psych

MDPI

Article

A Simulation and Empirical Study of Differential Test Functioning (DTF)

Güler Yavuz Temel

Faculty of Educational Sciences, University of Hamburg, 20146 Hamburg, Germany; gueler.yavuz.temel@uni-hamburg.de

Abstract: Detecting and understanding DTF is very important under various DIF conditions. In this study, the performance of DTF, DRF, SIBTEST, and CSIBTEST approaches in detecting DIF effects was investigated using a simulation study and a real dataset. It was observed that different DIF conditions (uniform, non-uniform), the proportion of DIF items in the test, the DIF size, and the DIF effect (balanced, unbalanced) affected the performance of the methods, and especially in the case of the non-uniform DIF condition, the power rates of sDTF, sDRF, and SIBTEST statistics were low. In addition, according to the DTF estimations with the balanced/unbalanced DIF effect condition, in some cases, the effect of DIF on the overall test could be negligible. However, it was clearly emphasized in this study that DTF analyses should accompany DIF studies since DTF analysis may change with different DIF and test conditions.

Keywords: differential test functioning; effect size; differential item functioning; SIBTEST; crossing-SIBTEST

Citation: Yavuz Temel, G. A Simulation and Empirical Study of Differential Test Functioning (DTF). *Psych* **2023**, *5*, 478–496. https://doi.org/10.3390/psych5020032

Academic Editor: Okan Bulut

Received: 29 April 2023
Revised: 27 May 2023
Accepted: 31 May 2023
Published: 5 June 2023

1. Introduction

Differential item functioning (DIF), differential bundle functioning (DBF), and differential test functioning (DTF) are important concepts in the field of educational and psychological testing. DIF refers to the situation where an item on a test functions differently for different groups of test-takers, even if they have the same level of ability. DBF occurs when items that contain small amounts of DIF can provide a significant differential effect at the level of clusters or bundles of items at the cumulative level. DTF occurs if true differences on the latent variable are held constant across two or more groups, but the function relating expected test scores to the latent variable differs across groups [1]. Several researchers have devoted considerable attention to the study of DTF because the effects of DIF on overall test scores may be negligible or it might be that there are large DIF effects in one direction for some items, but these effects are canceled out by DIF in the opposite direction for other items [1–6]. In some cases, examining the DTF in isolation or in combination with DIF analyses can be meaningful and informative for test administrators [1]. According to some researchers e.g., [4,5], DTF may be more critical than DIF because decisions are made based on the result of an entire test rather than the result of an individual item. If the DTF appears negligible, in some situations there is no need to discard items with DIF [1,7].

Numerous statistics have been developed to detect DTF using different approaches. Chalmers et al. [1] described the area-based signed (sDTF) and unsigned (uDTF) approaches, which have subtle but important differences compared to sample-based statistics proposed by Raju et al. [8]. Similarly, Chalmers [9] has described the differential response functioning (DRF) statistics (compensatory and non-compensatory statistics) by extending and improving the statistics of area-based DTF measures. DRF provides compensatory differential response functioning statistics (sDRF) and non-compensatory differential response functioning statistics (uDRF) when testing DIF, DBF, and DTF, and provides an effective approach to diagnose conditional differential functioning (i.e., at specific θ values), which are

straightforward for isolated hypothesis testing, as well as graphically presenting the range of differences between the response functions [9]. The DRF statistics can be interpreted as IRT generalizations of the item bias test (SIBTEST) [10,11] and its non-compensatory counterpart, Crossing-SIBTEST (CSIBTEST) [9,12]. The two-step differential item and test (DFIT) statistics are also the most widely used approaches to study DTF [8,13]. However, each of these methods has its own strengths and limitations, and for this reason, different researchers have investigated the DTF estimation performance of the methods with different test conditions by various Monte Carlo simulation studies [1,7,9,14,15]. The main objective of this study is to compare methods, including area-based signed (sDTF) and unsigned (uDTF) approaches [1], DRF statistics [9], and SIBTEST [10,11] and CSIBTEST [9,12], by simulating data from models with known IRT parameters and examining DTF estimation results under different test conditions. Numerous test conditions, DIF conditions and models (e.g., 3PL models), and methods/approaches have been specifically defined for optimal DIF detection using simulation studies. For example, for reference and focus groups, the sample size of the test (balanced/unbalanced, large (e.g., N = 1000 or 2500), or relatively small (e.g., N = 250 or 500)), the length of the test (n = 10, 20, or 40), the DIF direction (balanced/unbalanced), DIF type (uniform, non-uniform DIF), or the DIF size (e.g., small (0.25) large (0.60)). In this study, similar to previous studies [1,9,14,16], we simulated important DIF and test conditions that may have a significant impact on the detection of the DIF effect.

The performance of the sDTF and uDTF statistics was investigated using the 3PL and the graded response model by Chalmers et al. [1]. The performance of the statistics was examined with different test conditions (e.g., sample size, test length, DIF direction, and DIF effect size). The results showed that the performance of both DTF statistics was better with larger sample sizes (e.g., N = 3000), larger DIF effect sizes, and shorter test lengths (e.g., n = 10). Rejection rates were highest when there was DIF in both the slope (a) and the intercept (d) parameters. However, when the slope parameters alone contained DIF, the likelihood of detecting a significant sDTF was five times less than when the d parameters alone contained DIF [1]. The DRF statistics were compared to the SIBTEST and CSIBTEST by Chalmers [9], and the results of these studies indicated that DRF statistics were more optimal effect size estimates of marginal response bias than the SIBTEST family and were the most optimal when studying differential bundle and test bias. The DRF statistics were also compared to the SIBTEST and differential functioning of items and tests (DFIT) frameworks in another study [14]. The results from the DBF and DTF simulations clearly demonstrated that the DRF framework is superior in terms of Type I error control and power to detect DTF and DBF. The results of the study indicated that the DFIT framework demonstrated extremely liberal Type I error rates. The simulation study showed that more than ten anchor items are required to obtain acceptable Type I error rates with SIBTEST, and the dDRF family of statistics provided the highest power in the item bundles studied. Among all studied compensatory statistics, the sDRF family demonstrated the lowest power rates for the type of composite response bias generated. Recently, the performance of the area-based signed (sDTF) and unsigned (uDTF) was compared to a number of effect size statistics, including the DIF variance estimate from a random effects model, both unweighted ($\hat{\tau}^2$) and weighted ($\hat{\tau}^2_w$) [16], Cohen's d for the between-group average DIF [17,18], the average variance associated with DIF from logistic regression $\overline{R^2_\Delta}$ [17,19] and the Doebler´s statistics [20], including a measure of the variance in differences between group difficulty parameters (I^{2*}_{DIF}), and a standardized DIF variance measure (S−DIF−V) for the difference between group difficulty parameters [15]. The results of this study indicated that across study conditions, $\hat{\tau}^2_w$, I^{2*}_{DIF}, and d were consistently the most accurate measures of the DIF effects. Although DTF statistics have been investigated in simulation studies with different test conditions, there are inconsistencies in the performance of DTF estimation statistics for some test conditions. For example, non-compensatory statistics have been proposed as an effective DTF estimation technique when the non-uniform DIF is due to slope parameters only, while inflated Type I error rates have also been emphasized

for these statistics in the non-DIF case. It may be essential to evaluate these statistics in the absence and in the presence of DIF due to slope parameters. Similarly, in the case of unidirectional (unbalanced)/bidirectional (balanced) DIF, it may be important to consider how the power rates of the DTF statistics may differ under different test conditions.

1.1. Statistics for Differential Test Functioning

Chalmers et al. [1] described two area-based statistics for detecting DTF, referred to as sDTF (signed) and uDTF (unsigned) measures of the DTF. These DTF statistics were defined based on the group-specific test score function. These DTF measures were described as follows in Equation (1), Equation (2), and Equation (3), respectively:

$$sDTF = \int [T(\theta, \psi_R) - T(\theta, \psi_F)]g(\theta)d\theta \tag{1}$$

$$sDTF \cong \sum_{q=1}^{Q} [T(X_q, \psi_R) - T(X_q, \psi_F)]g(X_q) \tag{2}$$

$$uDTF = \int [T(\theta, \psi_R) - T(\theta, \psi_F)]g(\theta)d\theta \tag{3}$$

where θ represents the participant's value on the latent variable, $g(\theta)$ is a weighting function with the property that $\int g(\theta)d\theta = 1$, and ψ_R, ψ_F are the reference and focus group item parameters. The sample estimation sDTF is obtained by evaluating the function in Equation (1), where X_q is a quadrature node and $g(X_q)$ is the associated weight. Equation (1) expresses the average amount of test-scoring bias between the response curves and the range from $-$TS to TS, where TS represents the highest possible test score. The negative values indicate that the reference group scored lower than the focus group on average and the positive values indicate that the focus group had lower mean scores. The uDTF captures the average area between the two test curves and ranges from 0 to TS, because the area between the curve equals zero when the test-scoring functions have exactly the same functional form. $uDTF\%$ (Equation (4)) is defined as a suitable standardized effect size metric for better interpretation, and it represents the percent scoring difference for the overall test:

$$uDTF\% = \frac{uDTF}{TS} \cdot 100 \tag{4}$$

1.2. Statistics for Differential Response Functioning

The effect size estimates with DRF were defined by Chalmers [9] as follows:

$$sDRF = \int [S(C|\psi^{(R)}, \theta)S(C|\psi^{(F)}, \theta)]f(\theta)d\theta \tag{5}$$

$$uDRF = \int \left| S(C|\psi^{(R)}, \theta)S(C|\psi^{(F)}, \theta)]f(\theta)d\theta \right| f(\theta)d\theta \tag{6}$$

and

$$dDRF = \sqrt{\int [S(C|\psi^{(R)}, \theta)S(C|\psi^{(F)}, \theta)]^2 f(\theta)d\theta} \tag{7}$$

where $S(\cdot)$ are the scoring equations used to evaluate the model-implied difference between the focal and reference groups. The f(θ) terms can either be estimated from the posterior via an empirical histogram approach (default) or can use the best-fitting prior distribution that is obtained post-convergence (default is a Gaussian distribution). Similar to DTF statistics, compensatory differential response function statistics (sDRF), non-compensatory differential response function statistics (uDRF) and dDRF statistics have been described in Equation (5), Equation (6), and Equation (7), respectively.

Another important issue to consider when determining the DIF effect, or when measuring DTF and DBF statistics, is how differences across response functions should be

aggregated considering the θ parameters. Integrating (or marginalizing) over the latent trait distributions is often carried out to help automate the detection of differential item, bundle, and test effects, rather than focusing on particular θ locations. In the case where the response functions between the groups intersect at one or more locations of θ, practitioners must decide whether the total differential effects should be allowed to compensate across the response functions or whether a definition based on the total magnitude of the difference should be adopted. Based on this framework, it has been argued that, across latent trait distributions, differences between latent trait values can be canceled out if the functions intersect in one or more locations (compensatory) or they are not canceled out (non-compensatory). In the literature, non-compensatory item bias is the primary target for differential function statistics. [14].

Differential response functioning statistics have been obtained as sample-based instantiations of the compensatory and non-compensatory response bias functions and the estimations. These statistics have been described by Chalmers [9] and have been referred to as the compensatory and non-compensatory differential response functioning statistics, or CDRF and NCDRF, respectively. The overall details of the model-based compensatory and non-compensatory bias measures have been discussed by Chalmers [8]. In addition, DRF statistics share a common lineage with a number of previously proposed statistics, such as SIBTEST [11], CSIBTEST [12], and Wainer's [21] standardized impact measures for DIF and DFIT statistics [8,22]. The SIBTEST was designed to detect DIF at both the item and test levels [11], and it may be defined as a bridge between CTT and IRT in that it uses traditional non-parametric statistical mathematics with an IRT approach to analysis [7]. It obtains the marginal $\hat{f}(\theta)$ density by conditioning on the unweighted sum scores from a set of anchors (matching subsets) items, which are restricted to be monotonically related to the measured latent trait in order to compute coefficient α [23]. The compensatory response bias in the studied item set (suspect item set) is approximated after applying a regression-correction technique to equate the observed scores. SIBTEST and CSIBTEST statistics are closely related to the CDRF and NCDRF, respectively. The differences between them are explained by Chalmers [9] as follows: SIBTEST requires only observed scores to be used in computations, while CDRF utilizes fitted probabilistic item response models and the marginal distribution of the latent trait directly. Both CSIBTEST and NCDRF statistics are non-compensatory response bias measures, but they slightly differ in their population definitions. DRF, SIBTEST, CSIBTEST, and DTF statistics were used in this study.

2. Materials and Methods

The aim of this study was to investigate the performance of the effect size measures for the detection of the DTF. A Monte Carlo simulation study was conducted to evaluate and compare the DRF (sDRF, uDRF, dDRF), SIBTEST, CSIBTEST, and DTF (sDTF, uDTF) statistics. Similar to the DIF detection methods, the performance of the DTF methods are influenced by factors such as DIF type (i.e., uniform or non-uniform), DIF magnitude (i.e., low, medium), test length, sample size (i.e., equal or unequal), equal/unequal ability distributions, direction of the DIF (e.g., balanced, unbalanced), and percentages of the DIF items. The performance of DTF statistics was investigated by Chalmers et al. [1] using different sample sizes (500, 1000, and 3000), test lengths (30, 40, and 50 items for the 3PLM design, and 20, 25, and 30 items for the GRM design), and different test conditions, such as DIF types (e.g., DIF in intercepts, in slope, and in both parameters) and DIF directions. The results of this study highlighted the particularly poor performance of DTF statistics in detecting the non-uniform DIF effect. The performance of DRF statistics was compared with SIBTEST and CBISTEST with different test lengths (20, 30, and 40), sample sizes (500 to 3000), number of anchor items (five and ten anchors), and different latent ability distributions [9], and this study emphasized that DRF statistics outperformed SIBTEST and CSIBTEST. Overall, the performance of DTF statistics needs to be examined in terms of different IRT models, especially in terms of 2PL, which is the preferred model for the majority of IRT applications. In addition, the performance of DTF and DRF statistics may

also be examined, especially with respect to the non-uniform DIF effect. In particular, the direction of DIF and the effect of the sample ratio, the differences in ability between the focal and reference groups, should be investigated in the presence and absence of DIF. This study provides a comparison of DTF and DRF statistics with SIBTEST and CSIBTEST using various DIF conditions (balanced/unbalanced DIF), DIF sizes (low, medium), and DIF types (uniform and non-uniform) in a simulation study. This study also provides a comparison of the Type I error rates of the statistics in the absence of DIF conditions. Simulated data were generated using the two-parameter logistic (2PL) model [24]. Similar to the previous studies [1], the intercepts (d) were drawn from a standard normal distribution (d~N (0, 1)), and the slope parameters (a) were drawn from a log-normal distribution (a~log N (0.2, 0.2)). The ability parameters (θ) was set to a standardized normal distribution (~N (0, 1)). In addition, similar to the previous studies [7,15], the group mean differences in the latent trait (impact) were also evaluated in this study. The mean abilities of the reference and focus groups were varied by 0/0 and 0/−1.0 to simulate no and large differences, respectively. The item parameters of the DIF and anchor items are included in the Supplementary Materials (Table S1).

Balanced/unbalanced DIF is an important test characteristic that was expected to impact DTF detection and has been highlighted in several studies [1,7,15,25]. In some balanced DIF cases, the effects of the DIF items on the test-scoring functions cancel out to create a negligible DTF [1]. The sample sizes of the reference and focus groups were simulated at a variety of levels, and equal and unequal sample sizes (R250/F250, R500/F500, R1000/F1000 and R2000/F2000, R1000/250, R1000/F500, R2000/250, R2000/F500) were used to evaluate the DTF methods. In addition to these general test factors, the percentage (0%, 10%, and 20%), magnitude (low (0.25), medium (0.50)), and type of DIF effect (uniform, non-uniform DIF) were simulated. One test length was used in this study (I = 20). The uniform DIF condition was simulated by generating data with the DIF magnitude added to the intercept (d) parameters. The non-uniform DIF condition for the DIF pattern was created by generating data with DIF magnitude added to only the slope parameters (a), without differences in the intercept parameters (d). Two different DIF percentages with 20 items were used in the simulation study. When the DIF percentage was 10%, 18 items were considered as anchor items and 2 items were generated as DIF items. Similarly, for 20% DIF items, 16 items were generated as anchor items and 4 items were generated as DIF items. In the case of unidirectional (unbalanced) DIF and when the DIF percentage was 10%, both items had higher intercept/slope values for the reference group, while in the case of bidirectional (balanced) DIF, the DIF magnitude was added to a single item for both groups, creating a balanced DIF situation.

2.1. Data Analysis and Evaluation Measures

The Type I error and power rates or reject rates (correct identification of DTF) for the statistics were evaluated. In addition to the Type I error and power rates, the mean effect size values and the empirical standard errors for the effect sizes were evaluated to investigate the study outcomes. Similar to the previous studies [1,15], 1000 replications were implemented, and for each combination of the study conditions, the standard deviation of each effect size measure, and likewise, the mean effect size value for each effect size statistic, were calculated across the 1000 parameter estimates for each combination of conditions. The overall analysis of the simulation study was performed using the statistical software R [26], mirt [27], and the doParallel package [28]. The integration range for θ was set to [−6, 6], with 1000 independent imputations. Unless specified otherwise, all detection rates were calculated using a nominal α rate of 0.05, and the liberal Bradley robustness criteria [29] (0.025 < type I error < 0.075) was used to interpret Type I errors.

2.2. Real Data Illustration

The real data used for this study consisted of VERA 8 (Vergleichsarbeiten) [30], a mathematics test that was administered to 8th grade students in Germany. The results of

the VERA tests were used to inform teachers about their students' relative performance in Germany. The DIF effect of the VERA 8 mathematics test (Booklet 2) was examined using the DTF measures that were also used in the simulation study. Language groups (German, non-German (those students whose everyday language is not German)) were used to evaluate DIF at the item and test levels. Since the symmetrical treatment of groups can be advantageous when evaluating DIF across two groups, in the study, weighted sampling was used, and the sample sizes of both the German and non-German students were equal, at 1248. Before the DTF analyses, DIF analyses were performed with multigroup IRT, and the uniform and non-uniform DIF items that were derived from slope, intercept, and both slope and intercept parameters are reported in Table S2, Table S3, and Table S4, respectively.

3. Results

3.1. When No Items Contained DIF

The empirical *p*-values for the statistics were compared to the nominal α values 0.1, 0.05, and 0.01 in the study, and the results of the Type I error rates in the absence of the DIF condition are presented in Table 1. According to the results in the table, the lowest Type I error rates were obtained with sDRF, sDTF, and SIBTEST, and the highest were estimated with uDRF. Table 1 indicates that in the test condition where the mean ability distributions of the reference and focus groups were equal and zero, the Type I error rates for all statistics were not systematically affected by the sample size. In the test condition where the mean ability distributions were not equal, larger Type I error rates were estimated, especially with small and unbalanced sample sizes. However, in the test condition where the sample size was large (R5000/F5000), similar Type I error rates were also estimated from unequal ability distributions for sDTF, sDRF, and uDRF. However, for SIBTEST and CSIBTEST, the Type I error rates were higher for the unequal mean ability distribution condition. When the $p < 0.05$ condition was examined, the Type I error rates for sDTF, sDRF, and SIBTEST were consistently inside the liberal Bradley robustness criteria (0.025 < type I error < 0.075). The inflated type I error rates were obtained for CSIBTEST, especially for uDRF. For uDRF, in particular, these values were not within an acceptable range. In the case of unequal ability distributions, Type I error rates were slightly higher for sDTF, sDRF, and SIBTEST than in the case of equal ability distributions, but still inside the Bradley robustness criteria, whereas Type I error rates for CSIBTEST and uDRF were not inside the Bradley robustness criteria. The mean and standard errors of the DTF effect size statistics in the absence of the DIF condition are provided in Tables 2 and 3, respectively. When the standard error rates in the no DIF condition were examined, it was observed that larger standard errors were estimated with small sample sizes (250/250) and unbalanced sample sizes (e.g., 1000/250, 2000/250) for all the statistics when the ability distributions were equal, and the smallest errors were obtained in the test condition when the sample size was 5000/5000. Standard errors of similar values were estimated for sDRF, sDTF, SIBTEST, and CSIBTEST, and similar values were also estimated for uDTF, uDRF, and dDRF. However, in the test condition where the ability distributions were not equal, the standard errors for all statistics increased.

Table 1. Type I error rates when no DIF items contained DIF.

Ability	N [1]	sDTF			sDRF			uDRF			SIBT			CSIBT		
		$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.10$	$p < 0.05$	$p < 0.01$
$\mu\theta R = \mu\theta F = 0$	250/250	0.10	**0.05**	0.01	0.08	**0.04**	0.01	0.32	**0.19**	0.07	0.12	**0.06**	0.01	0.16	**0.08**	0.02
	500/500	0.09	**0.04**	0.01	0.09	**0.05**	0.00	0.26	**0.16**	0.06	0.1	**0.05**	0.01	0.13	**0.07**	0.01
	1000/250	0.11	**0.06**	0.01	0.1	**0.04**	0.00	0.27	**0.19**	0.05	0.1	**0.06**	0.02	0.15	**0.09**	0.02
	1000/500	0.10	**0.05**	0.00	0.06	**0.02**	0.00	0.24	**0.16**	0.04	0.1	**0.05**	0.02	0.12	**0.06**	0.02
	1000/1000	0.08	**0.04**	0.01	0.16	**0.08**	0.02	0.28	**0.18**	0.09	0.11	**0.05**	0.01	0.14	**0.08**	0.02
	2000/250	0.11	**0.06**	0.01	0.13	**0.06**	0.01	0.32	**0.23**	0.07	0.10	**0.05**	0.01	0.15	**0.08**	0.02
	2000/500	0.08	**0.04**	0.01	0.14	**0.08**	0.02	0.27	**0.17**	0.1	0.12	**0.07**	0.01	0.16	**0.09**	0.02
	2000/2000	0.09	**0.05**	0.01	0.13	**0.06**	0.01	0.3	**0.23**	0.13	0.10	**0.05**	0.00	0.14	**0.07**	0.01
	5000/5000	0.11	**0.05**	0.01	0.14	**0.06**	0.01	0.27	**0.17**	0.06	0.09	**0.04**	0.01	0.12	**0.07**	0.01
$\mu\theta R = 0.0/\mu\theta F = -1.0$	250/250	0.12	**0.06**	0.01	0.11	**0.04**	0.00	0.21	**0.15**	0.07	0.12	**0.07**	0.02	0.17	**0.10**	0.03
	500/500	0.11	**0.06**	0.02	0.12	**0.08**	0.02	0.3	**0.19**	0.04	0.14	**0.08**	0.02	0.19	**0.11**	0.03
	1000/250	0.12	**0.08**	0.03	0.07	**0.04**	0.00	0.21	**0.11**	0.02	0.14	**0.07**	0.02	0.20	**0.12**	0.04
	1000/500	0.12	**0.07**	0.02	0.2	**0.14**	0.04	0.33	**0.22**	0.11	0.14	**0.08**	0.01	0.20	**0.11**	0.03
	1000/1000	0.09	**0.05**	0.01	0.09	**0.04**	0.01	0.28	**0.18**	0.09	0.14	**0.08**	0.02	0.21	**0.12**	0.03
	2000/250	0.13	**0.08**	0.03	0.17	**0.06**	0.01	0.22	**0.15**	0.03	0.14	**0.07**	0.03	0.20	**0.12**	0.05
	2000/500	0.12	**0.06**	0.02	0.11	**0.07**	0.01	0.26	**0.10**	0.02	0.14	**0.08**	0.02	0.19	**0.12**	0.04
	2000/2000	0.10	**0.05**	0.01	0.07	**0.01**	0.00	0.18	**0.12**	0.03	0.14	**0.09**	0.02	0.24	**0.15**	0.05
	5000/5000	0.10	**0.05**	0.01	0.10	**0.03**	0.00	0.25	**0.15**	0.05	0.15	**0.09**	0.02	0.33	**0.22**	0.09

[1] The abbreviation N indicates the sample sizes of the reference and focus groups, respectively.

Table 2. Mean effect size values when no items contained DIF.

Ability	N	sDTF	%sDTF	uDTF	%uDTF	sDRF	uDRF	dDRF	SIBT	CSIBT	seSIBT
μθR = μθF = 0	250/250	**0.003**	**0.014**	**0.09**	**0.448**	**−0.015**	**0.105**	**0.117**	**0.003**	**0.108**	**0.092**
	500/500	0.002	0.01	0.062	0.310	−0.005	0.07	0.079	0	0.07	0.064
	1000/250	**0.004**	**0.019**	**0.071**	**0.354**	**0.011**	**0.079**	**0.088**	**0.003**	**0.082**	**0.073**
	1000/500	0.001	0.005	0.054	0.271	0.002	0.061	0.068	−0.003	0.06	0.055
	1000/1000	−0.001	−0.006	0.043	0.215	0.001	0.052	0.059	0	0.051	0.045
	2000/250	**0.006**	**0.028**	**0.065**	**0.326**	**0**	**0.078**	**0.088**	**0.001**	**0.078**	**0.069**
	2000/500	**0.002**	**0.01**	**0.048**	**0.238**	**0.004**	**0.059**	**0.065**	**−0.002**	**0.059**	**0.05**
	2000/2000	0.001	0.003	0.03	0.152	−0.005	0.038	0.042	−0.001	0.036	0.031
	5000/5000	0	−0.001	0.019	0.097	−0.001	0.022	0.025	−0.001	0.022	0.02
μθR = 0.0/μθF = −1.0	250/250	**0.013**	**0.064**	**0.113**	**0.564**	**0.006**	**0.103**	**0.116**	**0.014**	**0.12**	**0.102**
	500/500	0.009	0.043	0.081	0.407	0.002	0.079	0.09	0.002	0.084	0.07
	1000/250	**0.02**	**0.101**	**0.101**	**0.505**	**−0.006**	**0.092**	**0.105**	**−0.004**	**0.101**	**0.086**
	1000/500	0.015	0.076	0.077	0.386	0.003	0.078	0.089	−0.004	0.077	0.063
	1000/1000	0.004	0.019	0.054	0.271	−0.015	0.053	0.061	−0.003	0.062	0.049
	2000/250	**0.023**	**0.116**	**0.098**	**0.49**	**−0.001**	**0.106**	**0.12**	**−0.007**	**0.101**	**0.085**
	2000/500	0.009	0.046	0.07	0.349	0.015	0.071	0.082	−0.006	0.076	0.062
	2000/2000	0.001	0.007	0.04	0.2	0.005	0.035	0.04	−0.002	0.047	0.034
	5000/5000	0.00	0.00	0.025	0.125	−0.003	0.023	0.026	−0.005	0.035	0.022

Table 3. The empirical standard errors when no items contained DIF.

Ability	N	sDTF	%sDTF	uDTF	%uDTF	sDRF	uDRF	dDRF	SIBT	CSIBT	seSIBT
μθR = μθF = 0	250/250	**0.071**	**0.355**	**0.05**	**0.252**	**0.086**	**0.051**	**0.056**	**0.094**	**0.064**	**0.004**
	500/500	0.049	0.244	0.035	0.175	0.057	0.034	0.038	0.063	0.043	0.001
	1000/250	**0.056**	**0.279**	**0.040**	**0.202**	**0.071**	**0.04**	**0.044**	**0.075**	**0.049**	**0.003**
	1000/500	0.043	0.215	0.029	0.143	0.045	0.03	0.033	0.056	0.036	0.001
	1000/1000	0.034	0.169	0.023	0.117	0.048	0.03	0.033	0.044	0.031	0.001
	2000/250	**0.052**	**0.260**	**0.038**	**0.192**	**0.074**	**0.047**	**0.052**	**0.069**	**0.047**	**0.003**
	2000/500	**0.038**	**0.188**	**0.027**	**0.135**	**0.056**	**0.032**	**0.035**	**0.053**	**0.035**	**0.001**
	2000/2000	0.024	0.118	0.016	0.079	0.03	0.022	0.024	0.031	0.022	0.000
	5000/5000	0.016	0.078	0.010	0.054	0.022	0.014	0.015	0.019	0.013	0.000
μθR = 0.0/μθF = −1.0	250/250	**0.103**	**0.515**	**0.065**	**0.326**	**0.092**	**0.057**	**0.062**	**0.108**	**0.071**	**0.006**
	500/500	0.076	0.380	0.048	0.242	0.071	0.038	0.043	0.077	0.053	0.003
	1000/250	**0.097**	**0.485**	**0.064**	**0.322**	**0.082**	**0.042**	**0.049**	**0.093**	**0.061**	**0.006**
	1000/500	0.075	0.375	0.049	0.244	0.078	0.044	0.05	0.07	0.046	0.003
	1000/1000	0.051	0.257	0.033	0.165	0.047	0.03	0.034	0.054	0.037	0.001
	2000/250	**0.094**	**0.472**	**0.063**	**0.313**	**0.104**	**0.051**	**0.059**	**0.094**	**0.061**	**0.007**
	2000/500	**0.068**	**0.338**	**0.043**	**0.216**	**0.065**	**0.039**	**0.046**	**0.068**	**0.047**	**0.004**
	2000/2000	0.038	0.188	0.024	0.120	0.028	0.018	0.02	0.038	0.029	0.001
	5000/5000	0.023	0.113	0.014	0.069	0.011	0.011	0.012	0.024	0.02	0.000

3.2. The Results of DTF Statistics When the Items Contained Uniform DIF

In this study, in addition to the Type I error rates, the power rates of the statistics for the DTF detection were also calculated. The DTF power rates of the statistics are reported in Table 4 with different test conditions for the uniform DIF condition. In the unidirectional (unbalanced)/ bidirectional (balanced) DIF conditions, we examined the power rates estimated by DTF statistics for different DIF magnitudes and proportions. In the case of unidirectional (unbalanced) DIF, with balanced and large sample sizes, power rates were 1.0 or close to this value and increased for all statistics when the DIF size increased from 0.25 to 0.50 and the DIF proportions increased from 0.10 to 0.20. Additionally, it was observed that the highest power rates (e.g., in the case where the DIF magnitude was 0.25) were predicted by the uDRF statistic, although it varied by sample size. It should be noted again that the uDRF statistic (see Table 1) also provided high rejection rates in the no DIF condition. In the bidirectional (balanced) DIF condition, the obtained power rates resembled the Type I error rates in the no DIF condition and did not change considerably with increasing the DIF magnitude or with increasing the DIF rates. However, it was observed here that the power rates also tended to increase with larger sample sizes.

Table 4. Power rate estimates for DTF statistics when the items contained uniform DIF.

		10% DIF (Two Items Contained DIF)										20% DIF (Four Items Contained DIF)									
		sDTF		**sDRF**		**uDRF**		**SIBT**		**CSIB**		**sDTF**		**sDRF**		**uDRF**		**SIBT**		**CSIB**	
	N	**L** [2]	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**	**L**	**M**
UB [1]	250/250	0.30	0.84	0.38	0.88	0.44	0.86	0.32	0.86	0.34	0.86	0.50	0.99	0.60	1.0	0.63	0.99	0.58	0.99	0.58	0.99
	500/500	0.59	0.98	0.53	0.99	0.54	0.99	0.62	0.99	0.62	0.99	0.85	1.0	0.89	1.0	0.85	1.0	0.88	1.0	0.86	1.0
	1000/250	0.47	0.98	0.49	0.98	0.50	0.97	0.50	0.98	0.50	0.98	0.73	1.0	0.79	1.0	0.81	1.0	0.80	1.0	0.80	1.0
	1000/500	0.73	1.0	0.71	1.0	0.74	0.99	0.71	1.0	0.71	1.0	0.93	1.0	0.93	1.0	0.92	1.0	0.95	1.0	0.95	1.0
	1000/1000	0.90	1.0	0.90	1.0	0.92	1.0	0.90	1.0	0.89	1.0	0.99	1.0	0.99	1.0	0.99	1.0	1.0	1.0	1.0	1.0
	2000/250	0.53	0.98	0.60	0.99	0.60	0.99	0.54	0.98	0.55	0.98	0.79	1.0	0.83	1.0	0.86	1.0	0.82	1.0	0.82	1.0
	2000/500	0.79	1.0	0.82	1.0	0.80	1.0	0.82	1.0	0.82	1.0	0.96	1.0	0.99	1.0	0.99	1.0	0.97	1.0	0.97	1.0
	2000/2000	1.0	1.0	1.0	1.0	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	5000/5000	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
B [1]	250/250	**0.04**	**0.03**	**0.05**	**0.03**	**0.18**	**0.22**	**0.05**	**0.05**	**0.07**	**0.07**	**0.04**	**0.06**	**0.09**	**0.09**	**0.15**	**0.22**	**0.05**	**0.07**	**0.08**	**0.09**
	500/500	**0.05**	**0.05**	**0.07**	**0.05**	**0.18**	**0.20**	**0.05**	**0.07**	**0.07**	**0.09**	**0.07**	**0.11**	**0.05**	**0.10**	**0.20**	**0.16**	**0.06**	**0.07**	**0.08**	**0.08**
	1000/250	**0.04**	**0.04**	**0.02**	**0.05**	**0.17**	**0.19**	**0.05**	**0.05**	**0.06**	**0.07**	**0.05**	**0.08**	**0.02**	**0.09**	**0.13**	**0.23**	**0.06**	**0.05**	**0.08**	**0.07**
	1000/500	**0.05**	**0.06**	**0.07**	**0.07**	**0.16**	**0.26**	**0.05**	**0.05**	**0.07**	**0.08**	**0.06**	**0.12**	**0.04**	**0.06**	**0.18**	**0.25**	**0.06**	**0.06**	**0.08**	**0.09**
	1000/1000	**0.06**	**0.07**	**0.03**	**0.02**	**0.13**	**0.21**	**0.05**	**0.06**	**0.07**	**0.09**	**0.08**	**0.16**	**0.02**	**0.07**	**0.19**	**0.23**	**0.06**	**0.08**	**0.09**	**0.09**
	2000/250	**0.04**	**0.05**	**0.06**	**0.05**	**0.18**	**0.20**	**0.05**	**0.05**	**0.07**	**0.07**	**0.06**	**0.09**	**0.06**	**0.07**	**0.19**	**0.24**	**0.06**	**0.06**	**0.09**	**0.09**
	2000/500	**0.05**	**0.07**	**0.08**	**0.11**	**0.22**	**0.27**	**0.05**	**0.07**	**0.08**	**0.10**	**0.07**	**0.14**	**0.06**	**0.07**	**0.16**	**0.29**	**0.07**	**0.07**	**0.09**	**0.09**
	2000/2000	**0.07**	**0.13**	**0.04**	**0.05**	**0.19**	**0.25**	**0.04**	**0.07**	**0.07**	**0.10**	**0.11**	**0.31**	**0.02**	**0.13**	**0.18**	**0.31**	**0.06**	**0.09**	**0.07**	**0.11**
	5000/5000	**0.09**	**0.23**	**0.06**	**0.05**	**0.18**	**0.32**	**0.07**	**0.10**	**0.09**	**0.13**	**0.21**	**0.67**	**0.04**	**0.11**	**0.15**	**0.52**	**0.06**	**0.12**	**0.09**	**0.14**

[1] UB: unidirectional (unbalanced)/bidirectional (balanced) DIF. [2] L: low DIF size (0.25), M: medium DIF size (0.50).

3.3. The Results of DTF Statistics When the Items Contained Non-Uniform DIF

In this study, the performance of DTF statistics with uniform DIF as well as with non-uniform DIF was analyzed. When the DIF effect was unidirectional (unbalanced), the non-uniform DIF power rates with sDTF, sDRF, or SIBTEST were quite low, and showed no systematic pattern depending on the test conditions (e.g., when DIF magnitude increased (from 0.25 to 0.50) or when DIF proportions increased (from 0.10 to 0.20)). However, uDRF and CSIBTEST statistics yielded power rates consistent with the test conditions. For example, when the sample size, DIF magnitude, and DIF percentage increased, the power rates of the uDRF and CSIBTEST statistics increased, and the power rates obtained with uDRF were higher than the power rates obtained with CSIBTEST. In the test condition where the DIF effect was bidirectional (balanced), power rates were lower in the case of non-uniform DIF, where only the slope parameter contained DIF. For sDTF, sDRF, and SIBTEST, there was no difference in the power rates with increasing DIF magnitude values, except in the test conditions where the sample size was larger (2000/2000 and 5000/5000). In summary, the results in Tables 4–7 indicated that all statistics had very high power to detect a uniform DIF effect and Type I error rates within acceptable ranges in the absence of DIF, especially when the sample size was large (e.g., 1000 to 5000), the DIF size was moderate (0.50), the DIF conditions were unbalanced, and the percentage of DIF items was high (20%) (except CSIBTEST and uDRF, which showed extreme Type I error rates in the absence of DIF). In practice, it can be said that in the presence of an unbalanced DIF effect, almost all statistics provided higher performance rates for the detection of a uniform DIF. However, in the case of a non-uniform DIF effect, the performance of the statistics was quite different. The non-uniform DIF effect could be reasonably detected only with non-compensatory statistics (uDRF and CSIBTEST), even with an unbalanced DIF, large sample size, and a medium DIF size. Researchers may prefer these statistics when examining a non-uniform DIF effect, but it should be noted that it was also identified in this study that these statistics led to higher Type I error rates in the absence of DIF. This was also observed with the real dataset application, as presented below.

Table 5. Power rate estimates for DTF statistics when the items contained non-uniform DIF.

DIF	N	10% DIF (Two Items Contained DIF)										20% DIF (Four Items Contained DIF)									
		sDTF		sDRF		uDRF		SIBT		CSIB		sDTF		sDRF		uDRF		SIBT		CSIB	
		L [2]	M	L	M	L	M	L	M	L	M	L	M	L	M	L	M	L	M	L	M
UB [1]	250/250	0.04	0.04	0.07	0.01	**0.34**	**0.52**	0.06	0.06	**0.10**	**0.20**	0.03	0.04	0.06	0.05	**0.41**	**0.8**	0.05	0.06	**0.14**	**0.39**
	500/500	0.05	0.05	0.07	0.04	**0.51**	**0.86**	0.04	0.05	**0.15**	**0.38**	0.05	0.06	0.05	0.03	**0.65**	**1.0**	0.05	0.06	**0.27**	**0.70**
	1000/250	0.05	0.04	0.05	0.05	**0.45**	**0.80**	0.05	0.06	**0.10**	**0.28**	0.06	0.06	0.03	0.06	**0.65**	**0.96**	0.06	0.06	**0.22**	**0.58**
	1000/500	0.05	0.06	0.06	0.05	**0.57**	**0.93**	0.06	0.06	**0.16**	**0.48**	0.05	0.08	0.05	0.07	**0.73**	**1.0**	0.05	0.06	**0.32**	**0.83**
	1000/1000	0.06	0.06	0.05	0.06	**0.64**	**0.98**	0.05	0.06	**0.26**	**0.72**	0.05	0.07	0.09	0.09	**0.9**	**1.0**	0.06	0.07	**0.50**	**0.97**
	2000/250	0.04	0.05	0.04	0.09	**0.44**	**0.91**	0.04	0.05	**0.11**	**0.30**	0.05	0.04	0.07	0.05	**0.66**	**0.96**	0.05	0.06	**0.22**	**0.62**
	2000/500	0.03	0.06	0.06	0.02	**0.62**	**0.96**	0.04	0.06	**0.21**	**0.54**	0.06	0.08	0.12	0.02	**0.94**	**1.0**	0.06	0.06	**0.40**	**0.90**
	2000/2000	0.06	0.09	0.06	0.06	**0.91**	**1.0**	0.06	0.07	**0.47**	**0.96**	0.06	0.09	0.05	0.09	**1.0**	**1.0**	0.05	0.07	**0.80**	**1.0**
	5000/5000	0.09	0.14	0.03	0.03	**0.99**	**1.0**	0.06	0.07	**0.86**	**1.0**	0.09	0.20	0.06	0.08	**1.0**	**1.0**	0.07	0.10	**1.0**	**1.0**
B [1]	250/250	0.05	0.06	0.04	0.02	**0.18**	**0.16**	0.05	0.05	**0.07**	**0.07**	0.04	0.06	0.07	0.09	**0.16**	**0.16**	0.06	0.04	**0.08**	**0.06**
	500/500	0.04	0.07	0.02	0.05	**0.15**	**0.21**	0.06	0.06	**0.08**	**0.08**	0.05	0.10	0.04	0.02	**0.18**	**0.20**	0.05	0.07	**0.07**	**0.08**
	1000/250	0.06	0.05	0.06	0.06	**0.13**	**0.18**	0.05	0.06	**0.07**	**0.08**	0.03	0.07	0.05	0.05	**0.21**	**0.21**	0.06	0.06	**0.10**	**0.09**
	1000/500	0.06	0.07	0.05	0.03	**0.16**	**0.24**	0.06	0.07	**0.07**	**0.10**	0.05	0.12	0.07	0.02	**0.20**	**0.25**	0.05	0.06	**0.08**	**0.09**
	1000/1000	0.04	0.07	0.02	0.05	**0.14**	**0.19**	0.04	0.06	**0.07**	**0.09**	0.07	0.14	0.11	0.04	**0.25**	**0.21**	0.07	0.07	**0.09**	**0.08**
	2000/250	0.04	0.06	0.07	0.05	**0.18**	**0.20**	0.07	0.05	**0.10**	**0.07**	0.04	0.06	0.05	0.03	**0.33**	**0.19**	0.05	0.07	**0.08**	**0.09**
	2000/500	0.06	0.07	0.02	0.04	**0.19**	**0.17**	0.06	0.06	**0.08**	**0.09**	0.05	0.10	0.03	0.06	**0.21**	**0.17**	0.07	0.08	**0.10**	**0.10**
	2000/2000	0.07	0.13	0.04	0.08	**0.21**	**0.29**	0.07	0.06	**0.09**	**0.09**	0.13	0.24	0.06	0.08	**0.21**	**0.38**	0.05	0.10	**0.08**	**0.12**
	5000/5000	0.13	0.22	0.02	0.12	**0.21**	**0.47**	0.06	0.08	**0.09**	**0.12**	0.23	0.60	0.08	0.13	**0.24**	**0.78**	0.10	0.16	**0.10**	**0.18**

[1] UB: unidirectional (unbalanced)/bidirectional (balanced) DIF. [2] L: low DIF size (0.25), M: medium DIF size (0.50).

Table 6. The mean effect size values and the empirical standard errors for DTF statistics when the items contained uniform/non-uniform DIF.

Stat.	Dif Mag.	Mean Effect Size Values								Empirical Standard Errors							
		Uniform DIF				Non-Uniform DIF				Uniform DIF				Non-Uniform DIF			
		10% DIF		20% DIF		10% DIF		20% DIF		10% DIF		20% DIF		10% DIF		20% DIF	
		UB [1]	B	UB	B	UB	B	UB	B	UB	B	UB	B	UB	B	UB	B
sDTF	**0.25**	−0.032	−0.003	−0.072	−0.009	−0.002	0.003	−0.004	0.009	0.005	0.005	0.007	0.007	0.004	0.004	0.007	0.009
	0.50	−0.064	−0.006	−0.145	−0.018	−0.003	0.005	−0.007	0.015	0.005	0.005	0.008	0.008	0.004	0.004	0.007	0.007
%sDTF	**0.25**	−0.158	−0.015	−0.361	−0.045	−0.01	0.027	−0.021	0.045	0.023	0.023	0.037	0.036	0.022	0.022	0.035	0.035
	0.50	−0.318	−0.028	−0.726	−0.09	−0.017	0.024	−0.036	0.077	0.024	0.024	0.04	0.039	0.021	0.02	0.035	0.035
uDTF	**0.25**	0.032	0.008	0.072	0.016	0.029	0.009	0.073	0.014	0.005	0.004	0.007	0.006	0.006	0.004	0.012	0.006
	0.50	0.064	0.009	0.145	0.022	0.05	0.011	0.125	0.019	0.005	0.004	0.008	0.006	0.006	0.004	0.011	0.005
%uDTF	**0.25**	0.159	0.04	0.362	0.078	0.144	0.044	0.366	0.071	0.023	0.02	0.037	0.032	0.032	0.02	0.059	0.028
	0.50	0.318	0.047	0.726	0.108	0.248	0.054	0.624	0.093	0.024	0.019	0.04	0.028	0.031	0.021	0.054	0.025
sDRF	**0.25**	0.093	−0.004	0.196	0.008	0.001	−0.004	0.004	−0.009	0.013	0.013	0.02	0.018	0.012	0.012	0.02	0.012
	0.50	0.184	−0.007	0.39	0.013	0.005	−0.006	0.006	−0.017	0.013	0.011	0.019	0.021	0.012	0.015	0.017	0.011
uDRF	**0.25**	0.093	0.016	0.196	0.023	0.053	0.016	0.118	0.024	0.013	0.007	0.02	0.01	0.012	0.007	0.02	0.012
	0.50	0.184	0.018	0.39	0.029	0.099	0.019	0.212	0.03	0.013	0.008	0.019	0.013	0.013	0.007	0.017	0.011
dDRF	**0.25**	0.099	0.017	0.203	0.027	0.058	0.018	0.13	0.028	0.013	0.008	0.021	0.011	0.013	0.008	0.022	0.013
	0.50	0.193	0.02	0.404	0.035	0.108	0.022	0.233	0.035	0.014	0.009	0.02	0.013	0.014	0.007	0.019	0.011
SIB	**0.25**	−0.095	0.003	−0.081	0.108	−0.003	0.004	−0.007	0.01	0.013	0.014	0.013	0.013	0.013	0.013	0.021	0.021
	0.50	−0.189	0.008	−0.16	−0.017	−0.005	0.006	−0.012	0.018	0.014	0.013	0.013	0.019	0.013	0.013	0.02	0.020
CSIB	**0.25**	0.095	0.015	0.081	0.108	0.046	0.015	0.098	0.024	0.013	0.009	0.013	0.013	0.013	0.009	0.02	0.014
	0.50	0.189	0.017	0.16	0.024	0.084	0.017	0.181	0.026	0.014	0.01	0.013	0.024	0.013	0.009	0.021	0.015

[1] UB: unidirectional (unbalanced)/bidirectional (balanced) DIF.

Table 7. The results of the real data illustrations.

DTF Measures		Intercept (d)	Slope (a)	Intercept (d) and Slope (a)
sDTF	sDTF	0.074	−0.069	0.075
	p	0.174	0.213	0.303
	CI_97.5	0.183	0.042	0.225
	CI_2.5%	−0.030	−0.171	−0.062
uDTF	uDTF	0.077	0.203	0.468
	CI_97.5	0.183	0.394	0.703
	CI_2.5%	0.038	0.078	0.268
%sDTF	%sDTF	0.142	−0.132	0.143
	CI_97.5	0.352	0.081	0.434
	CI_2.5%	−0.058	−0.329	−0.119
%uDTF	%uDTF	0.077	0.390	0.901
	CI_97.5	0.183	0.759	1.351
	CI_2.5%	0.038	0.149	0.514
sDRF	sDRF	−0.132	0.016	−0.208
	p	0.160	0.591	**0.009**
	CI_97.5	0.048	0.071	−0.050
	CI_2.5%	−0.320	−0.043	−0.359
uDRF	uDRF	0.135	0.050	0.255
	p	**0.035**	0.218	**0.000**
	CI_97.5	0.315	0.157	0.397
	CI_2.5%	0.042	0.014	0.131
dDRF	dDRF	0.151	0.070	0.289
	CI_97.5	0.334	0.198	0.450
	CI_2.5%	0.051	0.026	0.154
SIBTEST	beta	0.113	0.129	0.237
	X2	1.885	4.106	8.504
	p	0.170	**0.043**	**0.004**
CSIBTEST	beta	0.107	0.129	0.237
	X2	2.662	4.106	8.504
	p	0.264	**0.043**	**0.004**

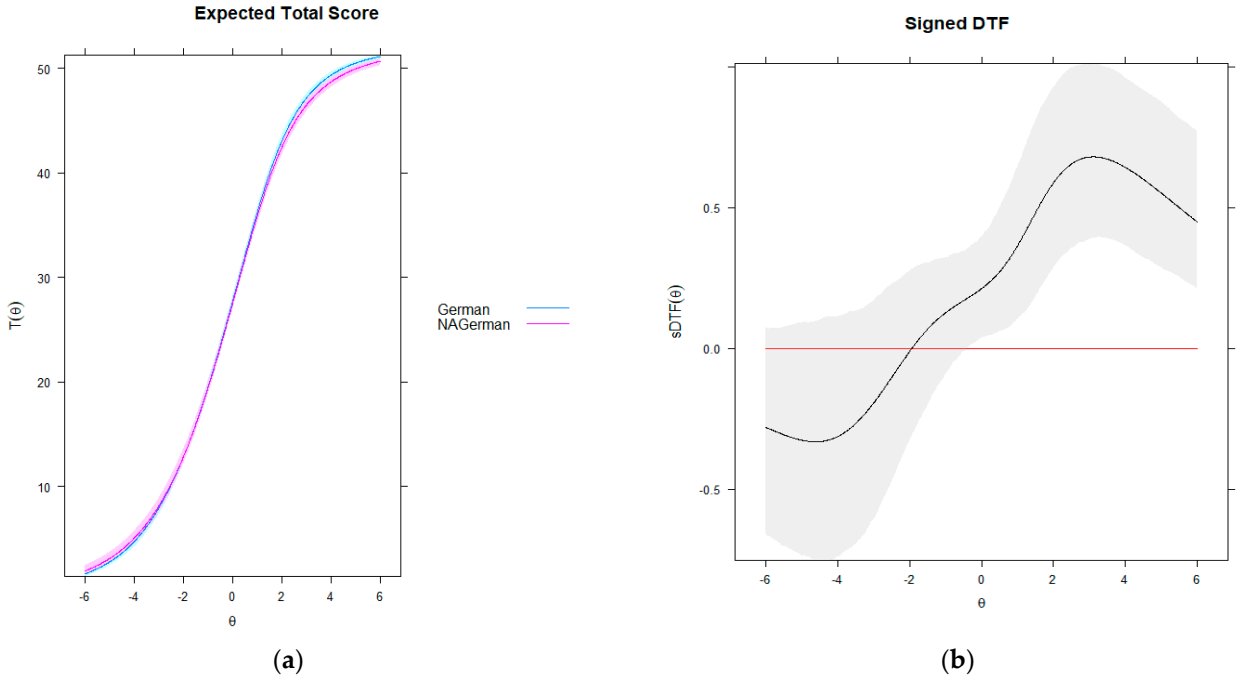
3.4. The Results of Mean Effect Size Values and the Empirical Standard Errors for DTF Statistics When the Items Contained Uniform/Non-Uniform DIF

These results are presented in Table 6. When examining the test condition where the DIF effect was unidirectional (unbalanced), it was found that for all statistics, the mean values of the estimated effect size yielded larger values when the magnitude and percentage of the DIF were larger. However, the increase in all statistics applicable to the uniform DIF was not valid for the non-uniform DIF condition. For example, when examining the mean effect size values estimated with the non-uniform DIF type, the increase was only seen in the values estimated with the uDTF, %uDTF, uDRF, dDRF, and CSIBTEST statistics when the DIF effect was unbalanced. However, when the DIF effect was balanced, similar to the uniform DIF condition, there was no systematic and significant increase in mean effect size values with the increasing DIF magnitude in the non-uniform DIF condition. Empirical standard errors for effect sizes by DIF magnitude and type are also presented in

Table 6. The combination of DIF magnitude and DIF direction had no effect on the empirical standard errors. Since the mean standard errors were estimated in the test condition where the sample size was 5000, equal for the reference and focus groups, they resembled those estimated in Table 3, where the ability distributions were equal, and the sample size was 5000/5000.

3.5. The Results of the Real Data Illustrations

In the last part of the study, the performance of DTF statistics was examined using a real dataset. In the real dataset, the sample sizes were equal, at 1248. In order to determine the effects of uniform and non-uniform DIF conditions on the performance of the DTF statistic, DIF items in the intercept, slope, and both intercept and slope parameters were identified, and the DTF statistics were calculated. When the values reported in Table 7 and Figure 1 were examined, it was first found that the DTF estimated by the DTF statistics (sDTF, uDTF) was not statistically significant ($p > 0$). Similarly, no significant effect was found for either the uniform or non-uniform DIF conditions. This can be clearly seen from the graphs shown in Figure 1. The test characteristic curves (TCC) of both groups are shown in Figure 1a. Since the number of items in the tests was 52, the graph shows the expected total score from 0 to 52. The TCCs can only cross if both statistics have the same value. However, as the graph shows, in this study, the TCCs for German and non-German students crossed. This can be better interpreted using the marked DTF plot together with the confidence limits presented in Figure 1b. The black curve on this graph shows how $sDTF_\theta$ varied with ability, and the 95% confidence limits on sDTF are shown as a grey area, while the red line indicates $sDTF_\theta = 0$. Examining this graph, when θ values were less than zero ($\theta < 0$), the sDTF was negative and the test favored non-German students (focus group), and when θ values were greater than zero ($\theta > 0$), sDTF was positive and the test favored German students (reference group). The situation was also indicated by the crossing TCCs. In contrast to the DTF statistics, the *p*-values in Table 7 indicated that the DIF effect determined by the DRF, SIBTEST, and CSIBTEST statistics was significant. When these statistics were examined, it was observed that both uniform and non-uniform DIF conditions, as determined by the DIF conditions in the intercept and/or slope parameters, had a significant effect.



(**a**) (**b**)

Figure 1. Empirical test-scoring functions with imputed confidence intervals: (**a**) Expected total score and (**b**) signed DTF.

4. Discussion

Methods and statistics for determining DIF/DTF have their own strengths and limitations, as often discussed in the literature. This simulation study examined the various methods and statistics commonly used in DTF measurement under different test conditions and with a real data application. In the first step of the study, the Type I error values in the no DIF case were examined under conditions where the sample sizes and the means of the group ability distributions were different. The results obtained in this part of the study demonstrated that the Type I error rates for all statistics were not affected by the sample size. These results were consistent with the approach described by Chalmers et al. [1]. However, it is important to note that the Type I error rates obtained for uDRF and CSIBTEST were quite high, even in the no DIF condition (e.g., when $p < 0.05$), and consistently outside the liberal Bradley robustness criteria (especially for uDRF). In the no DIF condition, where the means of group ability distributions were different, the Type I error rates for SIBTEST and CSIBTEST increased. In the case where no item contained DIF, standard errors were influenced by differences in the sample size and mean group ability distributions. In particular, values increased with small and unbalanced sample sizes.

In the second part of the simulation study, the effect of DIF items was investigated with different DIF conditions. In the study, uniform and non-uniform DIF were created by simply adding DIF magnitude to the intercept and slope parameters, respectively. In the case of uniform DIF, when the estimation statistics with unbalanced DIF were examined, it was observed that the power rates of the techniques were high and increased as the sample sizes, DIF magnitude, and the number of questions with DIF (percentage) increased. However, in the balanced DIF condition, the power rates did not change systematically depending on the test conditions. The highest power rates were obtained with uDRF, SIBTEST, and CSIBTEST. However, it should be noted that high Type I error rates were also estimated with these statistics in the no DIF condition. However, the performances of sDTF, sDRF, and SIBTEST statistics were not effective when the DIF effect was only due to the slope parameters; in other words, in the case of non-uniform DIF. The power rates estimated with these statistics did not vary systematically with DIF magnitude, proportion, or sample size, and they were also quite low. In contrast, uDRF and CSIBTEST yielded substantial and high-power rates. However, as mentioned before, the Type I error rates of these statistics in the no DIF condition should be considered.

The performance of the DTF statistic, which was also analyzed with a real dataset, was also consistent with the results of the simulation study. One of the main findings was that under the uniform DIF condition (considering only the DIF effect in the intercept parameters), no significant DTFs were estimated with any statistics except uDRF. In the non-uniform DIF condition (DIF in slope parameters only), a significant effect was observed with SIBTEST and CSIBTEST, but the effect was not statistically significant with the other statistics. This is again consistent with simulation studies, especially with the CSIBTEST findings. When we examined the items where the non-uniform DIF effect was due to both the slope and intercept parameters, it was found that the DIF effect was significant at the test level with sDRF and uDRF, and SIBTEST and CSIBTEST.

Considering all the findings in this study, these statistics were generally effective in identifying uniform DIF. However, non-uniform DIF should be investigated with different test conditions. Similar to the effect of DIF at the item level, it is important to investigate the effects of DIF items at the test level and evaluate their performance under different test conditions. In addition, only unidimensional 2PL models were used in this study, and the performance of these models can also be separately examined, especially for non-uniform DIF detection in multidimensional models.

5. Conclusions

The detection of DIF is an important step to improve the psychometric properties of achievement tests and scales developed in many fields (e.g., educational sciences, psychology, medicine) [31]. Many researchers from different fields have developed different

DIF determination techniques for analyzing DIF in test or scale items. Differential item functioning (DIF) analysis refers to procedures that evaluate whether an item's characteristics differ for different groups of examinees after controlling for overall differences in performance. Two types of DIF are often differentiated: uniform DIF and non-uniform DIF [32]. The basic definition of DIF, test fairness, and the basic overview of DIF have been explained in numerous studies e.g., [33–36]. DIF detection has been widely applied in the context of IRT models and observed score approaches. Based on both approaches, various methods for DIF detection have been proposed in the literature. Researchers have aimed to identify optimal DIF detection methods using simulation studies and the application of real datasets.

However, despite the vast literature on methods for detecting DIF, relatively little research has been conducted on the effects of DIF on subsequent outcomes, which is at least as important as the accurate identification of items with DIF. This study examined the performance of commonly used statistics for detecting the DIF effect under different DIF types (uniform, non-uniform), DIF conditions, and test conditions. One of the most important findings of the study was the superiority of non-compensatory statistics in detecting non-uniform DIF, but also their inflated Type I error rates in the absence of DIF. It was therefore recommended that studies be conducted to determine the effect of non-uniform DIF. It was also emphasized that the DIF and test conditions are very important in determining the DIF effect, and it was suggested to consider, in particular, the conditions in which the DIF effect is balanced and unbalanced. Moreover, this study emphasized that similar to DIF analyses, the DIF effect should be determined by researchers. The aggregative effect of DIF items on test scores, or the cumulative DIF effect of item sets, have an important effect on the improvement of the psychometric properties of tests or scales.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/psych5020032/s1, Table S1: The item parameters used for data generation. Table S2: Non-uniform DIF items in a real dataset (slope parameters only). Table S3: Uniform DIF items in a real dataset (intercept parameters only). Table S4: Non-uniform DIF items in a real dataset (both intercept and slope parameters).

Funding: This research was funded by Ministry of Science, Research and the Arts Baden-Württemberg with BRIGITTE-SCHLIEBEN-LANGE-PROGRAMM and Heidelberg University of Education within the framework of internal research funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The empirical datasets analyzed for this study cannot be made available because the dataset is an official dataset that requires approval from the "Institute for Educational Analyses Ba-den-Württemberg (IBBW)" at poststelle@ibbw.kv.bwl.de and the "Institute for Quality Development in Education (IQB)" at https://www.iqb.hu-berlin.de/institut/staff. All R codes used in this study will be available via e-mail in case of interest.

Acknowledgments: I gratefully thank Christian Rietz and Maya Machunsky for their contributions and suggestions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Chalmers, R.P.; Counsell, A.; Flora, D.B. It might not make a big DIF: Improved Differential Test Functioning statistics that account for sampling variability. *Educ. Psychol. Meas.* **2016**, *76*, 114–140. [CrossRef]
2. Drasgow, F. Study of the measurement bias of two standardized psychological tests. *J. Appl. Psychol.* **1987**, *72*, 19–29. [CrossRef]
3. DeMars, C.E. An analytic comparison of effect sizes for differential item functioning. *Appl. Meas. Educ.* **2011**, *24*, 189–209. [CrossRef]
4. Stark, S.; Chernyshenko, O.S.; Drasgow, F. Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *J. Appl. Psychol.* **2004**, *89*, 497–508. [CrossRef]

5. Pae, T.; Park, G. Examining the relationship between differential item functioning and differential test functioning. *Lang. Test.* **2006**, *23*, 475–496. [CrossRef]
6. Flora, D.; Curran, P.; Hussong, A.; Edwards, M. Incorporating measurement nonequivalence in a cross-study latent growth curve analysis. *Struct. Equ. Model. A Multidiscip. J.* **2008**, *15*, 676–704. [CrossRef]
7. Hunter, C. A Simulation Study Comparing Two Methods ff Evaluating Differential Test Functioning (DTF): DFIT and the Mantel-Haenszel/Liu-Agresti Variance. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2014. Available online: https://core.ac.uk/download/pdf/71425819.pdf (accessed on 3 April 2023).
8. Raju, N.S.; van der Linden, W.J.; Fleer, P.F. IRT-based internal measures of differential functioning of items and tests. *Appl. Psychol. Meas.* **1995**, *19*, 353–368. [CrossRef]
9. Chalmers, R.P. Model-Based Measures for Detecting and Quantifying Response Bias. *Psychometrika* **2018**, *83*, 696–732. [CrossRef]
10. Chang, H.-H.; Mazzeo, J.; Roussos, L. DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *J. Educ. Meas.* **1996**, *33*, 333–353. [CrossRef]
11. Shealy, R.; Stout, W. 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* **1993**, *58*, 159–194. [CrossRef]
12. Li, H.-H.; Stout, W. A new procedure for detection of crossing DIF. *Psychometrika* **1996**, *61*, 647–677. [CrossRef]
13. Oshima, T.C.; Raju, N.S.; Flowers, C.P.; Slinde, J.A. Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Appl. Meas. Educ.* **1998**, *11*, 353–369. [CrossRef]
14. Chalmers, R.P. A differential response functioning framework for understanding item, bundle, and test bias. Unpublished. Ph.D. Thesis, York University, Toronto, ON, Canada, October 2016. Available online: https://yorkspace.library.yorku.ca/xmlui/handle/10315/33431 (accessed on 3 April 2023).
15. Finch, W.H.; French, B.F. Effect Sizes for Estimating Differential Item Functioning Influence at the Test Level. *Psych* **2023**, *5*, 133–147. [CrossRef]
16. Camilli, G.; Penfield, R.A. Variance estimation for Differential Test Functioning based on Mantel-Haenszel statistics. *J. Educ. Meas.* **1997**, *34*, 123–139. [CrossRef]
17. Finch, W.H.; French, B.F.; Hernandez, M.F. Quantifying Item Invariance for the Selection of the Least Biased Assessment. *J. Appl. Meas.* **2019**, *20*, 13–26.
18. Hasselblad, V.; Hedges, L.V. Meta-analysis of screening and diagnostic tests. *Psychol. Bull.* **1995**, *117*, 167–178. [CrossRef]
19. Zumbo, B.D.; Thomas, D.R. *A Measure of Effect Size for a Model-Based Approach for Studying DIF (Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science)*; Canada University of Northern British Columbia: Prince George, BC, Canada, 1997.
20. Doebler, A. Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Appl. Psychol. Meas.* **2019**, *43*, 303–321. [CrossRef] [PubMed]
21. Wainer, H. Model-based standardized measurement of an item's differential impact. In *Differential Item functioning*; Holland, P.W., Wainer, H., Eds.; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, USA, 1993; pp. 123–135.
22. Oshima, T.C.; Raju, N.S.; Flowers, C.P. Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *J. Educ. Meas.* **1997**, *34*, 253–272. [CrossRef]
23. Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika* **1945**, *10*, 255–282. [CrossRef]
24. Lord, F.M.; Novick, M.R. *Statistical Theory of Mental Test Scores*; Addison-Wesley: Reading, MA, USA, 1968.
25. Robitzsch, A. Robust and Nonrobust Linking of Two Groups for the Rasch Model with Balanced and Unbalanced Random DIF: A Comparative Simulation Study and the Simultaneous Assessment of Standard Errors and Linking Errors with Resampling Techniques. *Symmetry* **2021**, *13*, 2198. [CrossRef]
26. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022. Available online: https://www.R-project.org/ (accessed on 3 April 2023).
27. Chalmers, R.P. Mirt: A multidimensional item response theory package for the R environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
28. Corporation, M.; Weston, S. *DoParallel: Foreach Parallel Adaptor for the 'Parallel' Package*, R package version 1.0.14; 2018. Available online: https://CRAN.R-project.org/package=doParallel (accessed on 3 April 2023).
29. Bradley, J.V. Robustness? *Br. J. Math. Stat. Psychol.* **1978**, *31*, 144–152. [CrossRef]
30. Schult, J.; Wagner, S. VERA 8 in Baden-Württemberg 2019. In *Beiträge zur Bildungsberichterstattung*; Institut für Bildungsanalysen Baden-Württemberg: Stuttgart, Germany, 2019; Available online: https://ibbw.kultus-bw.de/site/pbs-bw-km-root/get/documents_E56497547/KULTUS.Dachmandant/KULTUS/Dienststellen/ibbw/Systemanalysen/Bildungsberichterstattung/Ergebnisberichte/VERA_8/Ergebnisse_VERA8_2019.pdf (accessed on 3 April 2023).
31. Lee, S. Lord's Wald test for detecting DIF in multidimensional IRT models. Unpublished. Ph.D. Thesis, The State University of New Jersey, New Jersey, NJ, USA, 2015. Available online: https://doi.org/doi:10.7282/T3JH3P13 (accessed on 3 April 2023).
32. Wang, C.; Zhu, R.; Xu, G. Using Lasso and Adaptive Lasso to Identify DIF in Multidimensional 2PL Models. *Multivar. Behav. Res.* **2023**, *58*, 387–407. [CrossRef] [PubMed]
33. Zumbo, B.D. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Lang. Assess. Q.* **2007**, *4*, 223–233. [CrossRef]
34. Camilli, G. Test Fairness. In *Educational Measurement*, 4th ed.; Brennan, R., Ed.; American Council on Education and Praeger: Westport, CT, USA, 2006; pp. 221–256.

35. Penfield, R.D.; Camilli, G. Differential item functioning and item bias. In *Handbook of Statistics*; Rao, C.R., Sinharay, S., Eds.; Routledge: Oxford, UK, 2007; Volume 26, pp. 125–167.
36. Martinková, P.; Drabinová, A.; Liaw, Y.L.; Sanders, E.A.; McFarland, J.L.; Price, R.M. Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE Life Sci. Educ.* **2017**, *16*, rm2. [CrossRef] [PubMed]