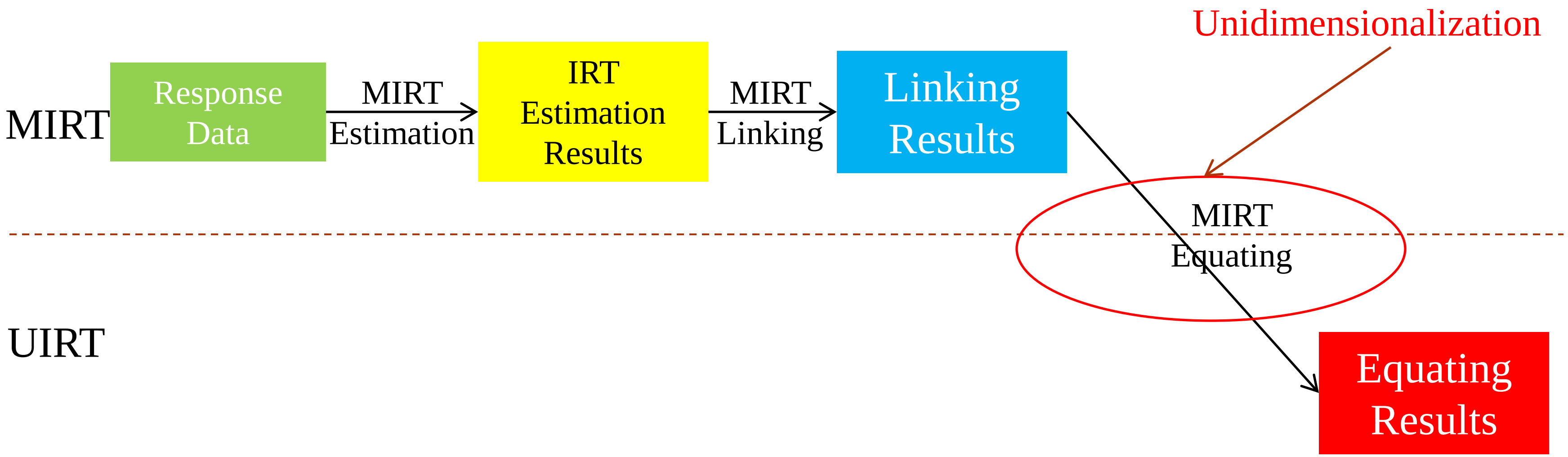
- Possible Procedure 4:
  - MIRT estimation - MIRT linking - MIRT Equating



Unidimensionalization at MIRT Equating stage

Unidimensionalization in MIRT Equating Procedures

- Possible Procedure 4:
  - MIRT estimation - MIRT linking - MIRT Equating



Unidimensionalization at MIRT Equating stage

Purpose of Study

- To evaluate the performance of the MIRT equating procedures under NEAT design.

Purpose of Study

- To evaluate the performance of the MIRT equating procedures under NEAT design.

- To explore how different MIRT linking methods interacting with MIRT equating procedures (Brossman, 2010) impact on the equating results, under various testing conditions.

Purpose of Study

- To evaluate the performance of the MIRT equating procedures under NEAT design.
- To explore how different MIRT linking methods interacting with MIRT equating procedures (Brossman, 2010) impact on the equating results, under various testing conditions.
- To provide a possible guidance to educational practitioners for their future MIRT equating application.

MIRT Linking Methods used in the Study

- Min's (M) Method (2003)
- Oshima, Davey and Lee's (ODL) Method (2000)
  - The direct method (OD)
  - The Test Characteristic Function method (TCF)
  - The Item Characteristic Function method (ICF)
- Reckase and Martineau (NOP) Method (2004)

- Coefficients Obtained from These MIRT Linking Methods
  - Rotation Matrix - **T**
  - Translation Vector - **m**
  - Dilation Vector - **K**

MIRT Equating Methods used in the Study

- MIRT Equating Methods (Brossman, 2010)
  - Full MIRT observed score equating method (MOSE)
    - (Possible procedure 4)
  - Unidimensional approximation of MIRT true score equating (ATSE)
    - (Possible procedure 3)
  - Unidimensional approximation of MIRT observed score equating (AOSE)
    - (Possible procedure 3)

MIRT Equating Methods used in the Study

- MIRT Equating Methods (Brossman, 2010)
  - Full MIRT observed score equating method (MOSE)
    - (Possible procedure 4)
  - Unidimensional approximation of MIRT true score equating (ATSE)
    - (Possible procedure 3)
  - Unidimensional approximation of MIRT observed score equating (AOSE)
    - (Possible procedure 3)

So, only Procedure **3** and Procedure **4** were applied in this study.

MIRT Equating Methods for This Study (cont.)

- Full MIRT Observed Score Equating Procedure
  - The full MIRT observed score equating method is a straightforward extension of UIRT observed score equating through the compound binomial recursion formula.

$$f(x) = \sum_{1}\sum_{2}\cdots\sum_{m} f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta})$$

  - or

$$f(x) = \int_{1}\int_{2}\cdots\int_{m} f(x \mid \boldsymbol{\theta})\psi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

  - where m is defined as the number of dimensions.

MIRT Equating Methods for This Study (cont.)

- Unidimensional Approximation of MIRT True Score Equating
- The UIRT true score equating procedure is utilized to equated composite true scores (T_α) on both multidimensional test forms. Thus,

$$irt_B(\tau_{\alpha Bi}) = \tau_B(\tau_{\alpha Ei}{}^{-1})$$

- and

$$func(\theta_{\alpha i}) = \tau_{\alpha A} - \sum_{j:A} p_{ij}(\theta_{\alpha i} \mid a_{\alpha j}, b_{\alpha j}, c_j)$$

- Finally, the composite true score on the base form $\tau_{\alpha B}(\theta_\alpha)$ associated with the composite true score on the equated form $\tau_{\alpha E}(\theta_\alpha)$ can be computed as

$$\tau_{\alpha B} = \sum_{j:B} p_{ij}(\theta_{\alpha i} \mid a_{\alpha j}, b_{\alpha j}, c_j)$$

MIRT Equating Methods for This Study (cont.)

- Unidimensional Approximation of MIRT Observed Score Equating

  - The conditional distributions for the unidimensional ability composite $f(x|\theta_\alpha)$ is determined at each composite ability level (θ_α) through the compound binomial recursion formula.

$$f(x) = \sum_{\theta_\alpha} f(x|\theta_{\alpha i})\psi(\theta_{\alpha i})$$

  - Then,

$$f(x) = \int_{\theta_\alpha} f(x|\theta_{\alpha i})\psi(\theta_{\alpha i})d\theta_\alpha$$

Simulation Design

- MIRT model used: M2PL (With D=1.7)
- Test length: total 40 items, 20 anchor items
- Test structure: Approximate simple structure (APSS) and complex structure (CS)
- Sample size: 2000
- Replication time: 200
- Population Design:
  - Null condition
  - Mean-difference
  - SD-difference
  - Correlation-difference
- MIRT estimation software: TESTFACT
- MIRT linking and MIRT equating: R

Evaluation Criteria

- Weighted average equating bias ($Bias_w$)

$$Bias_i = \frac{\sum_{k=1}^{N}\left[\hat{e}_{base_k}(x_i) - e_{base}(x_i)\right]}{N}$$

For the entire test: $$Bias_w = \sum_{x=1}^{39} Bias\left[\hat{e}_{base}(x_i)\right]P(x_i)$$

- Weighted Average Root Mean Square Deviation (ARMSDw)

$$RMSD_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[\hat{e}_{base_k}(x_i) - e_{base}(x_i)\right]^2}$$

For the entire test: $$ARMSD_w = \sum_{x=1}^{39} RMSD\left[\hat{e}_{base}(x_i)\right]P(x_i)$$

Results

Repeated ANOVA Analysis *(BIASw* and *ARMSDw)*

Statistic	Factors	Source	Partial ω^2	Statistic	Factors	Source	Partial ω^2
$ARMSD_w$	Between	test_str	0.0067	$Bias_w$	Between	test_str	0.02067
	Between	group	0.91944		Between	group	0.92557
	Between	test_str*group	0.02128		Between	test_str*group	0.00458
	Within	link	0.94089		Within	link	0.8497
	Within	link*test_str	0.03362		Within	link*test_str	0.00641
	Within	link*group	0.94122		Within	link*group	0.88045
	Within	link*test_str*group	0.15599		Within	link*test_str*group	0.06019
	Within	equat	0.57653		Within	equat	0.47878
	Within	equat*test_str	0.01727		Within	equat*test_str	0.01469
	Within	equat*group	0.58711		Within	equat*group	0.46236
	Within	equat*test_str*group	0.02497		Within	equat*test_str*group	0.00459
	Within	link*equat	0.38335		Within	link*equat	0.00185
	Within	link*equat*test_str	0.03872		Within	link*equat*test_str	0.00342
	Within	link*equat*group	0.40483		Within	link*equat*group	0.00873
	Within	link*equat*test_str*group	0.04714		Within	link*equat*test_str*group	0.00429

- The largest effect size: linking method * group distribution

Results

Repeated ANOVA Analysis (*BIASw* and *ARMSDw*)

Statistic	Factors	Source	Partial ω^2	Statistic	Factors	Source	Partial ω^2
$ARMSD_w$	Between	test_str	0.0067	$Bias_w$	Between	test_str	0.02067
	Between	group	0.91944		Between	group	0.92557
	Between	test_str*group	0.02128		Between	test_str*group	0.00458
	Within	link	0.94089		Within	link	0.8497
	Within	link*test_str	0.03362		Within	link*test_str	0.00641
	Within	link*group	0.94122		Within	link*group	0.88045
	Within	link*test_str*group	0.15599		Within	link*test_str*group	0.06019
	Within	equat	0.57653		Within	equat	0.47878
	Within	equat*test_str	0.01727		Within	equat*test_str	0.01469
	Within	equat*group	0.58711		Within	equat*group	0.46236
	Within	equat*test_str*group	0.02497		Within	equat*test_str*group	0.00459
	Within	link*equat	0.38335		Within	link*equat	0.00185
	Within	link*equat*test_str	0.03872		Within	link*equat*test_str	0.00342
	Within	link*equat*group	0.40483		Within	link*equat*group	0.00873
	Within	link*equat*test_str*group	0.04714		Within	link*equat*test_str*group	0.00429

- The largest effect size: linking method * group distribution
- The 2^{nd} largest effect size: equating method * group distribution

Results

Repeated ANOVA Analysis (*BIASw* and *ARMSDw*)

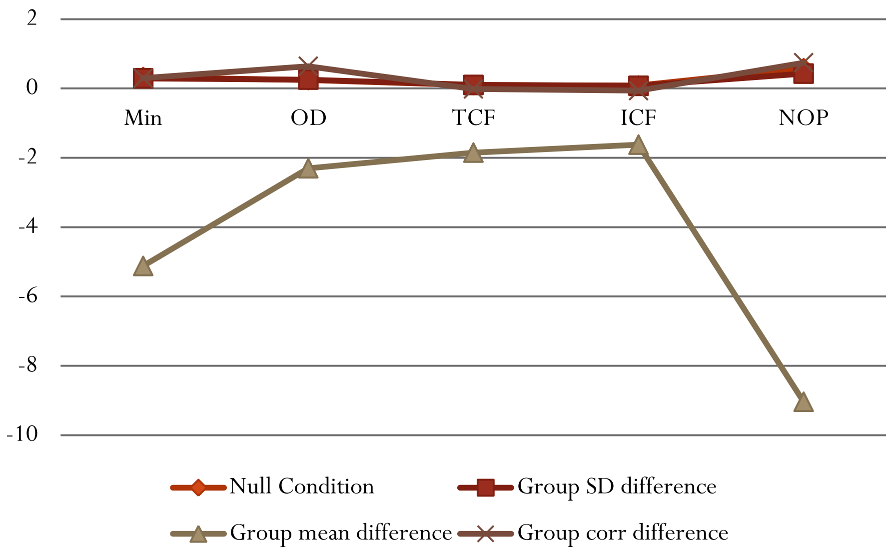
Statistic	Factors	Source	Partial ω^2	Statistic	Factors	Source	Partial ω^2
$ARMSD_w$	Between	test_str	0.0067	$Bias_w$	Between	test_str	0.02067
	Between	group	0.91944		Between	group	0.92557
	Between	test_str*group	0.02128		Between	test_str*group	0.00458
	Within	link	0.94089		Within	link	0.8497
	Within	link*test_str	0.03362		Within	link*test_str	0.00641
	Within	link*group	0.94122		Within	link*group	0.88045
	Within	link*test_str*group	0.15599		Within	link*test_str*group	0.06019
	Within	equat	0.57653		Within	equat	0.47878
	Within	equat*test_str	0.01727		Within	equat*test_str	0.01469
	Within	equat*group	0.58711		Within	equat*group	0.46236
	Within	equat*test_str*group	0.02497		Within	equat*test_str*group	0.00459
	Within	link*equat	0.38335		Within	link*equat	0.00185
	Within	link*equat*test_str	0.03872		Within	link*equat*test_str	0.00342
	Within	link*equat*group	0.40483		Within	link*equat*group	0.00873
	Within	link*equat*test_str*group	0.04714		Within	link*equat*test_str*group	0.00429

- The largest effect size: linking method * group distribution
- The 2nd largest effect size: equating method * group distribution
- Test structure and all the interactions including test structure- very small effect size

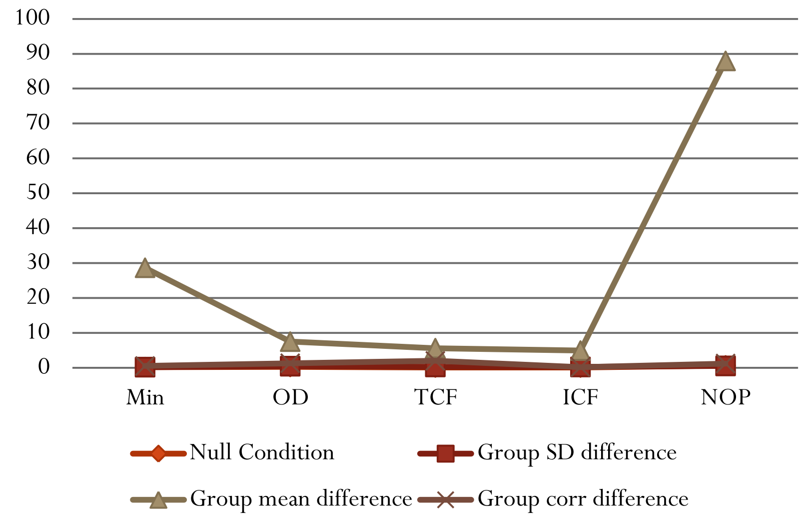
Results

- Comparison for the Linking Method x Group Distribution Interaction

$Bias_w$



$ARMSD_w$

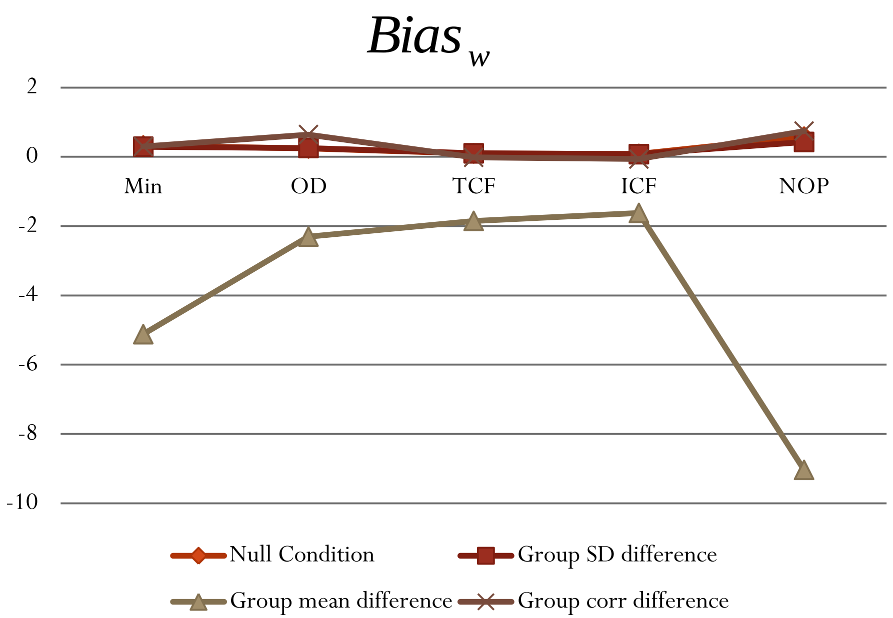


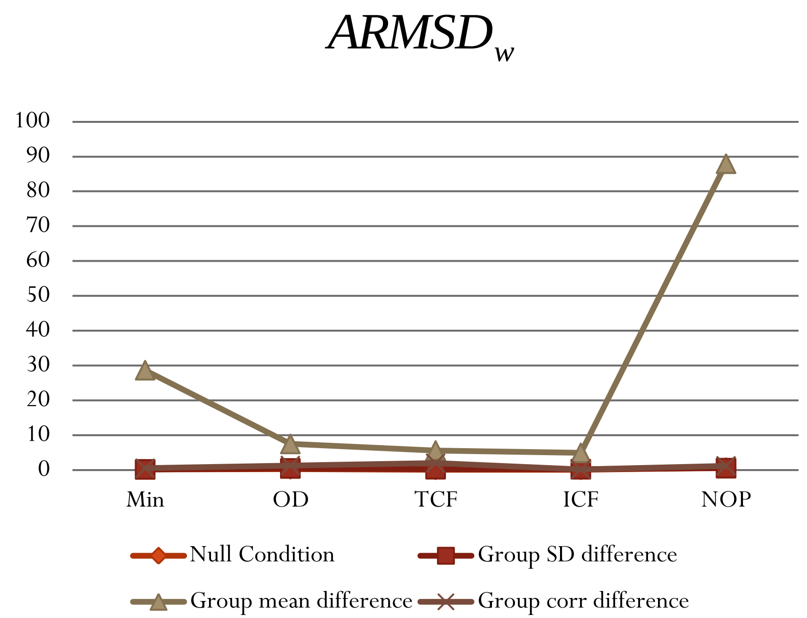
Bias	Min	OD	TCF	ICF	NOP	**ARMSD**	Min	OD	TCF	ICF	NOP
Null Condition	0.32926	0.24153	0.09908	0.08588	0.58308		0.23089	0.28789	0.11543	0.11071	0.69485
Group SD difference	0.29248	0.25659	0.10186	0.07712	0.42949		0.20884	0.48575	0.26935	0.26063	0.55671
Group mean difference	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351		28.6865	7.52435	5.64556	4.9614	87.9546
Group corr difference	0.29875	0.64432	-0.015	-0.0637	0.74563		0.54088	1.33343	2.00157	0.19166	1.24462

- Overall: TCF and ICF performed best across all group distribution conditions;

Results

- Comparison for the Linking Method x Group Distribution Interaction





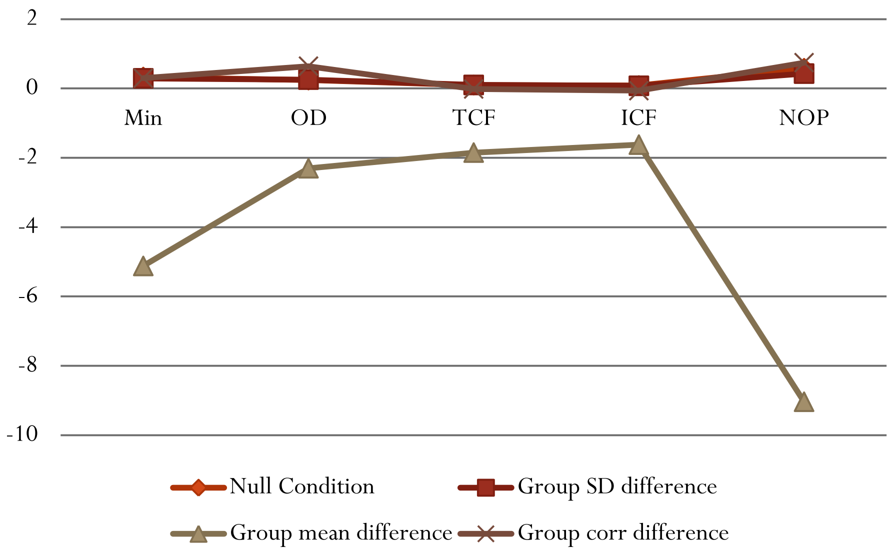
Bias	Min	OD	TCF	ICF	NOP	**ARMSD**	Min	OD	TCF	ICF	NOP
Null Condition	0.32926	0.24153	0.09908	0.08588	0.58308		0.23089	0.28789	0.11543	0.11071	0.69485
Group SD difference	0.29248	0.25659	0.10186	0.07712	0.42949		0.20884	0.48575	0.26935	0.26063	0.55671
Group mean difference	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351		28.6865	7.52435	5.64556	4.9614	87.9546
Group corr difference	0.29875	0.64432	-0.015	-0.0637	0.74563		0.54088	1.33343	2.00157	0.19166	1.24462

- Overall: TCF and ICF performed best across all group distribution conditions; OD and M methods' performances are next;

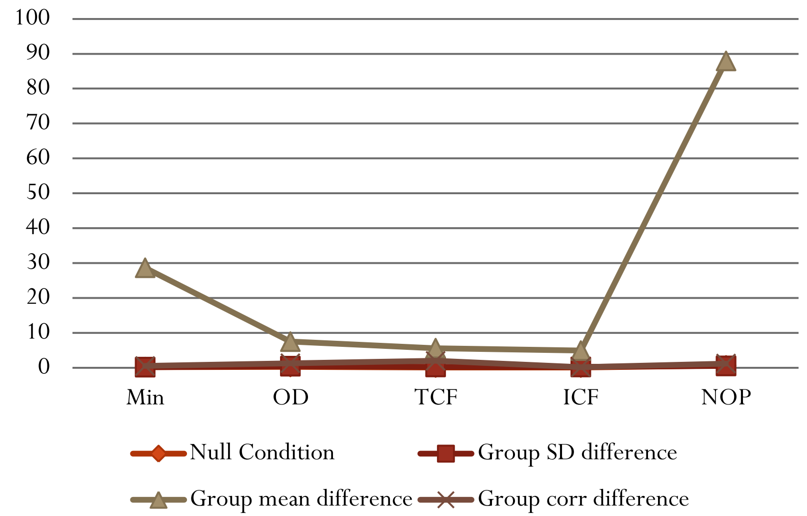
Results

- Comparison for the Linking Method x Group Distribution Interaction

$Bias_w$



$ARMSD_w$



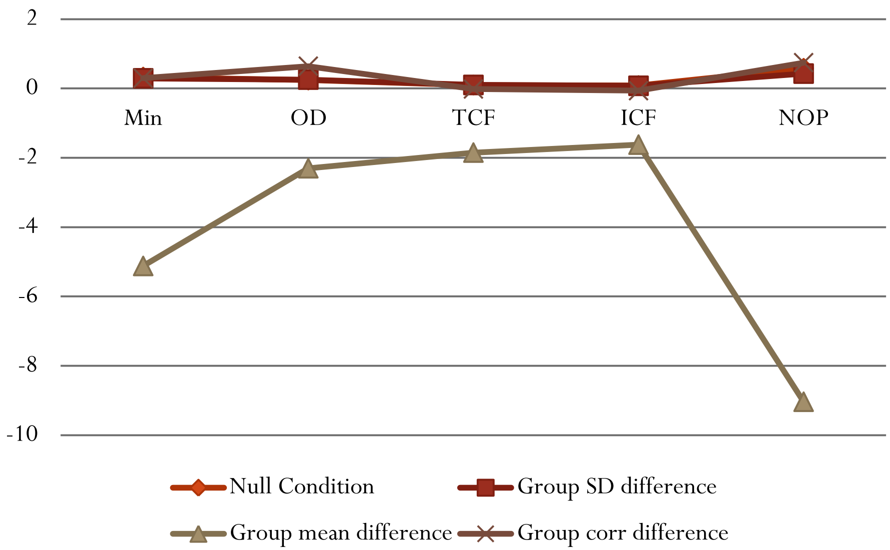
Bias	Min	OD	TCF	ICF	NOP	**ARMSD**	Min	OD	TCF	ICF	NOP
Null Condition	0.32926	0.24153	0.09908	0.08588	0.58308		0.23089	0.28789	0.11543	0.11071	0.69485
Group SD difference	0.29248	0.25659	0.10186	0.07712	0.42949		0.20884	0.48575	0.26935	0.26063	0.55671
Group mean difference	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351		28.6865	7.52435	5.64556	4.9614	87.9546
Group corr difference	0.29875	0.64432	-0.015	-0.0637	0.74563		0.54088	1.33343	2.00157	0.19166	1.24462

- Overall: TCF and ICF performed best across all group distribution conditions; OD and M methods' performances are next; NOP method performed worst among all 5 Linking methods.

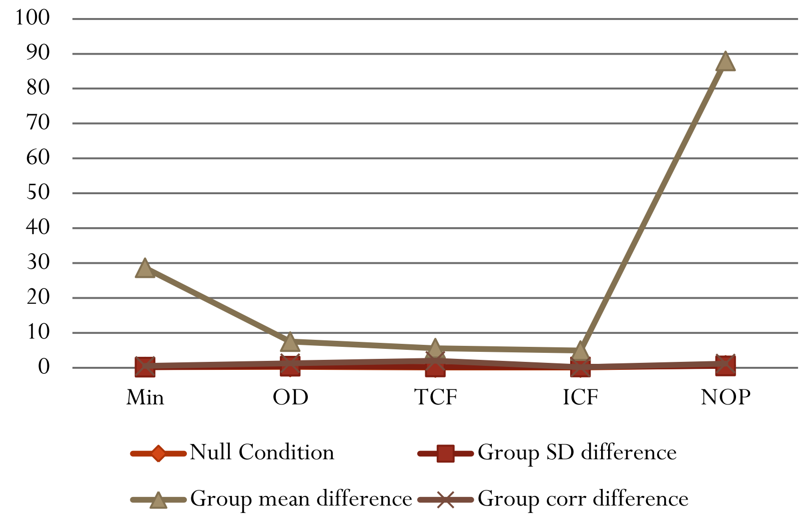
Results (cont.)

- Comparison for the Linking Method x Group Distribution Interaction

$Bias_w$



$ARMSD_w$



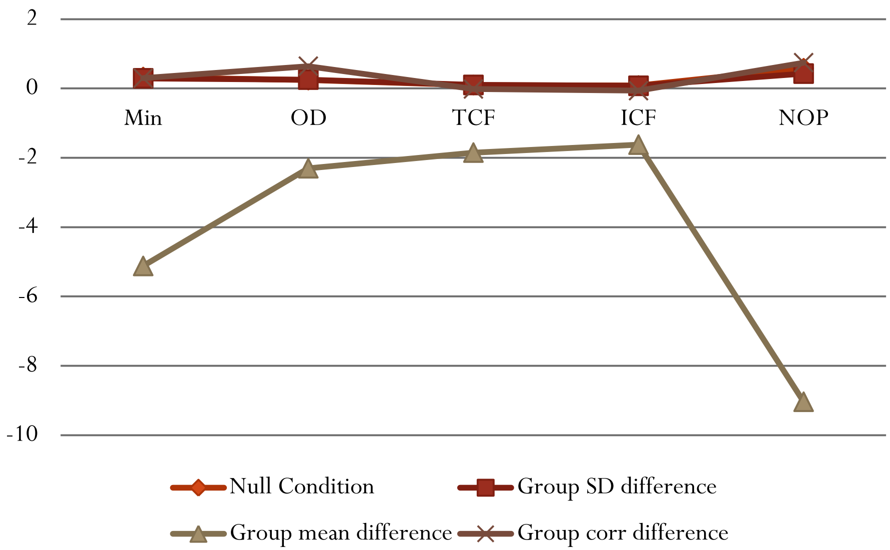
Bias	Min	OD	TCF	ICF	NOP	**ARMSD**	Min	OD	TCF	ICF	NOP
Null Condition	0.32926	0.24153	0.09908	0.08588	0.58308		0.23089	0.28789	0.11543	0.11071	0.69485
Group SD difference	0.29248	0.25659	0.10186	0.07712	0.42949		0.20884	0.48575	0.26935	0.26063	0.55671
Group mean difference	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351		28.6865	7.52435	5.64556	4.9614	87.9546
Group corr difference	0.29875	0.64432	-0.015	-0.0637	0.74563		0.54088	1.33343	2.00157	0.19166	1.24462

- Under the null condition, group SD difference, group corr difference, five MIRT linking methods have similar equating performances

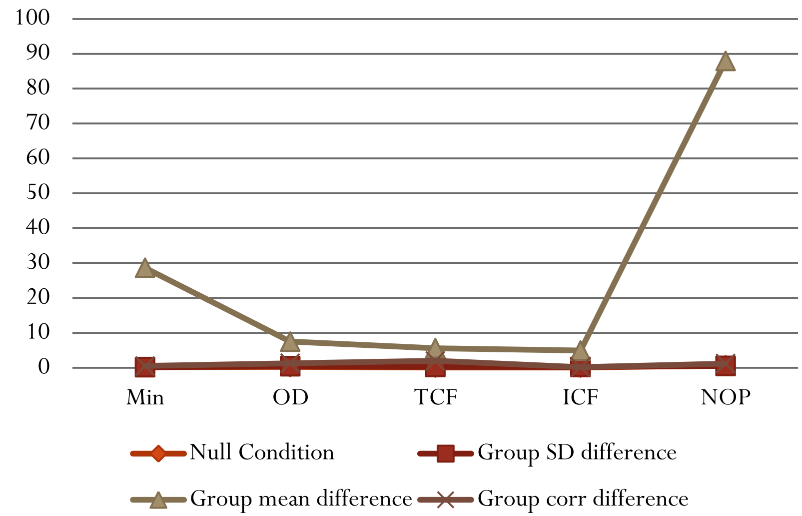
Results (cont.)

- Comparison for the Linking Method x Group Distribution Interaction

Bias$_w$



ARMSD$_w$

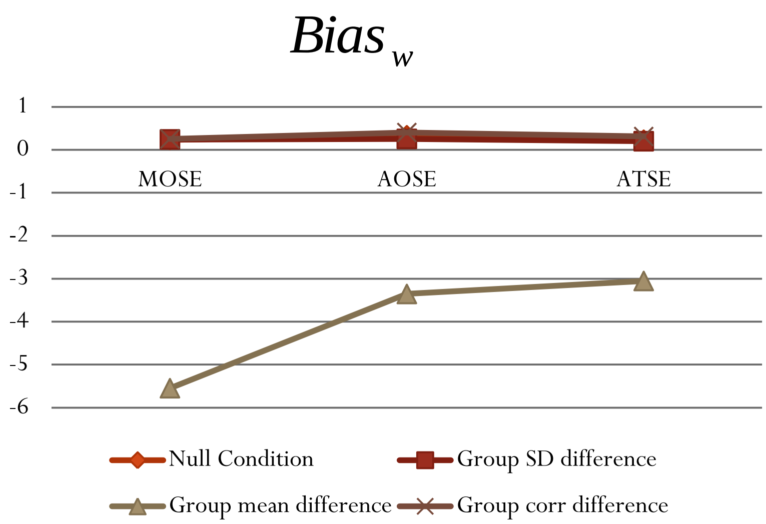


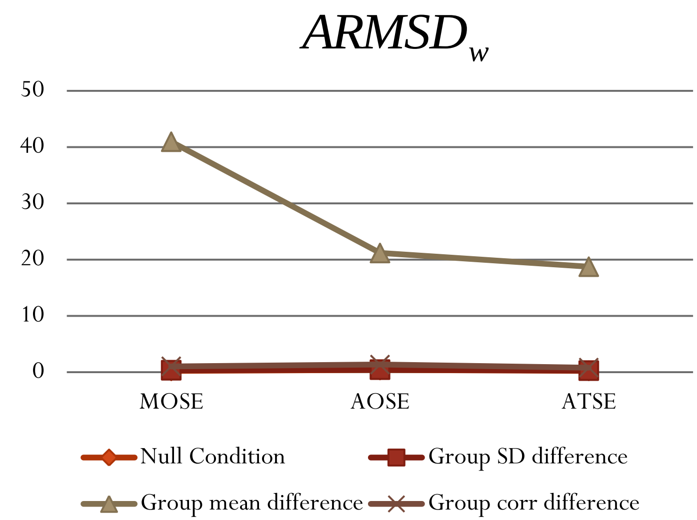
Bias	Min	OD	TCF	ICF	NOP	**ARMSD**	Min	OD	TCF	ICF	NOP
Null Condition	0.32926	0.24153	0.09908	0.08588	0.58308		0.23089	0.28789	0.11543	0.11071	0.69485
Group SD difference	0.29248	0.25659	0.10186	0.07712	0.42949		0.20884	0.48575	0.26935	0.26063	0.55671
Group mean difference	-5.1191	-2.3007	-1.8481	-1.6189	-9.0351		28.6865	7.52435	5.64556	4.9614	87.9546
Group corr difference	0.29875	0.64432	-0.015	-0.0637	0.74563		0.54088	1.33343	2.00157	0.19166	1.24462

- Under the null condition, group SD difference, group corr difference, five MIRT linking methods have similar equating performances
- Under the group mean difference condition, the magnitude of means of *BIASw* and *ARMSDw* for all five MIRT linking methods drastically increased. NOP method performed worst among all 5 Linking methods.

Results (cont.)

- Comparison for the Equating Method x Group Distribution Interaction



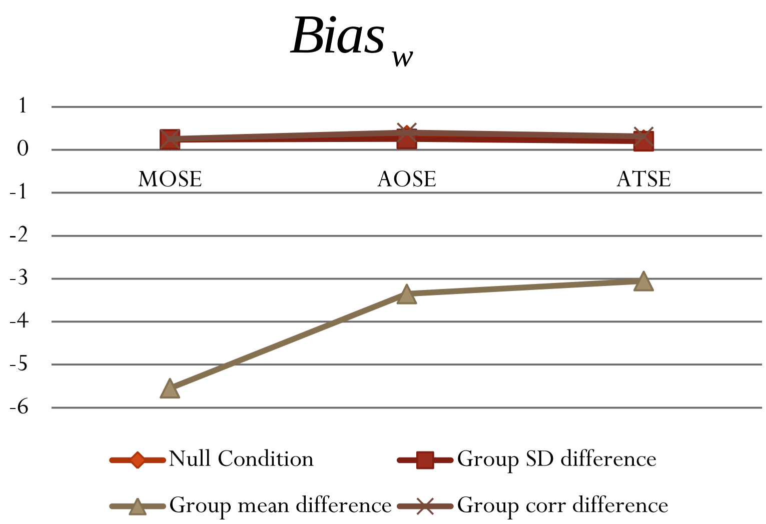


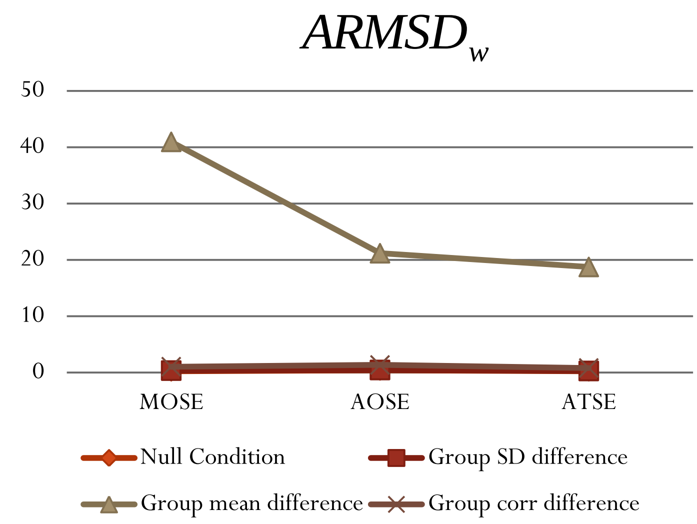
BIASw	MOSE	AOSE	ATSE	***ARMSDw***	MOSE	AOSE	ATSE
Null Condition	0.23211	0.33764	0.23354		0.23195	0.39008	0.24183
Group SD difference	0.24126	0.25349	0.19976		0.34226	0.45045	0.27605
Group mean difference	-5.5469	-3.3541	-3.0522		40.9469	21.1754	18.7411
Group corr difference	0.25488	0.40193	0.30917		1.0374	1.34202	0.80787

- Overall: All three MIRT equating methods performed comparatively well (no group mean difference)

Results (cont.)

- Comparison for the Equating Method x Group Distribution Interaction



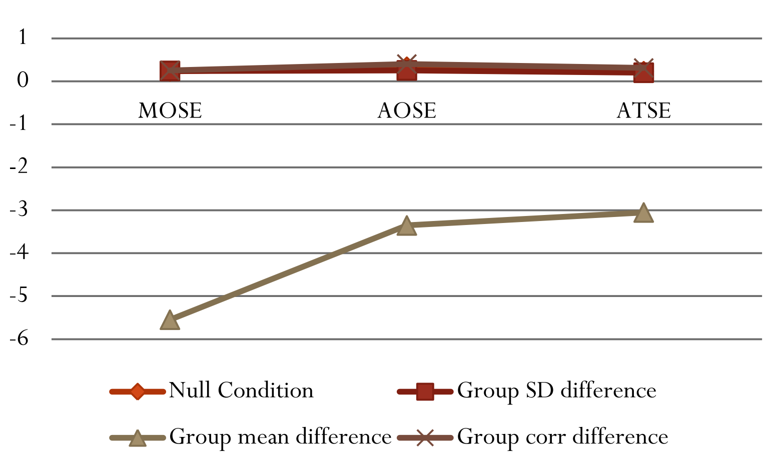


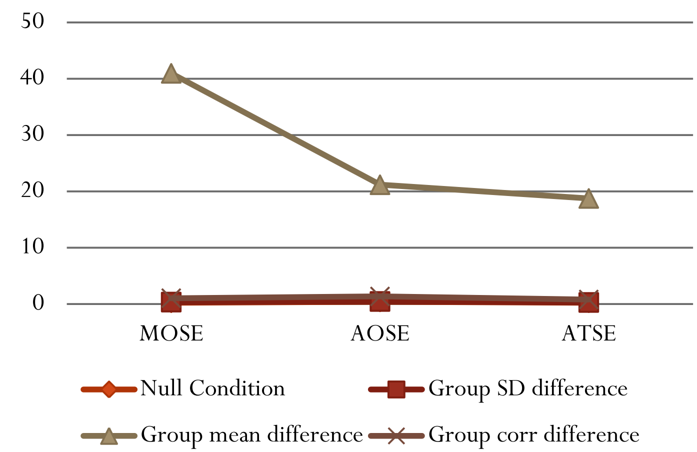
BIASw	MOSE	AOSE	ATSE	***ARMSDw***	MOSE	AOSE	ATSE
Null Condition	0.23211	0.33764	0.23354		0.23195	0.39008	0.24183
Group SD difference	0.24126	0.25349	0.19976		0.34226	0.45045	0.27605
Group mean difference	-5.5469	-3.3541	-3.0522		40.9469	21.1754	18.7411
Group corr difference	0.25488	0.40193	0.30917		1.0374	1.34202	0.80787

- Overall: All three MIRT equating methods performed comparatively well (no group mean difference)
- Equating performance: ATSE > AOSE > MOSE

Results (cont.)

- Comparison for the Equating Method x Group Distribution Interaction





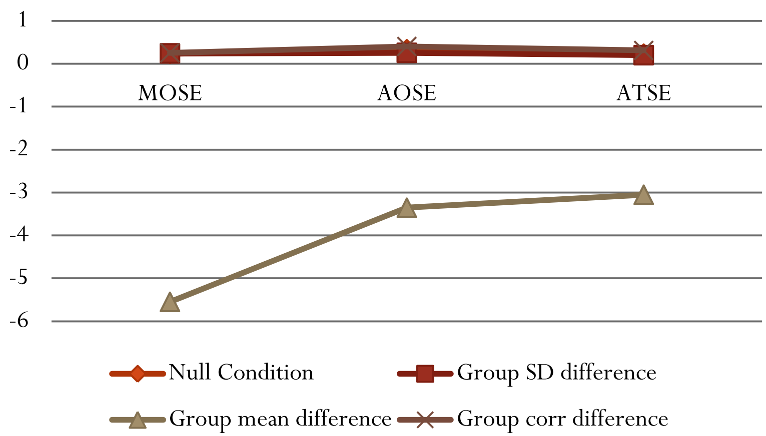
BIASw	MOSE	AOSE	ATSE	***ARMSDw***	MOSE	AOSE	ATSE
Null Condition	0.23211	0.33764	0.23354		0.23195	0.39008	0.24183
Group SD difference	0.24126	0.25349	0.19976		0.34226	0.45045	0.27605
Group mean difference	-5.5469	-3.3541	-3.0522		40.9469	21.1754	18.7411
Group corr difference	0.25488	0.40193	0.30917		1.0374	1.34202	0.80787

- Overall: All three MIRT equating methods performed comparatively well (no group mean difference)
- Equating performance: ATSE > AOSE > MOSE

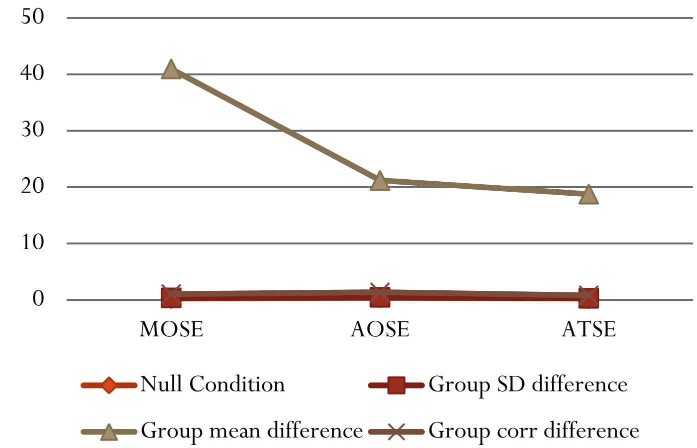
Results (cont.)

- Comparison for the Equating Method x Group Distribution Interaction

$Bias_w$



$ARMSD_w$



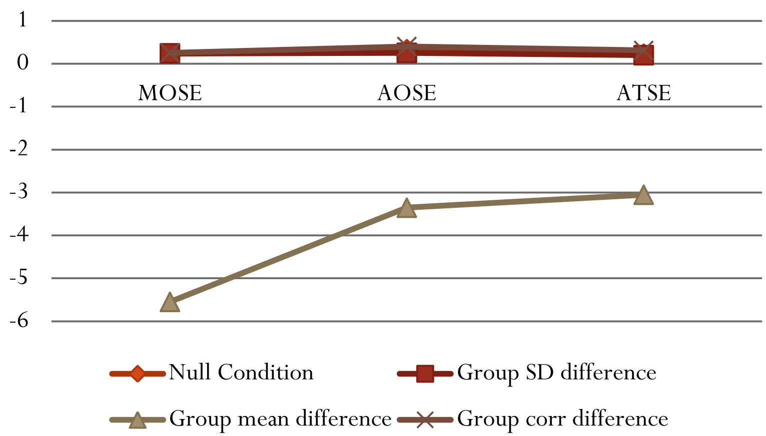
BIASw	MOSE	AOSE	ATSE	***ARMSDw***	MOSE	AOSE	ATSE
Null Condition	0.23211	0.33764	0.23354		0.23195	0.39008	0.24183
Group SD difference	0.24126	0.25349	0.19976		0.34226	0.45045	0.27605
Group mean difference	-5.5469	-3.3541	-3.0522		40.9469	21.1754	18.7411
Group corr difference	0.25488	0.40193	0.30917		1.0374	1.34202	0.80787

- Overall: All three MIRT equating methods performed comparatively well (no group mean difference)
- Equating performance: ATSE > AOSE > MOSE

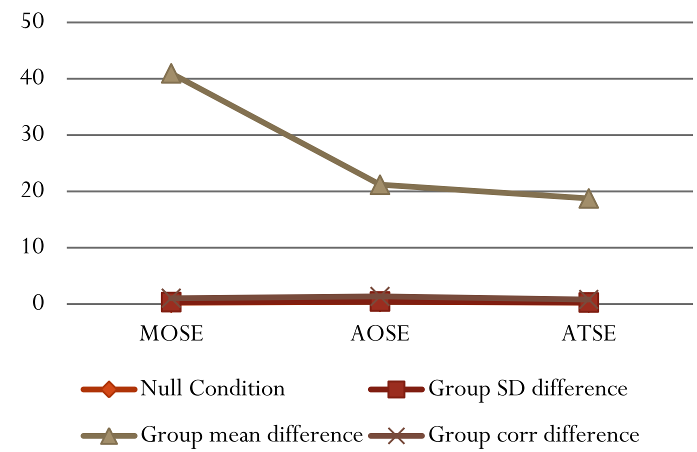
Results (cont.)

- Comparison for the Equating Method x Group Distribution Interaction

Bias$_w$



ARMSD$_w$



BIASw	MOSE	AOSE	ATSE	***ARMSDw***	MOSE	AOSE	ATSE
Null Condition	0.23211	0.33764	0.23354		0.23195	0.39008	0.24183
Group SD difference	0.24126	0.25349	0.19976		0.34226	0.45045	0.27605
Group mean difference	-5.5469	-3.3541	-3.0522		40.9469	21.1754	18.7411
Group corr difference	0.25488	0.40193	0.30917		1.0374	1.34202	0.80787

- Overall: All three MIRT equating methods performed comparatively well (no group mean difference)
- Equating performance: ATSE > AOSE > MOSE
- Under the group mean difference condition, the magnitude of means of *BIASw* and *ARMSDw* for all three MIRT equating methods drastically increased.

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.
  - Group correlation factor and standard deviation factor had a similar level of effect, but not as large as the group mean factor.

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.
  - Group correlation factor and standard deviation factor had a similar level of effect, but not as large as the group mean factor.
- Linking
  - MIRT equating procedures performed best under the TCF and the ICF linking methods (group distribution differences)

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.
  - Group correlation factor and standard deviation factor had a similar level of effect, but not as large as the group mean factor.
- Linking
  - MIRT equating procedures performed best under the TCF and the ICF linking methods (group distribution differences)
  - NOP method had the lowest robustness when there were group distribution shape differences.

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.
  - Group correlation factor and standard deviation factor had a similar level of effect, but not as large as the group mean factor.
- Linking
  - MIRT equating procedures performed best under the TCF and the ICF linking methods (group distribution differences)
  - NOP method had the lowest robustness when there were group distribution shape differences.
  - MIRT equating procedures performed: TCF ICF > OD M > NOP

Discussion and Conclusion

- Test Structure and Group distribution
  - Test structure and all the interactions including test structure had a very small effect on equating results.
  - Group mean factor influenced equating results the most.
  - Group correlation factor and standard deviation factor had a similar level of effect, but not as large as the group mean factor.
- Linking
  - MIRT equating procedures performed best under the TCF and the ICF linking methods (group distribution differences)
  - NOP method had the lowest robustness when there were group distribution shape differences.
  - MIRT equating procedures performed: TCF ICF > OD M > NOP
- Equating
  - ATSE procedure demonstrated, overall, the best equating performance as compared with the other two equating procedures (i.e., MOSE and AOSE) across all group distribution conditions.

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors
- Comparison between MIRT equating methods and UIRT equating methods

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors
- Comparison between MIRT equating methods and UIRT equating methods
- Comparison between MIRT equating methods and observed score equating methods

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors
- Comparison between MIRT equating methods and UIRT equating methods
- Comparison between MIRT equating methods and observed score equating methods
- IRT software Choice-TESTFACT, Mplus, IRTPRO, BMIRT (rotation)

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors
- Comparison between MIRT equating methods and UIRT equating methods
- Comparison between MIRT equating methods and observed score equating methods
- IRT software Choice-TESTFACT, Mplus, IRTPRO, BMIRT (rotation)
- No optimization is involved in the translation in current MIRT linking methods (may not work effectively) - New MIRT linking methods with translation optimization are needed

Limitation and Future Research

- The first simulation study to evaluate the performance of different MIRT equating procedures
- More comprehensive factors
- Comparison between MIRT equating methods and UIRT equating methods
- Comparison between MIRT equating methods and observed score equating methods
- IRT software Choice-TESTFACT, Mplus, IRTPRO, BMIRT (rotation)
- No optimization is involved in the translation in current MIRT linking methods (may not work effectively) - New MIRT linking methods with translation optimization are needed
- Orthogonal rotation vs. oblique rotation in MIRT linking influencing MIRT equating results needs further investigation

Key References

- Brossman, B. G. (2010). Observed score and true score equating procedures for multidimensional item response theory. *Unpublished doctoral dissertation*, University of Iowa. http://ir.uiowa.edu/etd/469.

- Davey, T. C., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *20*, 405-416.

- Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*, 115-138.

- Min, K. S. (2003). The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations. *Unpublished Dissertation*, Michigan State University, East Lansing, MI.

- Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *37*, 357-373.

- Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

- Simon, M. K. (2008). Comparison of concurrent and separate multidimensional IRT linking of item parameters. *Unpublished Dissertation*, University of Minnesota.

Thank you!

ou.zhang@pearson.com