International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 319–336
https://doi.org/10.21449/ijate.946978
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An Analysis of Differential Bundle Functioning in Multidimensional Tests Using the SIBTEST Procedure

Didem Ozdogan[1,*], **Hulya Kelecioglu**[2]

[1]Istanbul Kultur University, Faculty of Education, Department of Educational Sciences, İstanbul
[2]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation in Education, Ankara

ARTICLE HISTORY
Received: June 02, 2021
Revised: Jan. 06, 2022
Accepted: Jan. 24, 2022

Keywords:
Differential item functioning,
Differential bundle functioning,
Multidimensionality,
SIBTEST,
Type 1 error,
Power rate.

Abstract: This study aims to analyze the differential bundle functioning in multi-dimensional tests with a specific purpose to detect this effect through differentiating the location of the item with DIF in the test, the correlation between the dimensions, the sample size, and the ratio of reference to focal group size. The first 10 items of the test that is comprised of 30 items were acknowledged as the bundle. The data in line with the parameters were generated via SAS program as two categories (1-0) and multidimensional through an extended 2PL model. Differential bundle functioning was detected via the SIBTEST procedure. The results of the study were interpreted according to the criteria of the power rate and the type I error. When the results were reviewed, the analysis of the bundle revealed that the more the correlation between the two dimensions increased, relatively the less the power rates became. It was observed that the power rates, which were obtained according to two different sample sizes in the study, increased as the sample size increased. Another result as to the SIBTEST's power for detecting DIF was the highest when the ratio of reference to focal group size was equal. According to the results of the type I error rate, the error rate was observed to be relatively decreasing as the correlation between the dimensions increased and it was observed to be increasing as the sample size increased. Also, the highest error rate was obtained when the ratio of the samples was equal.

1. INTRODUCTION

In Classical Test Theory (CTT) and Item Response Theory (IRT), which are used to construct educational and psychological tests and to interpret scores, the assumption is that the individual has a characteristic or ability below the test performance. Uni-dimensional models included in IRT out of these theories comprise a single ability parameter that expresses the location of an individual on the tested characteristic. However, interactions between individuals and items are not generally easy enough to be expressed with these models. Answering a test item or solving a problem generally requires the use of multiple skills and abilities. For this reason, although one-dimensional IRT models are useful under certain conditions, IRT models which reflect the interaction between individuals and test items more accurately are needed. Such IRT models describe the interaction between individuals and characteristics of test items with multiple abil-

*CONTACT: Didem OZDOGAN ✉ reyhandidem@gmail.com Istanbul Kultur University, Faculty of Education, Department of Educational Sciences, İstanbul, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

ities. These models are defined as multidimensional IRT models for having more than one ability parameter for an individual (Reckase, 2009). The fact that only one ability is required for a test item to be answered is not well suited to actual test situations. For example, it is probable to encounter a two-dimensional structure in which a dimension in a mathematical problem reflects mathematical competency whereas the other reflects reading competency. In the related studies conducted it is stated that, besides the primary factors under the test performance of the individual, there is at least one secondary factor and that the tests generally show a multidimensional structure (Camilli, 1992).

Dimension can be defined as a characteristic that affects the probability of answering an item correctly. The main structure the test aims to measure is called as the primary dimension of the test. Since the cause of DIF is defined as the presence of multidimensionality in items showing DIF, such items measure at least one more dimension in addition to the primary dimension the items aim to measure (Cronbach, 1990; Wiley, 1990). In other words, when an item measures more than one dimension and the groups differentiate on the structure or structures that are not primarily measured by the item, the item shows DIF. If the groups do not differentiate on the dimension or dimensions that are not primarily measured, no DIF is detected even if the data are multidimensional (Ackerman, 1992a).

Other dimensions considered as the cause of DIF are called as secondary dimension. The secondary dimensions are the factors that may or may not be related to the dominant dimension. Each secondary dimension intended to be measured is called as the *auxiliary*; whereas each secondary dimension that is not intended to be measured is called as the *nuisance*. DIF caused by the auxiliary dimension is called as benign DIF (which refers to benign effect) because the test also aims to measure the auxiliary dimension. On the other hand, DIF caused by nuisance dimension is called as adverse DIF (which refers to bias) because the item is less valid for one group of individuals than the other for evaluating individual differences on the dimensions measured (Roussos & Stout, 1996).

Shealy and Stout (1993) suggested that DIF occurs due to the presence of two factors; namely,

(1) The item is sensitive not only to the structure θ, which the item aims to measure, but also to a secondary η structure;
(2) At a constant value of θ, there is a difference between the conditional distributions of groups of interest on the η structure.

In general, studies on the detection of DIF are carried out in two stages:
(1) The statistical determination of whether an item provides an advantage for a particular group;
(2) Evaluation of potentially biased items by substantive analysis methods to identify the source of DIF.

Since statistical methods provide limited information in detecting the sources of DIF, new methods have been developed. Roussos and Stout (1996) pointed to the failure of interpretation of substantive DIF analyses followed by statistical methods and Engelhard, Hansche and Rutledge (1990) pointed out in many cases the incompatibility of expert decisions in substantive analyses carried out after statistical methods. Roussos and Stout (1996) proposed a two-step approach by correlating the multidimensional model developed by Shelay and Stout (1993) for DIF to eliminate the inconsistency between statistical and substantive analyses. The first stage of this approach which is called as the multidimensionality-based DIF analysis paradigm is the substantive analysis in which DIF hypotheses are formed and the second stage is the statistical testing of DIF hypotheses. In other words, Shelay and Stout (1993) suggested reversing the process for DIF analyses; they thus stated that reliable and supportive explanations can be obtained about why DIF occurs. Some researchers argue that this problem between statistical and substantive analyses arises from the nature of analyzing individual items in DIF analyses and

that more information can be obtained by analyzing the items in groups rather than analyzing a single item at a time (Boughton et al., 2000; Douglas et al., 1996). Taking this into consideration, Douglas et al. (1996) introduced the DIF in the bundle concept and the applications of Differential Bundle Functioning (DBF) in order to identify the details of DIF.

DBF means that different groups of equal ability levels differ in their probability of correctly answering a bundle. A bundle is a dimensionally homogeneous item set that is not necessarily adjacent or related to a common text (Douglas et al., 1996). Many tests may appear to consist of independent items. However, when carefully analyzed, it might be observed that these items have common topics or similar content. Such items may be found scattered in the test, but they cause the DIF to increase at bundle or test level.

Gierl et al (2001) found that DIF-related properties may be more obvious in a multi-item pattern than a single item. Douglas et al. (1996) stated that the amount of DIF in small amounts at item level (at A-level) could not be statistically detected but that this amount can be detected by the DIF procedure in a bundle. Similarly, Nandakumar (1993) stated that the analysis of a bundle or a group of items would yield stronger results than a single item analysis at a time. Consequently, detecting the potential sources of DIF by identifying groups of biased items by Differential Bundle Functioning (DBF) procedure can provide significant contributions to assessment and test development processes in education (Ross, 2008).

A bundle is formed from items intended to measure a primary dimension (e.g. vocabulary) and a secondary dimension to be measured by the test (e.g. mathematical ability). A premise, which explains why one of the two groups of equal ability levels is more advantageous in the bundle, is developed. The hypothesis developed suggests that one of the groups in a bundle is more advantageous on the secondary dimension than the other group compared (Douglas et al., 1996). Since the bundles are formed based on a hypothesis, DBF analysis can be considered as confirmatory analysis. On the basis of DBF analysis, there is the assumption that a test that measures a particular trait, skill or ability consists of small bundles (Ross, 2008).

Multidimensionality is a concept related to the interaction between an item and its ability. When a test includes items consisting of more than one ability or combination of abilities, several problems might occur if it is not aimed to measure all abilities comprising the test. Ordering individuals accurately according to their abilities primarily requires a valid and reliable measuring. Considering that the items with DIF might weaken validity and that DIF in multidimensional tests is based on the ability differences of individuals on secondary dimension, DIF studies and addressing DIF and multidimensionality together become more important. No study has been found in our country on the concept of DBF and there are few studies outside of our country. The different condition analyzed in this study, different from the literature, is how the location of the items showing DIF in the test affects DBF. In this study, DBF and multidimensionality were analyzed together and tested under various conditions. It is believed that the results obtained would contribute to other studies aiming at detecting the source of DIF.

The purpose of this study is to analyze the DBF concept in multidimensional tests under various conditions. Variables included in the study were the correlations between the dimensions (0.10, 0.45, 0.80), the sample size (2000, 5000), and the ratio of reference to focal group size (1/3, 1/1, 3/1). All these variables were analyzed under the following conditions according to whether the items in the bundle and outside of the bundle were items with DIF.

1) When all items in the bundle show DIF,

SIBTEST power rates according to the conditions are as follows:

a) There are items showing DIF outside of the bundle and
b) There is no item showing DIF outside of the bundle.

2) When the items which show and do not show DIF are present together in the bundle,

SIBTEST power rates according to the conditions are as follows:

a) There are items showing DIF outside of the bundle and
b) There is no item showing DIF outside of the bundle.
3) When there are no items showing DIF in the bundle
a) For the condition in which there are items showing DIF outside of the bundle, SIBTEST Type I error is analyzed.
b) For the condition in which there is no item showing DIF outside of the bundle, SIBTEST Type I error is analyzed.

2. METHOD

2.1. Research Type

In this study, type I error and power rate changes in the bundle were analyzed by differentiating the variables; namely, the location of the item showing DIF in multidimensional tests, the correlations between the dimensions, the sample size, and the ratio of reference to focal group size. Therefore, this study was considered as a simulation study. In this study, the DIF was analyzed with a different approach which is believed to contribute to the theory. In this sense, it can be suggested that this study is a basic research (Karasar, 2020).

2.2. Data of the Study

Within this study, the DBF concept in multidimensional tests was analyzed by considering various conditions in which individuals who apply the tests might encounter in actual test situations. Since it was difficult to realize these conditions in an actual data set at the same time, simulation data were used.

A 30-item test was formed in the study and the first 10 items of this test were considered as the bundle. Type I error and power rates in the bundle were analyzed within the context of the location of the item showing DIF in multidimensional tests, the correlations between the two dimensions, the sample size, and the ratio of reference to focal group size. Item parameters used for generating the research data were generated in ITEMGEN (Ackerman, 1994) program. The item parameters which had equal scattering in this program were formed between the range of the angle values determined by the researcher. The angle values (α) are about the item discriminations. The value of α can vary from 0° to 90° based on the degree to which an item measures each trait. For the case of two dimensions, if an item measures only the first ability, θ_1, then the item direction (α) is 0°. If an item measures only the second ability, θ_2, then the item direction (α) is 90°. When α = 45°, an item measures two abilities (θ_1 and θ_2) equally. Therefore, calculating an item's angular direction (α) provides information about what items are really measuring (Ross, 2007).

When generating item parameters in this study, suitable angle values were determined for items that show and do not show DIF in accordance with the research conditions. It is generally considered in the conditions when the angle value for the items exceeds 20^0. If the secondary dimension is an unintended dimension, a threat to validity will arise (Ackerman, Gierl & Walker, 2003).

The dimension measuring the θ_1 ability in the study was determined as the dimension intended to be measured by the test, whereas the dimension measuring the θ_2 ability was determined as the dimension which is not intended to be measured by the test. The angle values for the items which primarily measure the θ_1 ability and do not show DIF were adjusted to alternate between the range of 5^0-20^0, whereas the angle values for items which primarily measure the θ_2 ability and show DIF were adjusted to alternate between the range of 70^0-85^0. The discrimination parameter of the model is a measure of the differential capability of an item. An item is considered valuable if it well discriminates subjects with ability levels in a range of interest for

the exam (UI Hassan & Miller, 2020). Multidimensional discrimination (MDISC) parameter was generated to alternate between the range of 0.8 and 1.8, whereas item difficulty parameter (d), which is a measure of the ease of the item, was generated to alternate between the range of -2 and 2. The same item parameters were used for reference and focus groups.

Item discrimination power is related to a certain angle value which the item has in latent ability space (Reckase & McKinley, 1991). Calculating this angle provides information about what the item measures. For multidimensional items, this angle can be calculated in terms of the latent axes. Accordingly, α angle alternates between 0^0-90^0. The angle between the discriminant vector and x axis (which refers to the primary dimension which the item aims to measure) can be calculated by Equation 1, which is a simplified version of the formula proposed by Reckase and McKinley (1991) (Ross, 2008).

$$\alpha = tan^{-1}\frac{a_2}{a_1} \quad (1)$$

a_1: The item discrimination power for the primary dimension;

a_2: The item discrimination power for the secondary dimension.

In this study, confirmatory factor analysis was performed in analyzing the accuracy of the structure formed. Also, α angles were checked by using the formula in Equation 1 and parameters were generated in accordance with the location of the item with DIF in the test.

2.3. Simulation Conditions

In accordance with the purpose of the study, simulation conditions determined in the DBF analysis in multidimensional tests are given below. In all conditions, test length, number of items in the bundle, and ability distributions of focus and reference groups were kept constant.

2.3.1. *Test length*

In this study, the length of the test was set at 30 items to represent a mid-length test.

2.3.2. *Number of items in the bundle*

There is one bundle in the test and the first 10 items of the test form the bundle.

2.3.3. *Ability distributions of the groups*

In the study, a test with the items with DIF was formed as θ_1 and θ_2 to measure two dimensions. The characteristic that was intended to be measured by the test was called θ_1, whereas the characteristic that causes the item to show DIF was called θ_2. Accordingly, the data were generated in a way that the items showing DIF would primarily measure the θ_2 dimension, whereas the items which do not show DIF would primarily measure the θ_1 dimension.

The ability distributions of the reference and focal groups for the first dimension were equal and the standard normal distribution was $\theta \sim (N_F(0,1))$ and $\theta \sim (N_R(0,1))$; for the second dimension, a difference of 0.50 was created between the ratio of reference to focal group size, considering the studies in the literature (Ross, 2008; Oshima & Miller, 1992; Russell, 2005, Walker & Şahin, 2016). In the second dimension, the distribution of focus and reference groups was determined as the non-standard normal distribution $\theta(\sim(N_F(-0.25,1))$ and $\theta \sim (N_R(0.25,1))$.

2.3.4. *Test location of item with DIF*

Six conditions were identified for this situation. In all conditions, the items which do not show DIF primarily measure the θ_1 dimension, whereas the items showing DIF primarily measure the θ_2 dimension.

1) All the items in the bundle are items with DIF and there are items showing DIF outside of the bundle: 10 items of the bundle have DIF and 5 test items outside of the bundle have DIF, whereas 15 items do not have DIF.
2) All items in the bundle are items with DIF and there are no items with DIF outside of the bundle: In this condition, 10 items forming the bundle show DIF, whereas 20 items outside of the bundle do not show DIF.
3) In the bundle, there are items with and without DIF, and outside of the bundle, there are items with DIF: 5 items in the bundle have DIF, while 5 items do not have DIF; 5 test items outside of the bundle have DIF, whereas 15 items do not.
4) In the bundle, there are items with and without DIF, and outside of the bundle, there are no items with DIF: 5 items in the bundle have DIF, while 5 items do not have DIF and 20 items outside of the bundle do not show DIF.
5) In the bundle, there are no items with DIF, and outside of the bundle there are items with DIF: 10 items of the bundle do not have DIF and 5 test items outside of the bundle have DIF, whereas 15 items do not.
6) There are no items with DIF in the bundle and outside of the bundle: 10 items forming the bundle do not have DIF; 20 items outside of the bundle do not have DIF.

2.3.5. *The correlation between the primary and secondary dimensions* ($r_{\theta 1 \theta 2}$)

One of the variables analyzed in this study is the effect of the correlations between the primary and secondary dimensions. For this purpose, the correlations between the dimensions were detected as 0.10, 0.45 and 0.80, to represent low, medium, and high correlation values, respectively.

2.3.6. *The sample size*

It is suggested in the literature to work with at least 1000-people samples in multi-dimensional structures (Bolt & Lall, 2003; Yao & Boughton, 2007). Ackerman (1994) stated that multidimensional calibrations require at least 2000 samples. When studies on the subject were reviewed, it was observed that the sample size generally varied between 500 and 5000. Two different sample sizes, 2000 and 5000, were determined for this study.

2.3.7. *The ratio of reference to focal group size (R/F)*

In DIF detection studies conducted, the ratio of reference to focal group size is generally preferred to be equal or close to each other. However, it can also be observed that the ratio of reference to focal group size differs from each other in actual test situations.
Shealy and Stout (1993) stated that at least 250 individuals in each group are required for the SIBTEST procedure. In this study, the reference and focus group sizes were analyzed by differentiating R/F rates, determined as 1/3 (500/1500; 1250/3750), 1/1 (1000/1000; 2500/2500), and 3/1 (1500/500; 3750/1250).

The variables and simulation conditions analyzed in the study are given in Table 1. As seen in Table 1, 108 conditions in total were analyzed: six for the location of the items with DIF, three for the correlation between the dimensions, two for the sample size, and three for the ratio of reference to focal group size (6x3x2x3). Each condition in the study was repeated 100 times. In the literature, it was stated that at least 25 replications are required for simulation studies (Harwell et al.,1996). In this study, a total of 10800 data sets were obtained with 100 replications done for each condition.

Table 1. *The variables and simulation conditions analyzed in the study.*

Variables	Simulation Conditions
Location of the DIF Item	1) All of the items in the bundle are items with DIF and there are items with DIF outside of the bundle 2) All of the items in the bundle are items with DIF and there are no items with DIF outside of the bundle 3) There are items with and without DIF in the bundle and there are items with DIF outside of the bundle. 4) There are items with and without DIF in the bundle and there are no items with DIF outside of the bundle. 5) There are no items with DIF in the bundle and there are items with DIF outside of the bundle 6) There are no items with DIF in the bundle and outside of the bundle
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	1) 0.10 2) 0.45 3) 0.80
Sample size	1) 2000 2) 5000
The ratio of the samples (R/F)	1) 1/3 2) 1/1 3) 3/1

2.4. Analysis and Evaluation Criteria of the Data

In line with the aims of the study, the item parameters were obtained in the ITEMGEN program regarding the location of the items with DIF in the test. According to these parameters, the data were generated in SAS program in accordance with the extended two-parameter logistic model for multidimensionality and two categories (1-0).

Differential Bundle Functioning was identified using the SIBTEST procedure. SIBTEST was developed as an extension of the multidimensional DIF model developed by Shealy and Stout (1993) and is a non-parametric procedure that models the relationship between the latency and item performance measured by the test. After completing the DBF analyses for 108 conditions addressed in the study by using SIBTEST program, SAS program was used to calculate the Type I error and power rates. The effect of the conditions in the study for detecting the DBF was evaluated by the Type I error and power rate criterion. Power rate gives a measure of how accurate the DIF is detected for each item and the bundle using the SIBTEST procedure. It is generally expected that the power rates obtained are equal to and greater than 0.80. However, Type 1 error occurs when the DIF is detected in the item and bundle that do not contain DIF. Generally, in DIF studies, the criterion of type 1 error is 0.05 nominal alpha value (Ross, 2007; Atalay Kabasakal et.al.)

In this study, a variance analysis was also performed to detect how the type I error and power rates obtained regarding the differentiating location of the item with DIF in the test changed according to the conditions studied.

3. FINDINGS

The findings obtained from the analysis of the data generated according to the conditions specified in this section are presented in the context of the location of the item showing DIF in the bundle.

3.1. The Situation When All the Items in the Bundle Show DIF

The situation in which all the items in the bundle show DIF was analyzed in two conditions: 1) there are items with DIF outside of the bundle and 2) there are no items with DIF outside of the bundle. The power of the test was calculated regarding the DIF results according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 2.

Table 2. *Power rates for the conditions in which all the items in the bundle show DIF.*

Condition			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Power Rate	Power Rate
0.10	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1
0.45	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1
0.80	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

In two different conditions (there are items and there are no items which contain DIF outside of the bundle) which were analyzed in the situation that all of the items in the bundle show DIF, SIBTEST detected the DIF in the bundle as 100% correct for all conditions. For this reason, the power rates of the bundle did not differ according to the variables analyzed in the study. All power rates obtained were above the acknowledged limit.

3.2. The Situation When the Items which Show DIF and do not Show DIF are Present Together in the Bundle

The situation when the items which show DIF and do not show DIF are present together in the bundle was analyzed in two conditions: 1) There are items showing DIF outside of the bundle and 2) There is no item showing DIF outside of the bundle. The power of the test was calculated regarding the DIF results according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 3.

Table 3. *Power rates for the conditions in which the items that show DIF and do not show DIF are present together in the bundle.*

Conditions			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Power Rate	Power Rate
0.10	2000	1/3 (500/1500)	0.33	1
		1/1 (1000/1000)	0.45	1
		3/1 (1500/500)	0.41	0.99
	5000	1/3 (1250/3750)	0.62	1
		1/1 (2500/2500)	0.84	1
		3/1 (3750/1250)	0.66	1
0.45	2000	1/3 (500/1500)	0.37	0.97
		1/1 (1000/1000)	0.30	1
		3/1 (1500/500)	0.35	0.97
	5000	1/3 (1250/3750)	0.58	1
		1/1 (2500/2500)	0.76	1
		3/1 (3750/1250)	0.69	1
0.80	2000	1/3 (500/1500)	0.26	0.97
		1/1 (1000/1000)	0.33	0.99
		3/1 (1500/500)	0.28	0.96
	5000	1/3 (1250/3750)	0.57	1
		1/1 (2500/2500)	0.74	1
		3/1 (3750/1250)	0.54	1

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

1) It was observed that the power rates obtained from SIBTEST varied between 0.26 and 0.84 when the items which show DIF and do not show DIF were present together in the bundle, and there are items with DIF outside of the bundle. The minimum power rate (0.26) was observed in the condition when the correlation between the two dimensions was 0.80, sample size was 2000, and sample ratio was 1/3. The largest power rate (0.84) was observed in the condition in which the correlation between two dimensions was 0.10, the sample size was 5000, and the ratio of reference to focal group size was 1/1.

In this condition analyzed regarding the differentiating location of the item with DIF, it was observed that the power rates obtained were detected below the acknowledged limit for other conditions except for one condition.

According to the variables analyzed in the study:

- The average power rates obtained from the bundle at different correlations between the dimensions were 0.55 when $r_{\theta 1 \theta 2}$ =0.10, 0.51 and when $r_{\theta 1 \theta 2}$ =0.45, and 0.45 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the DIF detection power of the SIBTEST in the bundle decreased as the correlation between the dimensions increased. In other words, the DIF in the bundle was more accurately detected using the SIBTEST method in the condition when the correlation between the dimensions was minimum.
- When the average power rates of the bundle at different sample sizes were examined, the results were obtained as 0.34 when the sample size was N=2000 and 0.67 and when the sample size was N=5000. DIF detection power of the SIBTEST in the bundle increased as the sample size increased. Using the SIBTEST procedure, the DIF in the bundle was more accurately detected in the large sample.

• When the average power rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.45 when R/F: 1/3; 0.57 and when R/F: 1/1 and 0.49 when R/F: 3/1. DIF detection power of the SIBTEST in the bundle was higher when the ratio of the reference and focal group size was equal. Using the SIBTEST procedure, the DIF was more accurately detected in the condition when the ratio of reference to focal group size was equal.

2) When the items which show DIF and do not show DIF were present together in the bundle and there was no item with DIF outside of the bundle, it was observed that power rates obtained from the SIBTEST were relatively lower compared to the condition when all of the items in the bundle were items with DIF. The power rates obtained varied between 0.96 and 1. The minimum power rate (0.96) was observed in the condition when the correlation between the two dimensions was 0.80, the sample size was 2000, and sample ratio was 3/1. The maximum power rate (1) was observed in all conditions when the sample size was 5000 and under some conditions when the sample size was 2000.

In this condition analyzed according to the differentiating location of the item with DIF in the test, the obtained power rates were above the acknowledged limit.

According to the variables analyzed in the study:

• The average power rates obtained from the bundle at different correlations between the dimensions were 1 when $r_{\theta 1 \theta 2}$ =0.10; 0.99 and when $r_{\theta 1 \theta 2}$ =0.45, and 0.99 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the highest DIF detection power of the SIBTEST in the bundle was obtained in the condition when the correlation between the dimensions was the lowest. In other words, the DIF in the bundle was most accurately detected using the SIBTEST procedure in the condition when the correlation between dimensions was the lowest.

• When the average power rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.98 when the sample size was N=2000 and 1 when the sample size was N=5000. DIF detection power of the SIBTEST in the bundle increased as the sample size increased. Using the SIBTEST procedure, the DIF in the bundle was more accurately detected in the large sample.

• When the average power rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.99 when R/F: 1/3; 1 when R/F: 1/1 and 0.99 when R/F: 3/1. DIF detection power of the SIBTEST in the bundle was relatively higher when the ratio of the reference to focal group size was equal. Using the SIBTEST procedure, DIF was more accurately detected in the condition when the ratio of reference to focal group was equal.

3.3. The Situation When No Items Show DIF in the Bundle

The situation in which no items in the bundle show DIF was analyzed in two conditions: 1) there are items with DIF outside of the bundle and 2) there are no items with DIF outside of the bundle. The Type I Error of the SIBTEST was calculated according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 4.

Table 4. *Error rates for the conditions in which there are no items showing DIF in the bundle.*

Conditions			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Error Rate	Error Rate
0.10	2000	1/3 (500/1500)	0.33	0.28
		1/1 (1000/1000)	0.46	0.42
		3/1 (1500/500)	0.39	0.35
	5000	1/3 (1250/3750)	0.65	0.74
		1/1 (2500/2500)	0.79	0.80
		3/1 (3750/1250)	0.69	0.62
0.45	2000	1/3 (500/1500)	0.32	0.34
		1/1 (1000/1000)	0.37	0.38
		3/1 (1500/500)	0.36	0.36
	5000	1/3 (1250/3750)	0.64	0.67
		1/1 (2500/2500)	0.66	0.77
		3/1 (3750/1250)	0.53	0.63
0.80	2000	1/3 (500/1500)	0.35	0.37
		1/1 (1000/1000)	0.34	0.31
		3/1 (1500/500)	0.34	0.28
	5000	1/3 (1250/3750)	0.57	0.66
		1/1 (2500/2500)	0.67	0.69
		3/1 (3750/1250)	0.55	0.62

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

1) It was observed that Type I errors in the bundle varied between 0.32 and 0.79 when there are no items which show DIF in the bundle and when there are items which show DIF outside of the bundle. The lowest error rate (0.32) was observed in the condition when the correlation between the two dimensions was 0.45, sample size was 2000, and sample ratio was 1/3. The largest error rate (0.79) was observed in the condition when the correlation between the two dimensions was 0.10, sample size was 5000, and the ratio of reference to focal group size was 1/1.

It was observed that the error rates obtained in this condition were significantly higher than the nominal alpha level (0.05). It is thought that this situation is caused by the fact that the DIF levels increase when the items are analyzed in a bundle, whereas the items do not show DIF when each item is analyzed individually.

According to the variables analyzed in this study:

- The average error rates obtained from the bundle at different correlations between the dimensions were 0.55 when $r_{\theta 1 \theta 2}$ =0.10; 0.48 and when $r_{\theta 1 \theta 2}$ =0.45 and 0.47 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the average Type I error rates in the bundle relatively decreased as the correlation between the dimensions increased. In other words, Type I error rate decreased as the test approached unidimensionality.
- When the average Type I error rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.36 when the sample size was N=2000 and 0.64 when the sample size was N=5000. It was observed that the error rates of the bundle increased as the sample size increased. A lower error rate was obtained in the small sample when detecting DIF in the bundle by using the SIBTEST procedure.
- When the average Type I error rates of the bundle were examined in terms of the ratio of reference to focal group size, the results were obtained as 0.48 when R/F: 1/3; 0.55 and when

R/F: 1/1 and 0.48 when R/F: 3/1. It was observed that the error rates obtained were higher when the ratio of reference to focal group was equal.

2) When there were no items which show DIF in the bundle and the test, it was observed that Type 1 errors of the bundle varied between 0.28 and 0.80. The lowest error rate (0.28) was observed in two different conditions when the sample size was 2000 and the correlation between the two dimensions was 0.10 and the sample ratio was 1/3, and the correlation between the two dimensions was 0.80 and the sample ratio was 3/1. The highest error rate (0.80) was observed in the condition when the correlation between the two dimensions was 0.10 and the sample size was 5000, and the ratio of reference to focal group size was 1/1.

In this condition analyzed regarding the differentiating location of the item with DIF in the test, it was observed that the error rates obtained were significantly higher than the nominal alpha level (0.05). It is thought as in the former sub-problem that this situation was caused by the fact that the DIF levels increase when the items are analyzed in a bundle, whereas the items do not show DIF when analyzed individually.

According to the variables analyzed in the study:

- The average error rates obtained from the bundle at different correlations between the dimensions were 0.53 when $r_{\theta 1 \theta 2}$=0.10; 0.52 and when $r_{\theta 1 \theta 2}$ =0.45 and 0.48 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the average Type I error rates in the bundle relatively decreased as the correlation between the dimensions increased. In other words, Type I error rate of the bundle decreased as the test approached unidimensionality.

- When the average Type I error rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.34 when the sample size was N=2000 and 0.69 when the sample size was N=5000. It was observed that the average error rates of the bundle increased as the sample size increased. A lower error rate was obtained in the small sample when detecting the DIF in the bundle by using the SIBTEST procedure.

- When the average Type I error rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.51 when R/F: 1/3; 0.56 and when R/F: 1/1 and 0.48 when R/F: 3/1. It was observed that the error rates obtained were higher when the sample ratio of the reference and focal groups was equal.

A variance analysis was performed to determine how the power rates of the bundle with DIF and type 1 error rates of the bundle with DIF differ in terms of the variables analyzed in the study. Since the power rates of the bundles did not differ in the conditions when all the items in the bundle contained DIF, a variance analysis could not be performed on these bundles. ANOVA results for the power rates are demonstrated in Table 5 and ANOVA results for the error rates are demonstrated in Table 6.

When the ANOVA results for the power rates in Table 5 were analyzed, the difference of power rates was found significant for the variables: the correlation between the dimensions, the sample size, and the ratio of reference to focal group size in the condition when the items that show DIF and do not show DIF was present together in the bundle and also there were items with DIF outside of the bundle. In this condition, the largest effect size belongs to the sample size. The sample size had a medium effect size over the power rates analyzed in this condition.

The variables, the correlation between the dimensions and the ratio of reference to focal group size, had a small effect size on the power rates. In this condition, only the sample size and sample ratio interaction were significant among the interaction effects between the variables (F=.704, η^2=.005). According to Post-hoc test results, the power rates that belong to the condition in which the correlation between the dimensions was 0.10 were significantly higher than the condition in which correlation was 0.80. No significant difference was found between the correlation 0.45 and other correlations. The power rates which belong to N=5000 value were

significantly higher than the power rates of N=2000 value. The rate R/F: 1/1 was determined to be significantly higher than the ratio of the other two samples. No significant difference was observed between the ratios R/F: 1/3 and 3/1.

Tablo 5. *ANOVA results for the power rates of the bundles.*

		$Bundle_{2a}$		$Bundle_{2b}$	
Effects	Sd	F	η^2	F	η^2
CD*SS*SR	4	.681**	.001	.408	.001
SS*SR	2	4.704**	.005	2.654	.003
CD*SR	4	1.273	.002	.408	.001
CD*SS	2	.124	.0001	2.654	.003
SR	2	9.558**	.009	2.654	.003
SS	1	215.533**	.11	15.309**	.008
CD	2	6.631**	.006	2.654	.003

Notes. **: p<0.05; $Bundle_{2a}$: The condition in which the items that show DIF and do not show DIF are present together in the bundle and also there are items with DIF outside of the bundle; $Bundle_{2b}$: The condition in which the items that show DIF and do not show DIF are present together in the bundle and also there are no items with DIF outside of the bundle; CD= Correlation between Dimensions; SS= Sample Size; SR=Sample Ratio (R/F).

The difference of power rates was found to be significant only for the main effect of the sample size in the condition in which the items that show DIF and do not show DIF were present together in the bundle and there were no items which show DIF outside of the bundle. Sample size had a small effect size on the power rates analyzed in this condition. The power rates when N=5000 were significantly higher than the power rates when N=2000. The power rates of the bundle analyzed in this condition did not make a difference for the interactions between the other main effects and variables.

Tablo 6. *ANOVA results for the type 1 error rates of the bundles.*

		$Bundle_{3a}$		$Bundle_{3b}$	
Effects	Sd	F	η^2	F	η^2
CD*SS*SR	4	.508	.001	.934	.002
SS*SR	2	1.373	.001	1.493	.001
CD*SR	4	.750	.002	.881	.002
CD*SS	2	.792	.0008	.314	.0003
SR	2	4.476**	.005	5.022**	.005
SS	1	150.087**	.08	245.214**	.12
CD	2	5.188**	.005	1.653	.002

Notes. **: $p<0.05$; $Bundle_{3a}$: ANOVA results for the Type I error rate in the condition in which there are no items showing DIF in the bundle and there are items showing DIF outside of the bundle; $Bundle_{3b}$: ANOVA results for the Type I error rate in the condition in which there are no items showing DIF in the bundle and outside of the bundle; CD= Correlation between Dimensions; SS= Sample Size; SR=Sample Ratio (R/F).

When the ANOVA results for the type 1 error rates in Table 6 were analyzed, the error rates were found significant for the variables: the correlation between the dimensions, the sample size, and the sample ratio in the condition in which there were no items showing DIF in the bundle and there were items showing DIF outside of the bundle. In this condition, the largest effect size belongs to the sample size. The sample size had a medium effect size on type 1 error rate analyzed in this condition. The variables, the correlation between dimensions, and the ratio of reference to focal group size, had a small effect size on the error rates. In this condition, the interaction effects between the variables were not found to be significant. According to Post-hoc test results, the error rates that belong to the condition in which the correlation between

dimensions was 0.10 are significantly higher than the error rates with other correlation values. No significant difference was found between the correlations 0.45 and 0.80. The error rates were significantly higher when N=5000 than the error rates when N=2000. The rate R/F: 1/1 was determined to be significantly higher than the ratio of the other two samples. No significant difference was observed between the rates R/F: 1/3 and 3/1.

In the condition in which there were no items showing DIF in the bundle and outside of the bundle, type 1 error rates were found to be significant for the main effects of the sample size and the ratio of the samples. The largest effect size belongs to the sample size. The sample size had a medium effect size on the type 1 error analyzed in this condition. The ratio of reference to focal group size had a small effect size on the type 1 error rate. In this condition, the interaction effects between the variables were not found to be significant. According to Post-hoc test results, R/F: 1/1 value had significantly higher error rates than R/F: 3/1. No significant difference was found in the other paired comparisons of the ratio of reference to focal group size. The power rates were significantly higher when N=5000 than the power rates when N=2000.

4. DISCUSSION and CONCLUSION

In this study, the results of the power rate and type 1 error rate were analyzed regarding the differentiating location of the item with DIF in the test in terms of the correlation between the dimensions (0.10, 0.45 and 0.80), the sample size (2000 and 5000), and the ratio of reference to focal group size (1/3, 1/1 and 3/1). The results obtained are discussed by reviewing the studies conducted on the topic.

In all conditions analyzed regarding the differentiating location of the item with DIF in the test, the power of SIBTEST to detect items with DIF was obtained as the highest in the conditions when the correlation between the two dimensions was the lowest. In general, the power rates of the bundle relatively decreased as the correlation between the dimensions increased. In other words, the power of SIBTEST to detect items with DIF decreased as the test approached unidimensionality. This result is expected when it is considered that DIF is caused by multidimensionality. Ross (2008), in her study, determined the correlation between the dimensions as 0.316, 0.632 and 0.837 and found that the power of SIBTEST to detect DBF relatively decreased as the correlation between the dimensions increased. This finding is similar with the results obtained in this study. In some of the DIF studies conducted at item level, the variation of the power rates is not very explicit in terms of the correlation between the dimensions (Walker & Şahin, 2016; Lee, 2004). All in all, it is possible to state that as the correlation between dimensions decreases, that is, as the test approaches unidimensionality, the power of determining DIF in the item cluster of SIBTEST increases.

Among all conditions analyzed regarding the location of the item with DIF in the test, it was found that the power rates of the bundle increased as the sample size increased. In the studies conducted by Finch (2012), Ross (2008), and Russell (2005), which addressed DIF in various conditions, the researchers analyzed the power rates of the bundle in various sample sizes and found that the power rates of the bundle increased as the sample size increased. These findings are similar with the results obtained in this study. In addition, in some of the DIF studies conducted at item level, the power rates increased as the sample size increased (Awuor, 2008; Lee, 2004; Bolt, 2002; Narayanan & Swaminathan, 1994; & Ackerman, 1992b).

Another condition examined in this study, besides the sample size, is the ratio of reference to focal group size. Among all conditions analyzed regarding the location of the item with DIF in the test, the highest power rates of the bundle were observed in the conditions in which the ratio of reference to focal group size was equal (R/F: 1/1). No clear pattern was observed in the conditions in which the ratio of reference to focal group size was R/F: 1/3 and 3/1. In their studies Finch (2012) and Ross (2008) analyzed DIF in the bundle and found that the power

rates of the bundle were higher in the conditions when the ratio of reference to focal group size was equal. These findings are similar to the findings obtained in this study. In some of the DIF studies conducted at item level, the power rates were found to be higher in the conditions when the ratio of reference to focal group size was equal (Awuor, 2008; Narayanan & Swaminathan, 1994). Sample size and the ratio of sample sizes of focus and reference groups is an important factor in DIF determination studies. It is suggested that the sample size should be at least 1000 in DIF determination methods based on ITC (Shepard et al., 1981), and that there should be at least 250 individuals in each of the focus and reference groups for the SIBTEST method (Shealy & Stout, 1993). Since SIBTEST is a method based on IRT in determining DIF, since the sample size increased, the power to detect DIF in the item cluster of SIBTEST increased as its sample size increased, and the highest power ratios were obtained under the conditions where the focus and reference group ratios were equal.

The study also analyzed the conditions in which there were items with DIF and there were no items with DIF outside of the bundle when there were no items with DIF in the bundle. It was observed in these conditions that the type 1 errors were notably higher than the nominal alpha level. In a study conducted by Russell (2005), the DIF in the bundle was analyzed in terms of the sample size, test length, and DIF size, and it was observed that the type 1 error rates of the bundle were notably higher than the nominal alpha level. In all conditions, the error rates varied between 0 and 47 in 50 repetitions carried out. In a study conducted by Ross (2008), the DIF in the bundle was analyzed in terms of the correlation between the dimensions, the sample size, the sample ratio, item angle, and DIF size.

The error rates obtained in all conditions in the study do not significantly exceed the nominal alpha level (it varies between 0.04 and 0.07). In a study conducted by Finch (2012), the DIF in the bundle was analyzed in terms of the sample size, the sample ratio, test length, and the item rate in the bundle. As a result, the error rates which do not significantly exceed the nominal alpha level were obtained (it varies between 0.053 and 0.072).

In this study, the error rates of the bundle relatively decreased as the correlation between the dimensions increased. In the study conducted by Ross (2008), the error rates did not show a significant difference in terms of the correlation between the dimensions.

Moreover, in the study, it was observed that the error rates increased as the sample size increased when the error rates were analyzed in terms of the sample size. This finding is not similar with the study results found by Finch (2012), Ross (2008), and Russell (2005).

Another important finding concerns the error rates analyzed in terms of the ratio of reference to focal group size in the study as it was observed that the highest error rates were obtained when the sample ratio of the groups was equal. This finding is similar with the results obtained by Finch (2012). However, the error rates of the bundle did not differ according to the sample ratio of the groups in the study conducted by Ross (2008).

In the study, the DIF detection power of SIBTEST was the highest when all the items forming the bundle showed DIF. When all the items forming the bundle showed DIF, the condition in which there were items with DIF or no items with DIF outside of the bundle in the rest of the test had no effect on the power rates obtained at bundle level.

In this study, the presence of items which do not show DIF in the bundle caused the power rates obtained at bundle level to slightly decrease. However, when the items which show and do not show DIF in the bundle were present together and there were no items which show DIF outside of the bundle, the power rates obtained at bundle level did not fall below the acknowledged limit. On the other hand, the power rates obtained at bundle level fell below the acknowledged limit when the items which show and do not show DIF in the bundle were present together and there were items which show DIF outside of the bundle.

In another condition analyzed in this study, it was observed that the error rates of the bundle were notably higher than the nominal alpha level when there were no items with DIF in the bundle and there were items with and without DIF outside of the bundle. This may be caused by the fact that the items do not show DIF when analyzed individually but they can cause DIF in high levels when they form a bundle.

In DIF studies, DIF does not occur in the item if the item is responsive to the secondary dimension and the ability distributions of the groups do not differentiate on the secondary dimension. Again, DIF does not occur in the item if the ability distributions of the individuals differentiate on the secondary dimension and the item is not responsive to the secondary dimension (Shealy & Stout, 1993; Ackerman, 1992b). In this study, the situations when the bundle and the test did not contain DIF were set by creating conditions in which the individuals had different ability distributions; however, the item was not responsive to the secondary dimension. This condition was assured with the angle values of the item. It was preferred due to one of the conditions analyzed in the study, which is, there are items which contain DIF outside of the bundle, whereas there are no items which contain DIF in the bundle. In future research for the conditions without DIF, the items can be created multidimensional and the error rates can be analyzed in the conditions created without differentiating the abilities of the individuals in the secondary dimension.

In the study, the first 10 items of the test were acknowledged as the bundle. In future research, the effect of the items, which form the bundle and are scattered in the test, for detecting DBF can be studied. Another condition for the items showing DIF in the study, the difference in ability distribution between the reference and focal groups, was kept constant in all conditions. In future research, the effect of the different ability distributions between the reference and focal groups for detecting DBF can be studied. Another suggestion is about the procedure. In this study, the DIF at bundle level was detected using SIBTEST; in future research, the DIF at bundle level can be studied using such procedures as MIMIC and DFIT.

Acknowledgments

This paper was produced from the first author's doctoral dissertation prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Didem Ozdogan: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Hulya Kelecioglu**: Methodology, Supervision, and Validation.

Orcid

Didem Özdoğan https://orcid.org/0000-0002-6631-3996
Hülya Kelecioğlu https://orcid.org/0000-0002-0741-9934

REFERENCES

Ackerman, T.A. (1992a). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Ackerman, T.A. (1992b). *An investigation of the relationship between reliability, power, and the type I error rate of the Mantel-Haenszel and simultaneous item bias detection*

procedures. Paper presented at the National Council on Measurement in Education (April 21-23), San Fransisco, CA. https://eric.ed.gov/?id=ED344937

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Atalay Kabasakal K., Arsan N., Gök, B., & Kelecioğlu H. (2014). Değişen madde fonksiyonunun belirlenmesinde mtk olabilirlik oranı sibtest ve mantel-haenszel yöntemlerinin performanslarının (i. tip hata ve güç) karşılaştırılması [Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning]. *Kuram ve Uygulamada Eğitim Bilimleri. 6*(14), 2175-2194. https://doi.org/10.12738/estp.2014.6.2165

Awuor, R.A. (2008). *Effect of unequal sample sizes on the power of dif detection: an irt based monte carlo study with SIBTEST and mantel-haenszel procedures.* [Doctoral dissertation, Virginia Polytechnic Institute and State University]. https://vtechworks.lib.vt.edu/handle/10919/28321

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141. https://doi.org/10.1207/S15324818AME1502_01

Bolt, D.M., & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414. https://doi.org/10.1177%2F0146621603258350

Boughton, K.A., Gierl, M.J., & Khaliq, S.N. (2000). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Paper presented at the Canadian Society for Studies in Education (May 24-27), Edmonton, Alberta, Canada. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.385.5167&rep=rep1&type=pdf

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129-147. https://doi.org/10.1177%2F014662169201600203

Cronbach, L.J. (1990). *Essentials of psychological testing* (5 ed.). Harper & Row.

Douglas, J.A., Roussos, L.A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Engelhard, G., Hansche, L., & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*(4), 347-360. https://doi.org/10.1207/s15324818ame0304_4

Finch, W.H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59. https://doi.org/10.1177%2F0146621611432863

Gierl, M.J., Bisanz, J., Bisanz, G.L., Boughton, K.A., & Khaliq, S.N. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36. https://doi.org/10.1111/j.1745-3992.2001.tb00060.x

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177%2F014662169602000201

Mahmood UI.H., & Frank, M. (2020). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics Simulatıon and Computation.* 1-21. https://doi.org/10.1080/03610918.2019.1705344

Karasar, N. (2020). *Bilimsel araştırma yöntemi, Kavramlar İlkeler (35. Baskı) Teknikler [Scientific Research Method, Concepts Principles Techniques (35 ed.)]*. Nobel Yayıncılık.

Lee, Y. (2004). *The impact of a multidimensional item on differential item functioning (DIF).* [Doctoral dissertation, University of Washington]. https://www.proquest.com/openview/2e24c73698bf27f10d35bd8b63e2cc31/1?pq-origsite=gscholar&cbl=18750&diss=y

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311. https://doi.org/10.1111/j.1745-3984.1993.tb00428.x

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328. https://doi.org/10.1177%2F014662169401800403

Oshima, T.C., & Miller, M.D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*(3), 237-248. https://doi.org/10.1177%2F014662169201600304

Reckase, M.D., & McKinley, R.L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373. https://doi.org/10.1177%2F014662169101500407

Ross, T.R. (2008). *The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test* [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/eps_diss/14/

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371. https://doi.org/10.1177%2F014662169602000404

Russell, S.S. (2005). *Estimates of type I error and power for indices of differential bundle and test functioning* [Doctoral dissertation, Graduate College of Bowling Green State University]. https://www.proquest.com/openview/25873a6f54d69f576b5c2d3ac61f3aa3/1?pq-origsite=gscholar&cbl=18750&diss=y

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*, 159-194.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375. https://doi.org/10.2307/1164616

Wiley, D.E. (1991). Test validity and invalidity reconsidered. In R. Snow & D.E. Wiley (Eds.), *Improving inquiry in social science: a volume in honor of Lee J. Cronbach.* Routledge.

Yao, L., & Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105. https://doi.org/10.1177%2F0146621606291559