A new method for detecting DTF using sum scores

Yutaro Sakamoto (Recruit Management Solutions Co., Ltd.)
Ryuichi Kumagai (Tohoku University)

When conducting differential item functioning (DIF) analyses in test development, item-level differences and differential test functioning (DTF) must be examined. With the renewed focus on the practical utility and interpretability of sum scores (Sijtsma, Ellis, & Borsboom, 2024), the need for effective DTF detection using these scores has increased. Current DTF detection methods rely on item response theory (IRT), leaving a gap in techniques that utilize sum scores. This study presents a new method for detecting DTF through sum scores. When two examinee groups are to be compared for DTF, first, a DIF analysis is conducted to identify non-DIF items designated as anchor items. Second, the sum score for the anchor items is calculated, and examinees are stratified by the sum scores. Within each stratum, the mean difference in overall test sum scores between the two groups is calculated. This difference is weighted according to the proportion of examinees in each stratum, and the final DTF index, interpreted as an effect size, is derived by summing these weighted values across all strata. The simulation studies demonstrated that the proposed DTF index correlates approximately 0.8 with an existing DTF index (Meade, 2010), affirming its validity. These findings suggest that DTF can be detected without the reliance on complex psychometric models like IRT. Furthermore, the proposed method is straightforward and requires only basic arithmetic operations, making it a practical tool for test developers and practitioners.

Introduction

- **Differential test functioning（DTF）**
  - When conducting differential item functioning (DIF) analyses in test development, item-level differences and differential test functioning (DTF) must be examined（Chalmers, Counsell & Flora, 2016 ; Temel, 2023）
- **Methods for detecting DTF using sum scores**
  - Recently, the practical utility and interpretability of sum scores have been re-evaluated (Sijtsma, Ellis, & Borsboom, 2024).
  - In this context, it is desirable to have methods for detecting DTF using sum scores.
  - One such method is SIBTEST (Shealy & Stout, 1993), which detects DTF using sum scores; however, it targets only the suspect subtest.
  - Ideally, for example, in a test with a maximum score of 20 points, it would be preferable to assess the magnitude of DTF directly on the "20-point scale.“
- **The purpose of the present study**
  - The purpose of this study is to propose a method for detecting DTF using sum scores and an index (Index S) for representing the magnitude of DTF.

Method

- **The proposed method（“Index S”）**
  - When there are two subgroups and the data are dichotomous, the procedure for detecting DTF is as follows:

Step 1 Identify the anchor items, which are items that do not exhibit DIF.
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).
Merge adjacent strata if either group’s sample size falls below N = 10, using one-point intervals of the sum scores.

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

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.

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

This index expresses the magnitude of DTF directly on the test’s original scale, such as a 20-point scale, and differs from existing DTF indices (Shealy & Stout, 1993).

Simulation study

- We assumed two subgroups and dichotomous data, and conducted simulation studies under the conditions shown in Table 1 using the 2PL model.

Table1 Summary of the simulation study

Design factor	Value
Respondents per group	$N_R = N_F = 500$
Replications per condition	500
Test length	20, 40
Proportion of DIF	10%, 30%, 50%
Type of DIF	uniform DIF and non-uniform DIF
Direction of DIF	“all” and “half & half”
Distribution of latent trait	$\theta_R \sim N(0,1)$ and $\theta_F \sim N(0,1)$
	$\theta_R \sim N(0,1)$ and $\theta_F \sim N(-0.5,1)$

Results

- We examined the correlation between Index S and the existing DTF detection index based on IRT, STDS (signed test difference in the sample; Meade, 2010) (Figures 1–4).

Figure 1 Relationship between Index S and STDS
(Test length = 20, Prop of DIF = 30%, Type of DIF = uniform DIF, Direction of DIF =“all”, $\theta_R \sim N(0,1)$ and $\theta_F \sim N(0,1)$)

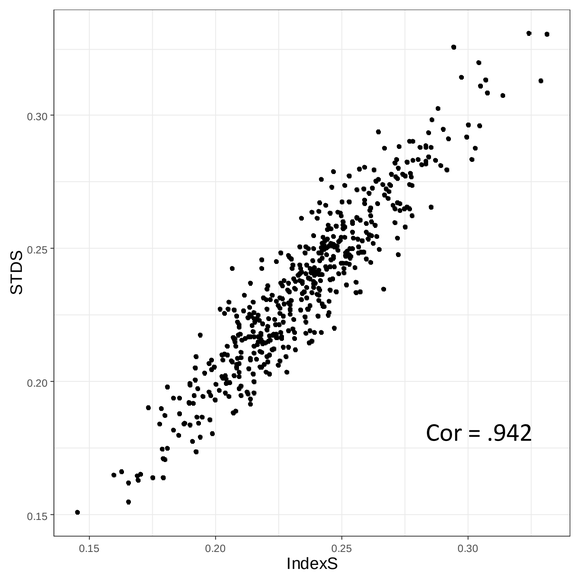


Figure 2 Relationship between Index S and STDS
(Test length = 40, Prop of DIF = 50%, Type of DIF = non- uniform DIF, Direction of DIF =“half & half”, $\theta_R \sim N(0,1)$ and $\theta_F \sim N(0,1)$)

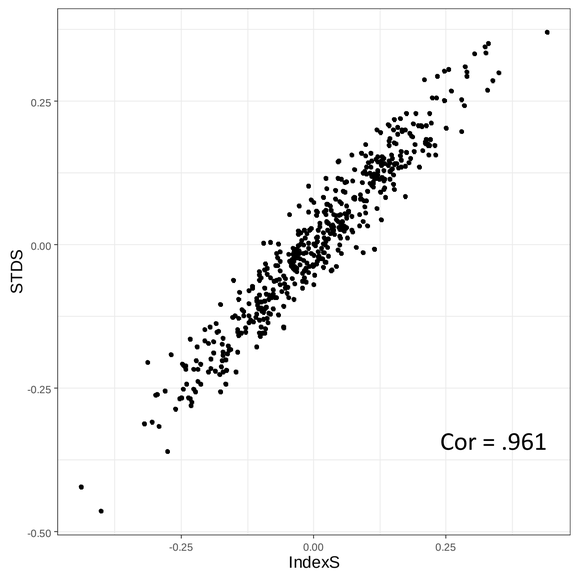


Figure 3 Relationship between Index S and STDS
(Test length = 20, Prop of DIF = 30%, Type of DIF = uniform DIF, Direction of DIF =“all”, $\theta_R \sim N(-0.5,1)$ and $\theta_F \sim N(0,1)$)

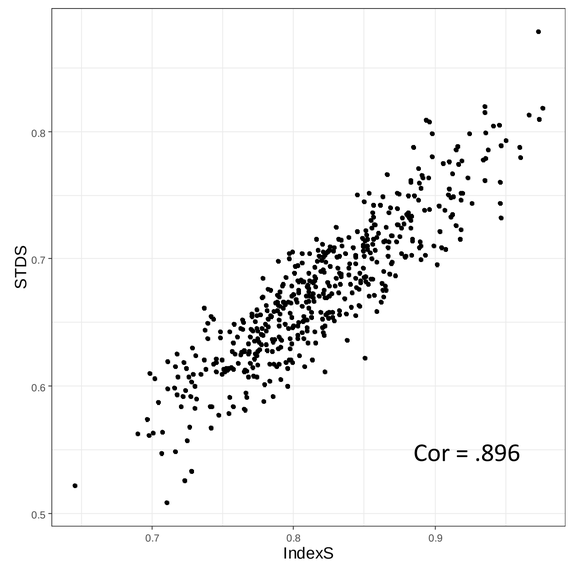
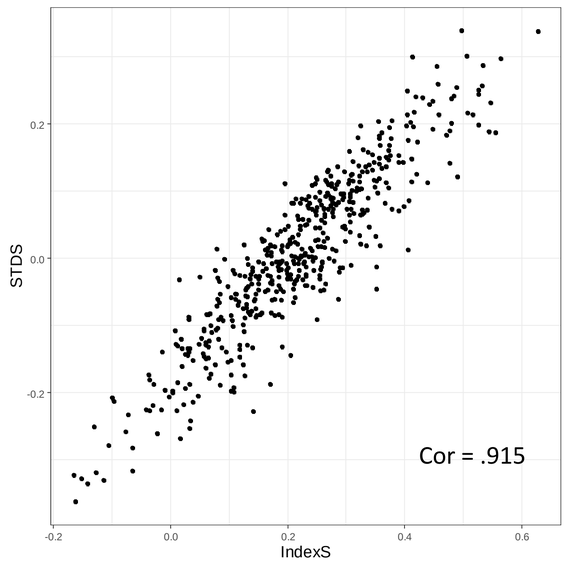


Figure 4 Relationship between Index S and STDS
(Test length = 40, Prop of DIF = 50%, Type of DIF = non- uniform DIF, Direction of DIF =“half & half”, $\theta_R \sim N(-0.5,1)$ and $\theta_F \sim N(0,1)$)



- When the distributions of the latent trait are both standard normal, the correlation between Index S and STDS is very high regardless of the type of DIF (Figures 1 and 2).
- In contrast, when the latent trait distributions differ and the number of test items is small, the correlation between Index S and STDS tends to be lower compared to when the distributions are aligned (Figures 3 and 4).
- ➡ This tendency may be influenced by the reduced number of examinees in each group after stratification based on anchor scores.

Discussion

- When the distributions of the latent trait can both be assumed to follow a standard normal distribution, Index S—although based on sum scores—may perform comparably to IRT-based indices.
  ➡ The proposed method provides practical utility for test developers and practitioners.
- However, such cases are rare, and in practical applications, flexible merging of adjacent strata based on anchor scores may help maintain stable estimates.

- In the future, challenges will include addressing cases where the sign of $\bar{S}_{RL} - \bar{S}_{FL}$ reverses when calculating d_L, as well as generalizing the method to comparisons involving three or more groups.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved Differential Test Functioning statistics that account for sampling variability. Educational and Psychological Measurement, 76(1), 114–140. https://doi.org/10.1177/0013164415584576 ; 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. https://doi.org/10.1037/a0018966 ; Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58(2), 159–194. https://doi.org/10.1007/BF02294572 ; Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the value of the sum score, psychometrics’ greatest accomplishment. Psychometrika, 89(1), 84–117. https://doi.org/10.1007/s11336-024-09964-7 ; Temel, G. Y. (2023). A simulation and empirical study of differential test functioning (DTF). Psych, 5(2), 478–496. https://doi.org/10.3390/psych5020032
This work was supported by JSPS KAKENHI Grant Number JP25K06765.