Correcting Between-Group Mean Differences Using a Sum-Score-Based Differential Test Functioning Index

Yutaro Sakamoto[1], Ryuichi Kumagai[2]

[1] Recruit Management Solutions Co., Ltd.
[2] Tohoku University

Original article introducing Index *S*
Sakamoto and Kumagai (2026) *Educational and Psychological Measurement*→

1 Introduction

- When conducting differential item functioning (DIF) analyses in test development, item-level differences and differential test functioning (DTF) must be examined（Drasgow et al., 2018）
- A sum-score-based index for quantifying the magnitude of DTF was recently proposed by Sakamoto and Kumagai (2026).
- Can a raw-score DTF index be used not only to describe differential test functioning, but also to adjust observed between-group mean differences?

2 Index *S* (Sakamoto & Kumagai, 2026)

When there are two subgroups and the data are dichotomous, the procedure for detecting DTF is as follows:

Step 1 Identify non-DIF items to use as anchors..
Step 2 Calculate the total sum score based on the anchor items. Stratify the data by each sum score value and compute the mean difference in total test scores between the two groups (d_L).

$$d_L = (\bar{S}_{RL} - \bar{S}_{FL}) \times W_L.$$

Here, L denotes the stratum, $\bar{S}_{RL}$ and $\bar{S}_{FL}$ represent the mean total test scores of subgroups R and F within stratum L, respectively, and W_L indicates the proportion of examinees belonging to stratum L.

Step3 We propose the following Index S as a new index of the magnitude of DTF.

$$\text{Index } S = \sum_L d_L .$$

3 Simulation Study

- Sakamoto and Kumagai (2026) evaluated recovery using *RMSE* and *Bias* in IRT-based simulations, but did not predefine the true between-group difference in raw-score units.
- In the present study, we conduct a score-based simulation study adapted from Kumagai and Sakamoto (2025) to examine whether Index *S* can recover the true mean difference between two groups.

Step 1. Specify the population distributions of sum scores
- Beta-binomial sum-score distributions were specified for both groups.
- Because their shape, expected value, and variance are controlled by *a* and *b*, the population mean difference can be defined directly.

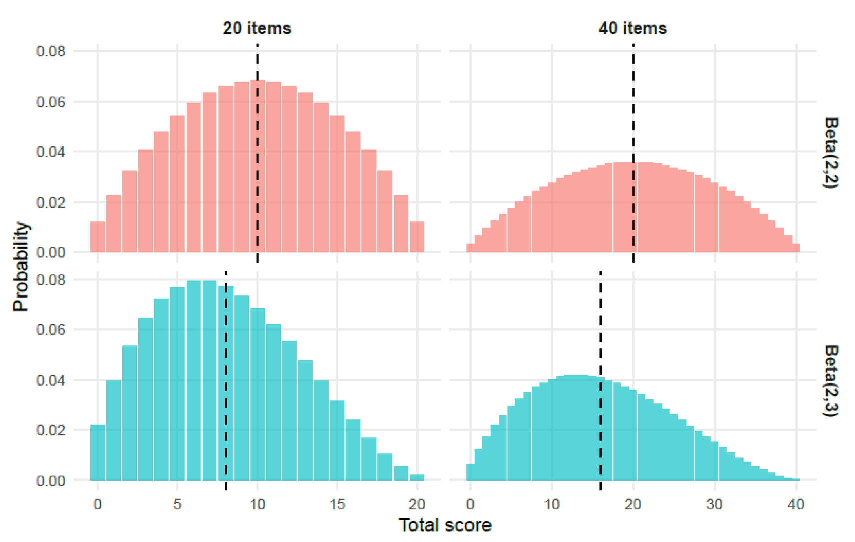

Figure 1. Examples of Beta-Binomial Distributions. Dashed Lines Indicate the Expected Total Scores.

Step 2. Generate sum scores
- Sum-score data were generated for each group using random numbers drawn from the beta-binomial distributions.
- Using these data, the sample mean difference between the two groups can be calculated.

Step 3. Generate dichotomous item-response data
- For each examinee, an item-length zero vector was created, and weighted random sampling was used to change 0s to 1s until the number of 1s matched the generated sum score.
- Item weights were repeated as 0.2, 0.4, 0.6, and 0.8.

Step 4. Introduce test bias through DIF
- DIF items were placed in the latter part of the test, and the remaining items were treated as anchors. For Group 2 DIF items, correct responses were randomly changed from 1 to 0 to reduce the proportion correct by n%, e.g., 10%, yielding the DTF-contaminated mean difference.

Figure 2. Illustration of DIF Manipulation. Items 15–20 are DIF items; Crosses Show Examples of 1-to-0 Response Changes.

Step 5. Compute Index *S* and *STDS* (Meade, 2010)
- Index *S* was computed and subtracted from the DTF-contaminated mean difference to obtain the corrected mean difference.
- For comparison, the IRT-based *STDS* was also computed.
- Steps 2 through 5 were repeated 500 times.

4 Result

Table 1. Summary of the Simulation Conditions.

Factor	Setting
Sample size	N_1/N_2 = 250/250, 250/500, 500/250, 500/500
Score distribution	G1: Beta-binomial ((2,2)); G2: Beta-binomial ((2,2)) or ((2,3))
Test length	20 or 40
DIF proportion	10%, 30%, 50%
DIF manipulation	Group 2 correct responses on DIF items were reduced by 10%
Correction	postAdj_Index *S* = preAdj - Index *S*; postAdj_*STDS* = preAdj - *STDS*

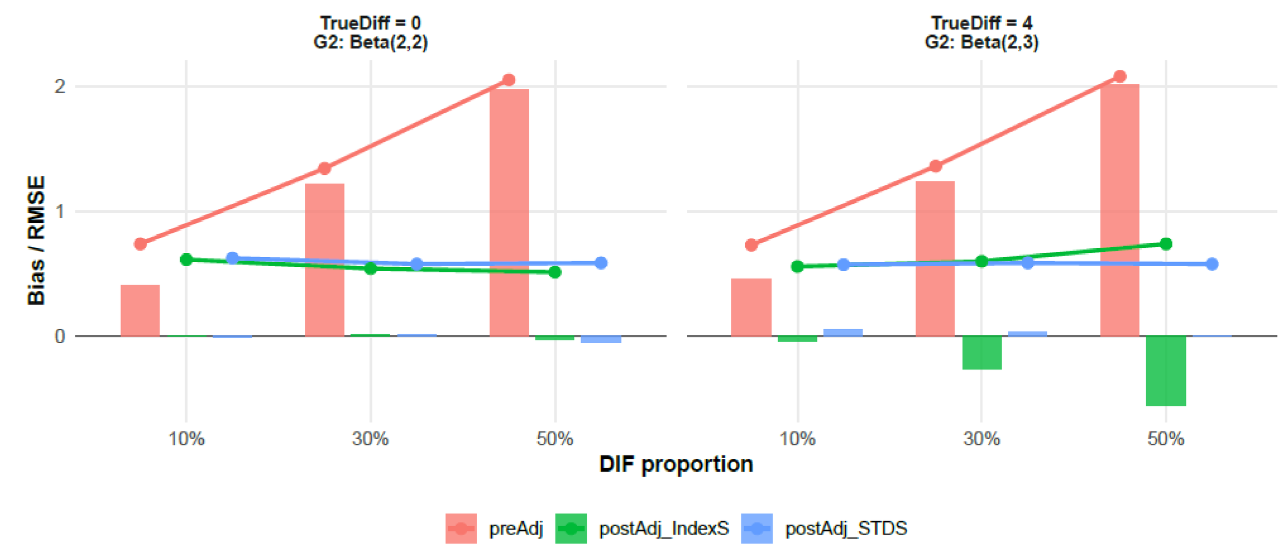

Figure 3. Recovery *Bias* and *RMSE* Across DIF Proportions: 40-Item Test. Bars Represent *Bias*; Lines and Points Represent *RMSE* ($N_R = 500, N_F = 500$).

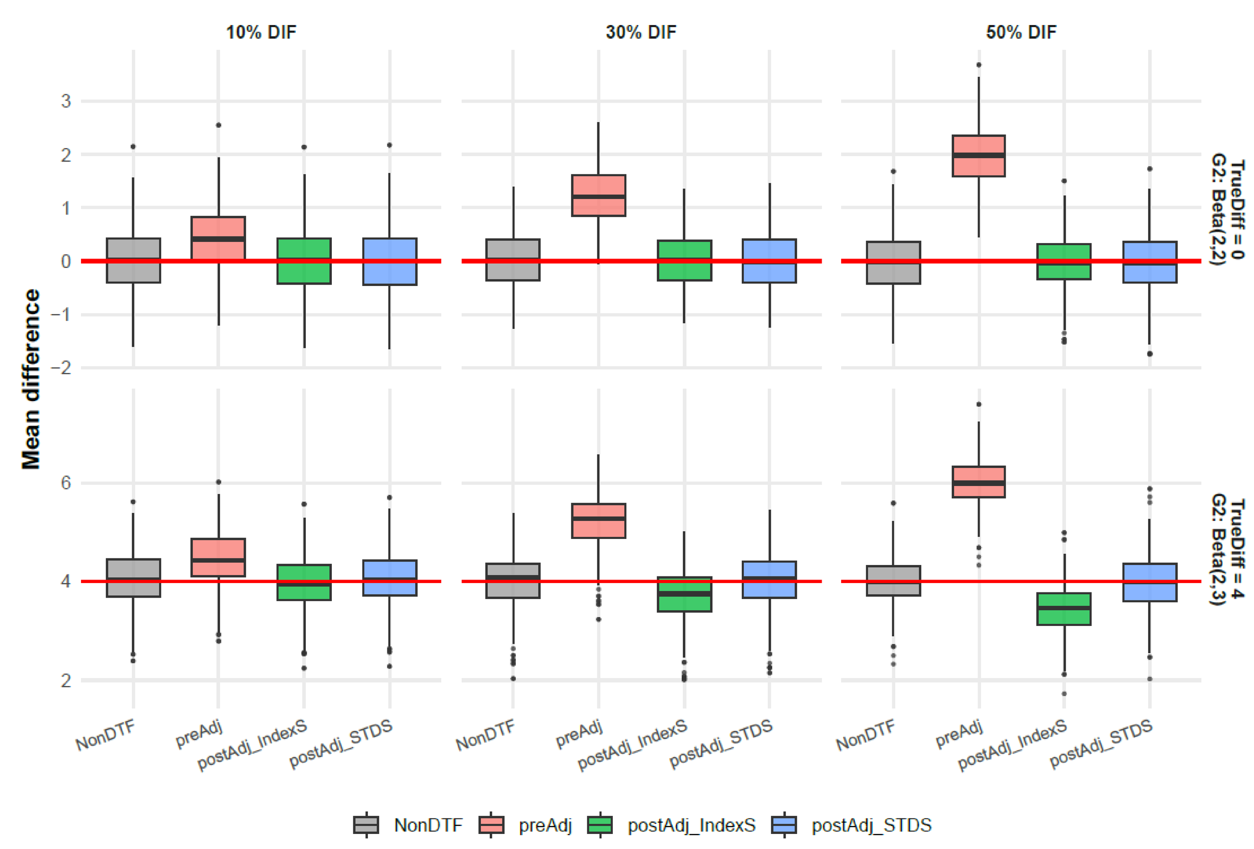

Figure 4. Boxplots of the Four Types of Mean Differences. Red Lines Indicate the Known Population Mean Difference Before DIF ($N_R = 500, N_F = 500$).

5 Discussion

- Across all conditions shown in Figure 3 and 4, Index *S* appeared to adjust the observed mean difference toward the true raw-score mean difference.
- However, when the population distribution of Group 2 was Beta(2,3), the adjustment by Index *S* tended to be slightly larger, indicating that it did not fully recover the true raw-score mean difference in all cases.
- The present study suggests that, even in such settings, Index *S* can function to some extent as a method for estimating the true raw-score mean difference after accounting for DTF effects.
- Future work should extend Index *S* to comparisons involving three or more groups and examine its performance with polytomous items.

References : Drasgow, F., Nye, C. D., Stark, S., & Chernyshenko, O. S. (2018). Differential item and test functioning. In P. Irwing, T. Booth, & D. J. Hughes (Eds.*), The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 885–899). Wiley. ; Kumagai, R., & Sakamoto, Y. (2025). Correction of between-group mean differences in a raw-score-based DTF detection method: A simulation study based on raw scores. *Nihon Kyoiku Shinri Gakkai Sokai Happyo Ronbunshu*, 67, 328. https://doi.org/10.20587/pamjaep.67.0_328 (In Japanese, translated by the authors of this poster.) ; Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology,* 95(4), 728–743. ; Sakamoto, Y., & Kumagai, R. (2026). A simple approach for differential test functioning based on sum scores. Advance online publication. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644261455414.

This work was supported by JSPS KAKENHI Grant Number JP25K06765.