مجلة الجامعة الإسلامية

للعلوم التربوية والاجتماعية

مجلة علمية دورية محكمة

تصدر أربع مرات في العام خلال الأشهر:

(مارس، يونيو، سبتمبر، ديسمبر)

العدد 19 - المجلد 36

ربيع الأول 1446 هـ - سبتمبر 2024 م

معلومات الإيداع في مكتبة الملك فهد الوطنية

النسخةالورقية :

رقم الإيداع: 1441/7131

تاريخ الإيداع: 1441/06/18

رقم ردمد : 1658-8509

النسخة الإلكترونية :

رقم الإيداع: 1441/7129

تاريخ الإيداع: 1441/06/18

رقم ردمد : 1658-8495

الموقع الإلكتروني للمجلة :

https://journals.iu.edu.sa/ESS

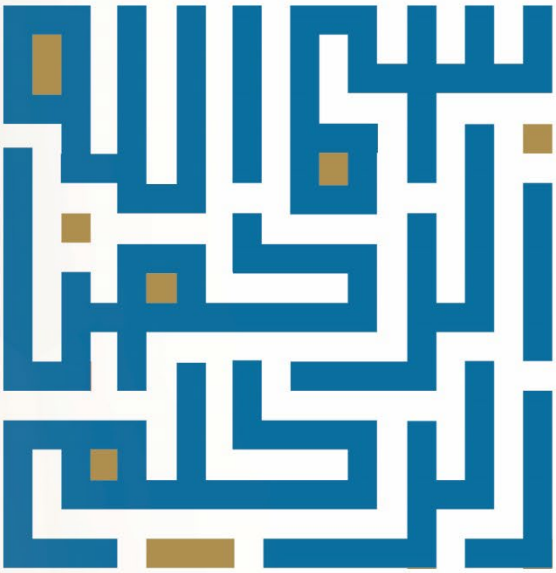
البريد الإلكتروني للمجلة :

ترسل البحوث باسم رئيس تحرير المجلة

iujournal4@iu.edu.sa

الجامعة الإسلامية بالمدينة المنورة
ISLAMIC UNIVERSITY OF MADINAH

البحوث المنشورة في المجلة تعبرعن آراء الباحثـين ولا تعبر بالضرورة عن رأي المجلة



قواعد وضوابط النشر في المجلة

- أن يتسم البحث بالأصالة والجدية والابتكار والإضافة المعرفية في التخصص.
- لم يسبق للباحث نشر بحثه.
- أن لا يكون مستلاً من أطروحة الدكتوراه أو الماجستير سواء بنظام الرسالة أو المشروع البحثي أو المقررات.
- أن يلتزم الباحث بالأمانة العلمية.
- أن تراعى فيه منهجية البحث العلمي وقواعده.
- أن لا تتجاوز نسبة الاقتباس في البحوث التربوية (25%)، وفي غيرها من التخصصات الاجتماعية لا تتجاوز (40%).
- أن لا يتجاوز مجموع كلمات البحث (12000) كلمة بما في ذلك الملخصين العربي والإنجليزي وقائمة المراجع.
- لا يحق للباحث إعادة نشر بحثه المقبول للنشر في المجلة إلا بعد إذن كتابي من رئيس هيئة تحرير المجلة.
- أسلوب التوثيق المعتمد في المجلة هو نظام جمعية علم النفس الأمريكية (APA) الإصدار السابع، وفي الدراسات التاريخية نظام شيكاغو.
- أن يشتمل البحث على : صفحة عنوان البحث، ومستخلص باللغتين العربية والإنجليزية، ومقدمة، وصلب البحث، وخاتمة تتضمن النتائج والتوصيات، وثبت المصادر والمراجع، والملاحق اللازمة مثل: أدوات البحث، والموافقات للتطبيق على العينات وغيرها؛ إن وجدت.
- أن يلتزم الباحث بترجمة المصادر العربية إلى اللغة الإنجليزية.
- يرسل الباحث بحثه إلى المجلة إلكترونيًا ، بصيغة (WORD) وبصيغة (PDF) ويرفق تعهدًا خطيًا بأن البحث لم يسبق نشره ، وأنه غير مقدم للنشر، ولن يقدم للنشر في جهة أخرى حتى تنتهي إجراءات تحكيمه في المجلة.
- المجلة لا تفرض رسوما للنشر.

الهيئة الاستشارية :

هيئة التحرير:

الجامعة الإسلامية بالمدينة المنورة

ISLAMIC UNIVERSITY OF MADINAH

فهرس المحتويات : ❋

❋ترتيب الأبحاث حسب تاريخ ورودها للمجلة مع مراعاة تنوع التخصصات

الجامعة الإسلامية بالمدينة المنورة

ISLAMIC UNIVERSITY OF MADINAH

Comparing the Effectiveness of Two Methods for Detecting Measurement Invariance at the Test Level (Dift and Sibtest) in Light of Differences in Ability Distribution and Sample Size

مقارنة فعالية طريقتين لبيان ثبات القياس على مستوى الاختبار (اختبار الانحراف والاختبار البعدي) في ضوء الاختلافات في توزيع القدرة وحجم العينة

إعداد

Dr. Abdulrahman A. Alnofei

Associate Professor of Psychological Measurement and Evaluation

Department of Psychology - Faculty of Education
Om ALqura University

د. عبد الرحمن عبد الله النفيعي

أستاذ علم النفس المشارك

قسم علم النفس – كلية التربية – جامعة أم القرى

Email: alnofei@gmail.com

DOI:10.36046/2162-000-019-019

المستخلص

هدفت الدراسة إلى مقارنة فاعلية أداء طريقتي DIFT وSIBTEST في الكشف عن تكافؤ القياس للاختبارات وفقاً لمتغيري حجم العينة وفرق توزيع القدرة، وباستخدام تصميم بحثي عاملي يتم من خلاله دراسة أثر التفاعل بين طريقة الكشف وحجم العينة وفوارق توزيع القدرة، وذلك من خلال فحص معدلات الخطأ من النوع الأول، وقوة الاختبار. ولتحقيق هذا الهدف تم إجراء دراستين: الأولى لدراسة معدلات الخطأ من النوع الأول، والثانية لدراسة معدلات القوة للاختبار الإحصائي، وذلك عند ضبط فوارق توزيع القدرة وحجم العينة، باستخدام التصميم التجريبي للقياسات المتكررة لثلاثة عوامل تجريبية، وقد تم تحليل البيانات باستخدام الأسلوب الإحصائي لكل طريقة من طريقتي الكشف، وذلك لاختبار الفرضية الصفرية التي تنص على عدم وجود أداء تفاضلي للفقرة، والحصول على معدلات الخطأ من النوع الأول ومعدلات القوة. وتم معالجة البيانات باستخدام الأسلوب الإحصائي تحليل التباين المختلط. وبناء على نتائج التحليلات الإحصائية تم التوصل إلى الاستنتاجات التالية:

- تميز طريقتي SIBTEST وDIFT بالفاعلية في الكشف عن الأداء التفاضلي للاختبار بصفة عامة.
- طريقة الأداء التفاضلي لحزم الفقرات والاختبار أكثر فاعلية من طريقة اختبار تحيز الفقرة المتزامن عند أخذ حجم العينة في الاعتبار عند استخدام حجم العينة الكبير (١٠٠٠/١٠٠٠) أو أكثر.
- طريقة اختبار تحيز الفقرة المتزامن أكثر فاعلية في الكشف عن الأداء التفاضلي لحزم الفقرات والاختبار في حالة عدم وجود فرق في توزيع القدرة، وفي حالة وجود فرق في توزيع القدرة فكلا الطريقتين غير فاعلة، حيث تعاني طريقة DIFT من ضعف قوة الاختبار الإحصائي، كما تعاني طريقة SIBTEST من تضخم الخطأ من النوع الأول، لذا يوصى باستخدام الطريقتين معا للكشف عن الأداء التفاضلي للاختبارات في حالة وجود فرق في توزيع القدرة بين المجموعتين.

الكلمات المفتاحية: تكافؤ القياس، الأداء التفاضلي للاختبارات، نظرية الاستجابة للمفردة الاختبارية، طريقة DIFT، طريقة SIBTEST.

Abstract

The aim of the current study was to compare the effectiveness of the DIFT and SIBTEST methods in detecting measurement invariance for tests according to sample size and differences in ability distribution. A factorial experimental design was used to look at how the detection method, sample size, and differences in ability distribution all affect each other. Examining type I error rates and test power served to accomplish this. Two studies were conducted, the first to examine Type I error rates and the second to examine test power while controlling for ability distribution differences and sample size. Data were analyzed using statistical methods for each detection method to test the null hypothesis of no differential performance and obtain Type I error rates and test power. The data were processed using mixed-variance analysis. Based on the results of the statistical analysis, a number of important findings were obtained, including: both the SIBTEST and DIFT methods were effective in detecting differential performance of the test in general; the differential item functioning (DIF) method was more effective than the simultaneous item bias test (SIBTEST) when considering sample sizes of 1000 or more. And the differential item bias test was more effective in detecting differential performance of items and tests in the absence of ability distribution differences. However, in the presence of ability distribution differences, both methods were ineffective, as DIFT suffered from low statistical power and SIBTEST suffered from inflated Type I error rates. Therefore, the study recommends using both methods together to detect differential test performance in the presence of ability distribution differences between groups.

Keywords: Measurement parity, differential performance of tests, test single response theory, DIFT method, SIBTEST method.

1. Introduction:

The field of psychological and educational measurement and evaluation is an important area needed by researchers in the behavioral sciences, as well as decision-makers related to individuals in various applied psychological and educational, social, administrative, industrial, military, and other fields centered around individuals. This is to help individuals recognize their abilities, potentials, and energies and make the most of them, and to develop those abilities and potentials in a way that allows for the best possible plans that contribute to achieving the goals that they seek and overcoming their various problems. However, the accuracy of these decisions related to individuals largely depends on the accuracy of the information obtained, which needs to use tools for gathering information that are accurate and reliable, like psychological and educational measures. Experts in measurement and evaluation work hard to create standards and criteria for validity, reliability, and item effectiveness coefficients to make sure that psychological and educational measures are accurate measures of what they are meant to measure.

But if you read information rules and peer-reviewed scientific journals in the field of measurement and evaluation, you'll notice that there has been a lot of interest in the differential item functioning (DIF) property of psychological and educational measures since the mid-1960s. DIF occurs when there are different probabilities of answering an item correctly for test takers from different racial or cultural groups after equating them in the basic ability measured by the test. This interest is evident in the hundreds of studies conducted and in the various statistical methods developed to detect DIF. Measurement experts, legal bodies, and test critics have considered item and test DIF a problem in educational, legal, and professional contexts.

Therefore, excluding DIF from tests and items has been considered a condition of good testing, which is taken into account by scientific associations concerned with psychological and educational measures, such as the American Psychological Association (APA) and the American Educational Research Association (AERA), who have

made it a condition for publishing psychological and educational measures.

Even though the differential item functioning (DIF) issue came up when Binet and Simon made the first psychological tests in 1951, the clear concept of DIF didn't come out until the beginning of the second half of the 20th century. This was thanks to Eells, whose work was the main reason why DIF detection methods based on different measurement theories kept getting better and better.

Methods based on traditional measurement theories, such as the transformed item difficulty (TID) method proposed by Angoff (1972) and the standardization method proposed by Doran's and Kulick (1983), as well as methods based on item response theory (IRT) models, such as the area measures method developed by Linn, Levine, Hastings, and Wardrop (1981), the Lord's chi-square method (1980), the LR-IRT method developed by Thissen et al. (1988), Raju et al. (1995), Raju (1995), and Raju's (1995) DIF and item-level test (DFIT), have emerged. Additionally, methods based on the unidimensional item response theory, such as Ramsey's (1991) TestGraf method, and methods based on probability tables, such as Holland and Thayer's (1988) Mantel-Haenszel (MH) method and Swaminathan and Rogers' (1990) logistic regression method, have also emerged. Furthermore, methods based on multidimensional item response theory, such as Shealy and Stout's (1993) simultaneous item bias test (SIBTEST), have also emerged.

All of the above methods are ways to find differential item functioning (DIF) at the item level, which means that the finding is done separately for each item. Analyzing DIF at this level is necessary and very helpful when making measurement tools because it helps to find items with DIF and then study their content to figure out if the DIF is due to the item's impact, in which case it should be kept, or if it is due to confusion, in which case it needs to be changed or deleted. Specialized organizations in psychological and educational measurement suggest doing three types of statistical analyses on measures before publishing them to get indices of reliability, validity, and DIF (Bufam, 2005). However, there are several observations on DIF detection at the item level, including:

First, it assumes that all test items are free of DIF except for the item being tested, and Raju et al. argue that this assumption is not true in most test situations (1995: 365).

Second, it fails to identify the possible sources and reasons for DIF. Roussos and Stout describe most attempts to explain DIF at the item level as "dismal failures" (1996: 360). Stark et al. also point out the difficulty of predicting which items will have DIF (2001: 949).

Third, in practice, people who use psychological and educational tests make decisions based on the sum of subscale scores or the total test score. DIF detection, on the other hand, is done at the item level, so the cumulative effect of DIF is not taken into account.

Fourth, ignoring the cumulative effect of DIF can lead test developers to leave out items that have DIF at the item level but don't have a big effect at the subscale or test level, which wastes time and effort (Nandakumar, 1993).

Even though it is important, previous research has led a number of researchers to say that differential item functioning (DIF) needs to be studied at levels other than the level of each item (Shealy & Stout, 1993). In response to this idea, two separate studies were done that had a big effect on the development of DIF detection at the item bundle and test levels. The first was conducted by Raju and others, who introduced new concepts distinguishing between two types of item-level DIF: non-compensatory DIF (NCDIF), which assumes that all test items are non-differential except for the item identified as DIF and does not take into account the DIF of other items, due to its lack of additivity, and compensatory DIF (CDIF), which assumes the possibility of other items exhibiting non-uniform DIF, thus taking into account the DIF of other items as it possesses the additivity property, allowing for the aggregation of DIF values across items (Raju et al., 1995). The second study was conducted by Roussos and Stout, who introduced the concept of bundle-level DIF (DBF), where a bundle is a group of related items that naturally occur within test specifications based on an educational framework, such as Bloom's taxonomy, where items measuring each level of the taxonomy (such as knowledge, comprehension, or application) form a bundle of related items (Roussos & Stout, 1996). The concept of bundle-level DIF

extends from the item-level DIF explained by Raju and others and includes two types:

The Compensatory DIF (CDIF) for bundles is where the differential item functioning values for items in each bundle are summed to obtain a value for the compensatory differential item functioning for the bundle, due to the additive property.

$$\text{Bundle CDIF} = \sum_{i=1}^{n} CDIF$$

The compensatory differential item functioning (CDIF) for the bundle is found by adding up the values of each item's differential item functioning. n represents the number of items in the bundle. The CDIF values for the bundle help in studying the inflation or cancellation of differential item functioning. Inflation occurs when differential item functioning is present for most or all items in the test against one group, often the targeted group. Cancellation occurs when the differential item functioning for some items is against the reference group, while for others it is against the targeted group, resulting in the cancellation of the differential item functioning values according to the additivity property (Raju et al., 2006).

The non-compensatory differential item functioning (NCDIF) for the bundle is found by finding the differential item functioning values for each item in the bundle at the same time, without adding the non-compensatory differential item functioning values for each item due to the lack of additivity property. It is calculated as follows:

$$bundleNCDIF = \left[\frac{1}{N_F} \sum_{S=1}^{n_F} \left(\sum_{i=1}^{n} P_{iF}(\theta_s) - \sum_{i=1}^{n} P_{iR}(\theta_s) \right)^2 \right] / n$$

Where:

N_F : the number of examinees in the target group.

N: the number of items in the bundle.

$P_{if}(\theta_s)$: the probability of person (s) in the target group (F) with ability (s) to answer item (i).

$P_{iR}(\theta_s)$: the probability of person (s) in the reference group (R) with ability (s) to answer item (i).

This type of bundle DIF helps in identifying the potential sources and reasons for it (Raju et al., 2006). This is what failed in detecting item DIF, perhaps because an item represents a small and unreliable sample of the behavior being measured and therefore is unable to identify sources and reasons for DIF. On the other hand, using a bundle of items provides a larger representation of behavior, allowing for the identification of potential sources and reasons for DIF (Gierl et al., 2001).

The results of the first two studies led to the creation of ways to find differential item functioning (DIF) at the item bundle and test levels. Based on the item response theory, Raju et al. (1995) came up with the first method in a more advanced framework for finding DIF in items and tests. They called it the Differential Functioning of Items and Tests (DFIT) method, which is used to detect DIF in dichotomous and polytomous data. In this method, the true score of an individual on a K-item dichotomous test is first estimated as follows:

$$T_s = \sum_{i=1}^{k} P_i(\theta_s)$$

Where:

$P_i(\theta_s)$: the probability of a correct response for person with ability on item .

The true score of a person on a test using this method and the theoretical framework of item response theory is the sum of the probabilities of a correct response for each item. According to this method, two separate true scores are estimated for each person, one when they belong to the reference group and another when they belong to the target group. The difference between the two true scores for each person is then calculated as follows:

$$D_s = (T_{sF} - T_{sR})$$

Finally, the differential item functioning (DIF) index for the test is obtained by computing the expected value (E) of the difference squared between the true scores of the reference and target groups, as follows:

$$DTF = E_F\,(T_{sF} - T_{sR})^2 = \sigma_D^2 + \mu_D^2$$

Similarly, the previous equation can also be rewritten as follows:

$$DTF = E_F\,(D^2) = E_F\left[\sum_{i=1}^{k}(diD)\right] = \sum_{i=1}^{k} E_F\,(diD) = \sum_{i=1}^{k}\left[Cov(di, D) + \mu_{di}\,\mu_D\right]$$

$$d_i = P_{iF}(\theta) - P_{iR}(\theta)$$

$$\sum_{i=1}^{k} d_i\, D = T_F - T_R$$

"The difference between the probability difference for paragraph (di) and the difference between the two true scores (D) (Raju et al., 1997) was calculated. Based on the cut-off points proposed by Raju et al. (2016), the different test scores were interpreted. When the value of the differential performance index is higher than the cut-off point, a test is said to have differential performance. The value of the test is the chi-square statistic, which is calculated as follows:"

$$X_{NF}^2 = \frac{N_F\,(\hat{DTF})}{\hat{\sigma}_D^2}$$

Where: the number of examinees in the target group. The estimated value of the test's differential performance The variance of the difference between the estimated true scores and the degree of freedom for the chi-square in the above equation (NF-1).

Raju et al. (1995) came up with two indices for the test's differential performance: the compensatory differential performance index and the non-compensatory differential performance index. These were based on the test's differential performance index and took

into account how different items worked. The first index goes as far as compensatory differential item functioning, which means that any item on the test can have different results and have an additive property that lets the different results cancel each other out. Its value can be positive or negative depending on the direction of the differential performance toward the reference group. It is calculated at the test level as follows:

$$CDIF = E_F(d_i D) = COV(di, D) + \mu_{di}\mu_D$$

The differential performance of this type of test makes it possible to figure out which parts of the test are most responsible for its high differential performance.So, the items with the most different scores are taken out until the test's differential performance index is no longer statistically important (Raju, 1999).

The other type is an extension of non-compensatory differential item functioning, where all other items except for the identified differential item are assumed to be non-differentially functioning, and thus, it does not have the additive property. (Raju et al., 2006) says that the value of each item is found by adding up the non-compensatory differential performance values for that item.

$$NCDIF = E_F\left[P_{iF}(\theta) - R_{iR}(\theta)\right]^2 = E_F(d\hat{i}^2) = \sigma_{di}^2 + \mu_{di}^2$$

A chi-square statistic with degrees of freedom (NF-1) is used to test the non-compensatory differential performance. Raju et al. (2006) found through simulation studies that the chi-square test for the non-compensatory differential performance index is too sensitive for large sample sizes. Therefore, they suggested an alternative significance level of 0.006.

The Differential Item Functioning (DIF) and Test Performance method has a number of benefits (Teresi & Fleishman, 2007):

1. The method is a parameter based on a strong theoretical framework with strong assumptions.

2. The estimated potential is a way to compare the reference group to the target group.

3. It has a statistical significance test.

4. It shows that there are two kinds of different performance: regular and irregular, as well as compensatory and non-compensatory.

5. They reveal the differential performance of the paragraphs and the test as a whole.

6. It deals with data with a dual and multiple response.

7. It is used with response theory models for the one-dimensional and multidimensional test vocabulary.

Roussos and Stout (1996) came up with the second way to find differences in performance at the item and bundle levels. This method is an extension of the way Shealy and Stout (1993) found differences in performance at the item level. The Simultaneous Item Bias Test (SIBTEST) is a non-parametric method based on multidimensional item response theory models that is used to figure out how biased an item is. It is assumed that all psychological and educational scale items measure the desired characteristic or ability (θ), but some items may measure another undesired ability called the nuisance parameter (η), meaning that the ability is multidimensional. This is reflected in the definition of differential item performance and bundle performance, which is "the difference in the probability of a correct response to the item or bundle between the reference and target groups, who are equal in the desired ability to be measured (η) and different in the nuisance ability to be measured (η)" (Roussos & Stout, 1996). This can be expressed mathematically as:

$$T_{iF}(\theta) \neq T_{iR}(\theta)$$

"Where: θ: refers to the ability being measured.

Where:

g: indicates the reference and target group.

$P_i(\theta,\eta)$: The probability of a correct answer for an individual with the ability (θ,η) to paragraph(i).

$f_g(\theta,\eta)$: refers to the conditional density function of the power distribution

when it is (η) when (θ)known.

The differential item functioning of a set of items can be detected using this method by dividing the items into two subgroups. The first subgroup is called the honest subgroup and consists of non-differential items that measure the intended construct of the item set(θ**).**

According to the text:

(1), the paragraphs in this group are identified in two ways: either by using the repeated refinement method for the same method or by using another method for detecting differential performance, such as the Mantel-Haenszel method. The second group of paragraphs is called the suspected subset and consists of the remaining paragraphs of the package.

that measure the targeted basic ability,(θ**)** "Additionally, the unwanted ability(η) Which can be measured"

(Roussos & Stout, 1996), the statistical indicator for finding concurrent item bias is as follows:

1. the package paragraphs for the truthful subset, represented by paragraphs from 1 to (K) out of the total package paragraphs (N), are placed. (Ui) represents the degree of paragraphs, which is zero or one. Therefore, the total score for an individual in the truthful subset becomes as follows:

$$X = \sum_{i=0}^{k} U_i$$

2.Consider the rest of the paragraphs from paragraph (K+1) to paragraph (N) as the paragraphs that represent the suspected subgroup and the individual's overall score in this group are:

$$X = \sum_{i=0}^{k} U_i$$

3. From paragraph (K+1) to paragraph (N), the remaining paragraphs make up the suspected subset. The total score for a person in this subset is:

$$Y = \sum_{i=K+1}^{N} U_i$$

4.To compare groups, the observed total score for the truthful subset X = 1,.... ,K is used to form equal-ability groups from the reference and targeted groups, creating levels of the criterion for comparison (3). The true score for individuals in the reference and targeted groups is estimated for each level of the criterion for comparison using the classical test theory by regression.

5. Finding the statistical indicator for differential item performance of the method ($\hat{\beta}_u$) "Through the following equation"

$$\hat{\beta}_u = \sum_{K=o}^{k} P_k \hat{} (\overline{Y_{R_k}}^{*} - \overline{Y_{F_k}}^{*}) = \sum_{K=o}^{k} (\overline{Y_{R_k}}^{*} - \overline{Y_{F_k}}^{*}) \frac{J_{RK} + J_{FK}}{\sum_{j=o}^{\eta} (J_{Rj} + j_{Fj})}$$

"Where:"

P_k **: "The proportion of examinees in the targeted group with the total score in the truthful subset X=k."**

$(\overline{Y_{R_k}}^{*} - \overline{Y_{F_k}}^{*})$ **:**

the difference between the mean adjusted scores (true score) of the suspected subset of items for both the reference and targeted groups (Nandakumar, 1993).

6. Testing the null hypothesis that the package of items does not exhibit differential performance, which is expressed as follows:"

$$H_o : \beta_u = o \quad vs. \quad H_a : \beta_u > o$$

"And this is done by finding the statistical test for the method (B), which is approximately normally distributed, using the following equation"

$$B = \frac{\hat{\beta}_u}{\hat{\sigma}_{\hat{\beta}_\varepsilon}}$$

represents the estimated standard error of the statistic, which is calculated using the following equation:

$$\hat{\sigma}_{\hat{\beta}_u} = \sqrt{\sum_{K=o}^{K} \hat{P}_K^2 \left(\frac{1}{J_{RK}} \hat{\sigma}^2 (Y|K,R) + \frac{1}{J_{FK}} \hat{\sigma}^2 (Y|K,F)\right)}$$

"Where:"

$\hat{\sigma}^2(Y|K,R)$: Represents the sample variance for the subset of examinees in the targeted group with the total score of K in the truthful subset

J_{RK}, J_{FK}: "Represents the sample size for the reference and targeted groups, respectively."

7. In the case of statistical significance and hence rejection of the null hypothesis, the test statistic $(\hat{\beta}_u)$ is used as a measure of effect size that reflects the magnitude and degree of differential performance, using the classification criteria proposed by Roussos and Stout (1996) for classifying paragraph and bundle differential performance as follows:

1) Negligible differential performance that can be overlooked or at the level of (A): rejection of the null hypothesis and $|\hat{\beta}_u| < 0.059$

2) Average differential performance or at the level of (B): rejection of the null hypothesis and $0.059 \leq |\hat{\beta}_u| < 0.088$

3) Significant differential performance or at the level of (C): rejection of the null hypothesis and $\left|\hat{\beta}_u\right| \geq 0.088$

The simultaneous item bias test method (Teresi & Fleishman, 2007) has the following features:

- A non-parametric approach based on a strong theoretical framework, the multidimensional item response theory.
- The model does not require specific assumptions about the data.
- Suitable for use with medium and short tests.
- can be used with multidimensional item response theory models.
- It has a statistical significance test and a measure of the size of the effect, along with a criterion for classifying the different levels of performance.
- It detects the differential performance of both items and item bundles.
- It reveals the causes of item and item bundle bias.
- The method is not complex and does not require extensive effort.
- The analysis using the DIF-T and SIBTEST methods at the level of item sets and tests contributed to overcoming the criticisms that were made to the analysis at the level of test items, as it has the following advantages:
  - The ability to study the inflation or cancellation of differential performance through compensatory DIF, which takes into account the possibility that many items in the test or item set have differential performance, is consistent with the nature of different test situations.
  - The ability to detect possible sources and causes of differential performance through non-compensatory DIF, where the item set

allows for a greater representation of behavior, thus providing greater opportunities to identify sources and causes

Several studies have been done to compare how well the DIF-T method and the SIBTEST method find differences in how items and tests perform. As both methods provide a statistical test for the null hypothesis that there is no differential performance of the item set or test, the type I error rate and statistical power are the two basic criteria for comparing their performance in studies that have compared them. A good method is one that keeps the type I error rate of its statistical test at or below the nominal alpha level and has good power rates. This is critical to ensuring the validity of the test hypothesis. If the null hypothesis of no differential performance is accepted, then it can be concluded that the item set or test does not contain differential performance, while if it is rejected, it can be concluded that the item set or test has differential performance. Maintaining the statistical test of the method with a type I error rate below the nominal alpha level is crucial from a statistical perspective, as the statistical power of the test is unknown unless the test maintains its type I error rate below the nominal alpha level (Shealy & Stout, 1993). In all comparative studies, a simulation method was used to examine the type I error rates and the experimental power rates of statistical tests for both methods. The simulation method generated study data for items with specific characteristics that made the items in the item set or test non-differential when examining type, I error rates, and differential when examining power rates. The data was repeated in light of some variables several times (usually a hundred times) for each variable, and the method was applied to those data to estimate the type I error rates and power rates of the statistical test.

Previous studies have compared the performance of two methods in terms of type I error rates and statistical test power based on several key variables that the data were generated under, and a comparison was made. Sample size for both reference and target groups is considered the most important variable that previous studies have focused on and controlled for. Sample sizes used in these studies range from n = 100 in the study by Roussos and Stout (1996) to n = 300 in the studies by Nandakumar (1993), Shealy (1993), and Stout (1993) for testing simultaneous item bias; n = 250 in the study by

Maurer et al. (1998); and approximately n = 10,000 in the study by Facteau and Craig (2001) for Differential Item and Test Functioning (DIFT) performance. Although most sample sizes ranged from n = 500 to n = 1000 for previous studies of testing for simultaneous item bias, positive relationships were found between sample size and type I error rates, particularly when there were differences in ability distribution between reference and target groups (Russell, 2005; Bolt, 2002; Roussos & Stout, 1996).

In the past, type I error rates and statistical power have been used to compare the two methods (Roussos & Stout, 1996). Sample size for both the reference and target groups was thought to be the most important variable and was controlled for. Sample sizes used in these studies ranged from n = 100 individuals in Roussos and Stout's study (Roussos & Stout, 1996) to n = 300 individuals in Nandakumar's and Shealy and Stout's studies (Nandakumar, 1993; Shealy & Stout, 1993) for testing simultaneous item bias, to n = 250 individuals in Maurer et al.'s study (Maurer et al., 1998), to approximately n = 10,000 individuals in Facteau and Craig's study (Facteau & Craig, 2001) for the differential item functioning However, most sample sizes ranged from n = 500 to n = 1,000 individuals. Previous studies have found a positive relationship between sample size and type I error rates, indicating that increasing sample size led to an increase in type I error rates, especially when there were differences in ability distribution between the reference and target groups (Russell, 2005; Bolt, 2002; Roussos & Stout, 1996). Meanwhile, Raju et al. and Russell found a negative relationship between sample size and differential item functioning performance for the differential item functioning test (Raju et al., 1995; Russell, 2005). Type II error rates, on the other hand, were found to increase with sample size for both methods. It is crucial to maintain a statistical test's type I error rates below the nominal alpha level to ensure that the test is reliable for its hypothesis. If the null hypothesis of non-differential performance is accepted, it can be inferred that the package or test does not contain differential performance. If it is rejected, it can be inferred that the package or test has differential performance. To examine the type I error rates and power for both methods, simulation was used in all comparative studies. Data were generated for paragraphs with specific characteristics that made them non-differential when examining type I

error rates and differential when examining power rates. The data were repeated several times (usually 100) for each variable, and the methods were applied to those data to estimate type I error rates and power rates for the statistical test. The second variable controlled for in studies comparing the two methods was the difference in ability distribution between the reference and target groups.

groups, which was expressed as a metric scale ($d_\theta = \mu_{\theta R} - \mu_{\theta F}$) ,(). Differences in ability distribution in previous studies ranged from no differences (($d_\theta = 0$)), to moderate differences ($d_\theta = -0.5$) to large differences ($d_\theta = -1$).

In recent studies, Raju et al. (1995) and Russell (2005) found that sample size and differential item functioning (DIF) performance for paragraph DIF and test DIF were related in a way that was not positive.Specifically, the studies found that the rates of type I error for the statistical indicators of the two DIF methods increased with smaller sample sizes and decreased with larger ones. Moreover, the studies found that the power rates increased with larger sample sizes for both methods.

In the studies that compared the two methods, the second variable that was controlled for was the difference in ability distribution between the reference and targeted groups, which was shown on a metric scale.The magnitude of the difference varied across the studies, ranging from no difference (Bolt, 2002), to small (Facteau & Craig, 2001; Robie et al., 2001), to moderate (Roussos & Stout, 1993), to large (Stark et al., 2001).

The third variable controlled for in the simulation studies was the size of DIF for the items in the test package, which was activated as a part of some or all items. The percentage of DIF items varied across the studies, ranging from a few items in some studies (Facteau & Craig, 2001; Robie et al., 2001) to half of the items in one study (Stark et al., 2001), and 20% of the items in most studies. In a study by Raju et al. (1995), a test of 40 items was generated with DIF rates of 5% (2 items), 10% (4 items), and 20% (8 items). The study found that the rates of type I errors for paragraph DIF and test DIF stayed the same

as DIF got bigger, but the rates of power errors went down as DIF got bigger.

In a study by Shealy and Stout (1993), subtests were generated with DIF items ranging in size from 0% to 12.5% of the total test. The study found that the rates of type I error for the simultaneous item bias test remained constant with increasing DIF size, while the power rates increased with increasing DIF sizee simultaneous item bias test remained constant with increasing DIF size, while the power rates increased with increasing DIF size. Similar results were reported in a study by Nandakumar (1993) with the same DIF sizes.

Previous studies comparing the performance of the paragraph differential item functioning (DIF) and simultaneous item bias test (SIBTEST) have used simple research designs to compare each variable separately. This study aims to compare the performance of paragraph DIF, SIBTEST, and test bias detection in detecting item-level DIF and test-level DIF based on the variables of sample size and ability distribution differences, using an advanced factorial research design to examine the interaction effects between test type, sample size, and ability distribution differences.

More specifically, the research question is: What effect do differences in test type, sample size, and ability distribution have on Type I error rates and statistical power for detecting test bias? This question leads to the following sub-questions:

1. **Type I error rates:**

- Do Type I error rates differ by test type?
- Do Type I error rates differ by the two-way interaction between test type and sample size?
- Do Type I error rates change when there is a two-way interaction between the type of test and the way people's skills are distributed?
- Do Type I error rates change based on how test type, sample size, and differences in how people's abilities are spread out interact with each other?

2. Statistical power:

- Does statistical power differ by test type?
- Does statistical power differ by the two-way interaction between test type and sample size?
- Does the two-way interaction between test type and differences in how well people do on tests change the statistical power?
- Does statistical power change based on how test type, sample size, and differences in ability distribution interact with each other?

Based on the raised questions about the study issue, the study hypotheses were formulated as follows:

First: Type I error rates:

- The experimental Type I error rate is the same no matter what method is used to find out how paragraph bundles and tests perform differently.
- The Type I error rate doesn't change based on how the detection method and sample size interact with each other.
- The Type I error rate does not change based on how the detection method and the difference in ability distribution interact.
- The experimental Type I error rate doesn't change based on how the detection method, sample size, and difference in ability distribution interact with each other.

Second: Statistical test power rates:

- The experimental power rates of the statistical test don't change based on the method used to find out how paragraph bundles and tests perform differently.
- The experimental power rates of the statistical test don't change based on how the detection method and sample size interact with each other in two ways.

- The experimental power rates of the statistical test don't change based on how the detection method interacts with the difference in ability distribution.
- The experimental power rates of the statistical test don't change based on how the detection method, sample size, and difference in ability distribution interact with each other.

2. THE METHODOLOGY AND PROCEDURES:

Study design: The study aimed to compare the performance of two testing methods: the simultaneous paragraph bias test and the differential performance of paragraph bundles and tests in detecting the differential performance of paragraph bundles and tests. The study examined the Type I error rates and statistical test power rates of the two methods when adjusting for the sample size and ability distribution difference between the reference and targeted groups. Therefore, the researcher used an experimental design to answer the study's questions. This design provides an understanding of the directed causal relationship between the independent variables controlled and adjusted through the simulation design and the dependent variable represented in the Type I error rates and statistical test power rates.

Research design: The study used a Three-Factor Experiment with Repeated Measures design (Winer, 1971) to measure three experimental factors. The dependent variables in this design are the Type I error rates in the first part of the study and the statistical test power rates in the second part of the study. The independent variables that were controlled and adjusted are the two methods of detecting differential performance of bundles and tests, which represent the repeated measures variable; sample size, which was adjusted and determined by two sample sizes; and ability distribution difference between the reference and targeted groups, which was adjusted and determined by three ability distribution differences. Since the design used was of the two between, one within factor type, a 2x2x3 factorial design was obtained. To control and adjust the independent variables according to the research design, a simulation study method was used to generate study data. The WinGen3 software (Han, 2007) was used

to generate datasets that will be analyzed using the specialized software of the two compared detection methods.

Steps of the study:

The study was conducted according to the following steps:

1. Determining the number of paragraphs in each package in the Type I error study, which consisted of thirty-two (32) non-equivalent paragraphs, and their characteristics, which will be generated in light of them (see Appendix 1),

2. Determining the number of paragraphs in each package in the statistical power study, which consisted of eight (8) paragraphs with differential performance and their characteristics, will be generated in light of them (see Appendix 2).

3. determining the sample sizes and numbers for each of the reference and target groups, where two sample sizes were selected for the reference and target groups, respectively (500/500, 1000/1000), which were selected because they are the most common in previous studies.

4. determining the characteristics of the ability distribution for the reference and target groups and the difference between them, where the ability for the reference group was fixed at the ability level of zero (), while it was varied for the target group to become (), thus producing three levels of ability distribution differences: no difference (), moderate difference (), and large difference ().

5. Calculating the number of packages that will be generated, which consist of the number of cells in the overall design (two sample sizes 3 ability distribution differences), results in six packages to study the type I error, and the same for the statistical power study.

6. generating data for each of the six packages for both the Type I error and statistical power studies using WinGen3 software (Han, 2007), based on the data mentioned in steps (1-4).

7. Repeating each of the six packages one hundred times for the Type I error study and the same for the statistical power study results

in (600) packages for the Type I error study and (600) packages for the statistical power study, each package in a separate file.

8. Using the DIFT and SIBTEST methods to look at the paragraph package files made in step 7, where:

a. The paragraph package files, which numbered (1200) files and were distributed according to sample size and ability distribution differences, were analyzed using SIBTEST 1.1 software (Stout & Roussos, 2005) to obtain the value of the statistical test for the SIBTEST method, which is used to test the null hypothesis, as well as the effect size measure. b. The same paragraph packages as in (a) were analyzed using DFIT8 software (Oshima et al., 2009) to obtain the value of the statistical test for the DIFT method, which is used to test the null hypothesis, as well as the effect size measure. Since the scores for the two groups are not on the same scale, as the method assumes, the scores for the reference and target groups were equated using IRTEQ 1.1 programming (Han, 2011) and placed on a common scale, and then the data was analyzed using DIFIT8 software.

b. Classifying each of the analyzed paragraph packages and obtaining their indicators as differential or non-differential performance, using the criterion shown in Table 1 for the SIBTEST method, where the paragraph is classified as having differential performance if the statistical test for the null hypothesis is statistically significant and the effect size is within level C or B, while it is classified as non-differential if otherwise.

9.Classifying each of the analyzed paragraph packages and obtaining their indicators as differential or non-differential performance, using the criterion shown in Table 1 for the SIBTEST method, where the paragraph is classified as having differential performance if the statistical test for the null hypothesis is statistically significant, and the effect size is within level C or B, while it is classified as non-differential if otherwise.

Table 1, Describes the criteria for classifying paragraph packets with differential performance according to the SIB method.

Way	Statistical testing	Scale of impact	Standard of classification according to the size of differential performance			Reference
			A (Small)	B (Mediam)	C (Strong)	
SIB	B	$\hat{\beta}_u$	$\lvert 0.059 \rvert >$	To me $0.059 \lvert$ $\lvert 0.088$	$\lvert 0.088 \rvert \leq$	Roussos & Stout (1996)

Regarding the DIFT method, the statistical test is represented by (Chi-square test)(X^2) of an statistical indicator, and the differential performance index is greater than the cutoff value generated by the DFIT8 program (Oshima, et al., 2009).

11. The results from steps 10 and 11 were entered into the SPSS 25 program, where two separate files were created, the first for studying Type I error and the second for studying statistical power.

12. The experimental Type I error rate for each method was calculated based on the number of times the package was classified as having differential performance compared to the other 100 packages and was divided by 100. It was then compared to the nominally expected Type I error rate, which was set at 0.05. The average power of the statistical test for each method was calculated after the number of times the package was classified as having differential performance compared to the other 100 packages and was divided by 100. The power rates were then classified according to Cohen's criteria (Gotzmann, 2001).

Weak force: $Power < 0.70$

Medium Strength : $0.80 > power \geq 0.70$

Great power: $power \geq 80$

13. A statistical method called "multivariate analysis of variance for repeated measures" was used to process and look at the data. The statistical method used was a mixed-design, three-factor experiment with repeated measures. Therefore, a multivariate analysis technique called repeated measures: two between subjects and one within subjects variable was used to analyze the data and answer research questions related to Type I error rates and statistical power after verifying the assumptions, which include the normality of the dependent variable for each population included in the analysis and equal variances of the error between any two levels of the factor within the groups and equal covariance matrices of the independent variables within the groups (Alam, 2003).

1) 1)Assumptions for statistical analysis include the normality of the dependent variable for each population included in the analysis.

2) 2) as well as equal variances of the error between any two levels of the factor within the groups

3) 3) equal covariance matrices of the independent variables within the groups. (Stevens, 2002).

Since the statistical significance of differences reflects their apparent significance and therefore is affected by sample size, practical significance was also determined for the statistical test, expressed by the effect size (ES) and represented by Partial Eta Squared, which is interpreted according to the following criterion (Cohen, 2005).

1.Small effect size:

$$0.06 > \eta_p^2.$$

2.Average effect size:

$$0.14 > \eta_p^2 \geq 0.06$$

2. The magnitude of the significant impact:

$$\eta_p^2 \geq 0.14$$

3. RESULTS AND DISCUSSION:

The study was designed to compare the effectiveness of the DIFT and SIBTEST methods in detecting differential item and test functioning, based on sample size and ability distribution differences. The impact of these factors on the Type I error rates and experimental test power for each method was examined and compared to the expected Type I error rates and test power.

Therefore, the results of the Type I error analysis will be presented and discussed first, followed by the results of the experimental test power analysis, and finally a summary of the overall results.

Table 2, Results of variance analysis for repeated measurements of the type 1 error study using the MEDA Wilkes test

Effects	The value of Wilkes LIMD	Value of p.	Degrees of freedom of assumptions	Degrees of freedom of error	Level of significance	Part of ETA
Method	0.986	8.326	1	594	0.004	0.067
Interaction with ability	0.979	6.472	2	594	0.002	0.085
Interaction with the sample size	0.989	6.375	1	594	0.012	0.064
Method interaction, capacity distribution and sample size	0.995	1.398	2	594	0.248	0.005

Presenting the results of the statistical analysis of the Type I error study and discussing them: To verify the impact of the differences in the methods of detecting differential performance of item bundles, sample size, and differences in ability distribution on Type I error rates the data was analyzed using a mixed-variance analysis method.

1. Firstly, the assumptions of the study were examined, and the results of the Kolmogorov-Smirnov, Shapiro-Wilk, and Box's

M tests showed no significant statistical differences at the (0.05) significance between the Type I error rates for the two detection methods, indicating that they followed a normal distribution. Furthermore, the Box's M test showed no significant statistical difference at the (0.05) significance, which confirmed the assumption of homogeneity of variance across the independent variables for the groups.

2.Secondly, a multivariate analysis was conducted using the Wilks' Lambda test, and the results are presented in Table (2).

Third, the statistical analysis results were used to answer the questions about type I error. This was done by applying the Wilks' lambda test to the means shown in the previous table. The first null hypothesis is: The statement that "there is no difference in the experimental type I error rate across different methods of detecting differential item and bundle functioning (DIF/DBF)", was examined for acceptance or rejection. Using the results presented in Table 2 for the method, it is noted that the Wilks' lambda test value was 0.986, which is statistically significant at the 0.01 level. This means that the experimental type I error rate differs across different methods of detection. To confirm the practical significance of the difference, the partial eta squared effect size indicator was computed, and the value was 0.067, indicating a medium effect size. Therefore, the null hypothesis was rejected and the alternative hypothesis, stating that "the experimental type I error rate differs across different methods of detecting DIF and DBF," was accepted.

Considering the average type I error rates for the SIBTEST and DIFT methods, which were 0.060 and 0.037, respectively, it is observed that the DIFT method had a lower type I error rate in general than the SIBTEST method. Moreover, it was less than the nominal alpha level of 0.05. However, the type I error rate for the SIBTEST method was greater than the nominal alpha level, but it was within the expected upper limit according to the proposed criterion by Narayanan and Swaminathan (1994), which is 0.635 for the previous cases. Therefore, both methods maintained the type I error rate within the expected limits, but the DIFT method was more accurate in detecting DIF and DBF than the SIBTEST method. The reason for the difference in effectiveness of detecting DIF and DBF between the two

methods may be due to the type of comparison standard used by each method. The DIFT method uses the latent and estimated ability according to the item response theory models as a comparison standard, while the SIBTEST method uses the true score estimate according to the traditional test theory as a comparison standard. Studies have confirmed that the estimated true score according to the item response theory is more accurate than the estimated true score according to the traditional test theory, which reflects the accuracy of both methods. Also, using the estimated true score as a comparison standard instead of the observed scores, which include measurement error, explains the ability of both methods to maintain the type I error rate within the expected limits of the type I error rates. This answers the first study question.

3.The second hypothesis of the study, which states that "the experimental type I error rate does not vary as a function of the interaction between the detection method and sample size"

was tested using the results from Table 2. The obtained value for the Wilks' Lambda test (0.989) was statistically significant at the alpha level of 0.012, indicating the presence of significant differences. The partial eta squared value of 0.064 indicated a medium effect size, suggesting that the differences were practically significant as well. These results confirm that the experimental type I error rates vary as a function of the interaction between the detection method and sample size, rejecting the null hypothesis and accepting the alternative hypothesis. To further explore the nature of the interaction and its effect on the type I error rates, the error rates for the two detection methods were calculated for each sample size and presented in Table 3.

Table 3, Experimental Type I error rates for the two detection methods according to sample size

The sample size	The Dift method	The sibtest method
R=500, f=500	0.047	0.050
R=1000, f=1000	0.027	0.077

From the previous table, we can see that as the sample size went up, the experimental type I error rate for the DIFT method went down.Also, for both sample sizes, it was less than the standard alpha level of 0.05.This result is the same as what other studies have found, which is that the sample size has the opposite effect on the rate of type I errors .Specifically, the type I error rates were higher for the smaller sample size of 500 individuals than for the larger sample size of 1000 individuals (Russell, 2005; Raju et al., 1995).

Also, as the sample size went up, the experimental type I error rate for the SIBTEST method went up .It was equal to the nominal alpha level of 0.05 for the smaller sample size of 500 individuals and larger than the nominal alpha level for the larger sample size of 1000 individuals. This result is also in line with the results of other studies that found a direct link between the size of the sample and the rate of type I errors. That is, as the sample size increases, the type I error rates also increase (Russell, 2005; Bolt, 2002; Roussos & Stout, 1996).

The reason for the difference in the two methods' responses to sample size may be attributed to the nature of the model and theoretical framework followed by each method. The DIFT method belongs to models of item response theory, which increase in accuracy in estimating latent ability with increasing sample size. This leads to a decrease in type I error rates. On the other hand, the SIBTEST method belongs to models of nonparametric item response theory that give accurate estimates for small sample sizes. This leads to an increase in type I error rates as sample size increases. This answers the second study's questions (Embretson & Reise, 2013).

The third hypothesis of the study on Type I error, which says that "the interaction between the detection method and the difference in ability distribution does not change the Type I error rate," was tested. The results shown in Table 2 were used, which are specific to the interaction between the detection method and the ability distribution difference. The value of the Wilks' lambda test was 0.979, which is statistically significant at the alpha level of 0.01 ($\lambda = 0.979$, $p < 0.01$). Also, the partial eta-squared value was 0.085, indicating a moderate effect size. These results confirm that Type I error rates differ due to the interaction between the detection method and the ability

distribution difference. To examine the nature of the interaction and its impact on type I error rates, the type I error rates of the two methods were calculated according to the ability distribution difference, and the results are presented in Table 4 (Johnson, 2009).

Table 4, Experimental Type I error rates for the two detection methods according to the power distribution difference

Distribution of capacity	THE DIFT METHOD	THE SIBTEST METHOD
$d_\theta = 0$	0.020	0.005
$d_\theta = -0.5$	0.025	0.075
$d_\theta = -1$	0.0625	0.110

It is observed from Table 4 that the experimental Type I error rate for both the DIFT and SIBTEST methods increases with an increase in the ability difference between the reference and target groups. This finding is consistent with previous studies that found a positive relationship between Type I error rate and ability difference (Russell, 2005; Bolt, 2002; Roussos & Stout, 1996; Shealy & Stout, 1993).

Moreover, the DIFT method maintained the Type I error rate within the expected limits of the criterion proposed by Narayanan and Swaminathan (1994). The Type I error rates for each ability difference level were less than the nominal level of alpha. On the other hand, the SIBTEST method maintained the type I error rate below the nominal level of alpha (0.05) in the case of no ability difference between the reference and target groups. However, if there was an ability difference, whether small or large, the type I error rates were inflated and greater than the nominal level of alpha.

The fact that the rate of Type I errors goes up as the difference between people's abilities gets bigger may be because it gets harder to tell when people's abilities are different.This is because both methods assume that all examinees in each ability level are equal in ability, and when comparing unequal ability distributions, this assumption is not met, resulting in an invalid comparison criterion. This leads to an incorrect interpretation of item effects as differential performance, resulting in an increase in Type I error rates.

The DIFT method is able to keep the type I error rate below the nominal level for all levels of ability difference because it uses a common scale to compare the scores of the reference group and the target group. The SIBTEST method doesn't do this. This procedure reduces the impact of ability differences, which are not addressed by the simultaneous item bias test. This provides an answer to the third question of the study's inquiries.

4.To verify the third hypothesis of the study, which states that the experimental type I error rate does not vary with the three-way interaction between the detection method, sample size, and ability distribution difference:

the results presented in Table 2 were used. It is observed that the Wilks' lambda test value was 0.995, which is not statistically significant as the significance level was 0.248, much larger than the nominal alpha level (0.05). This indicates the absence of statistically significant differences, thus accepting the null hypothesis that the type I error rates do not vary with the three-way interaction between the detection method, sample size, and ability distribution difference (Liu, 2018). This answers the fourth question of the study.

Statistical Analysis Results for Study Power Evaluation:

The mixed model ANOVA was used to investigate the effect of variations in the detection method, sample size, and ability distribution difference on the statistical power. Before conducting the analysis, the normality assumption was confirmed using the Kolmogorov-Smirnov and Shapiro-Wilk tests, and the homogeneity of variance assumption was confirmed using Levene's test. The results of the Box's M test also confirmed the assumption of variance homogeneity across the groups. The results of the Wilks' Lambda test are presented in Table (5).

Table 5, Results of variance analysis for repeated measurements to study statistical test strength rates using the Wilkes-LMEDA test

Effects	The value of the Wilkes LIMDA test	The value of the test p.	Degrees of freedom of assumptions	Degrees of freedom of error	Level of significance	Part of ETA
Method	0.683	276.182	1	594	Zero	0.317
Interaction with ability	0.961	12.098	2	594	Zero	0.076
Interaction with the sample size	0.996	2.643	1	594	0.105	0.004
Method interaction, capacity distribution and sample size	0.994	1.787	2	594	0.168	0.006

The results shown in the table were used to answer the questions of studying statistical test force rates as follows:

5. To verify the fifth hypothesis of the study, which states that "the experimental power rates of the statistical test do not differ with variations in the method of detecting differential item functioning"

The results presented in Table 5 were used. It can be observed that the Wilks' Lambda test value was 0.683, a statistically significant value at a significance level of less than 0.01. Additionally, the partial eta squared value was 0.317, indicating a large effect size, which means that the null hypothesis is rejected and the alternative hypothesis is accepted, indicating that the statistical power rate differs with variations in the method of detecting differential item functioning. The power rate means for SIBTEST and DIFT were 0.87 and 0.72, respectively, indicating that the power rate for SIBTEST was higher than that for DIFT and was a high power, while the power rate for DIFT was moderate. This answers the fifth question of the study.

6. To verify the sixth hypothesis of the study which states "that the experimental statistical power rates of the test do not differ with the interaction between the detection method and sample size,"

The results of the analysis presented in Table 5 were used. The Wilks' lambda test value for the interaction between the method and sample size was (0.05) indicating no statistically significant differences and therefore accepting the null hypothesis that the statistical power rates do not differ with the interaction between the detection method and sample size. In other words, there is no effect of sample size on the power rates, contrary to previous studies' findings. Russell (2005) and Raju et al. (1995) found that DIFT's power rates increase with an increase in sample size, and both Russell (2005) and Bolt (2002) and Roussos and Stout (1996) found that SIBTEST's power rates also increase with an increase in sample size. This is the answer to the study's sixth question.

7. **The seventh hypothesis of the study, which states that "the experimental power rates of the statistical test do not differ depending on the two-way interaction between the detection method and ability difference"**.

The results of the analysis in Table (5) were used to examine the interaction between the detection method and ability differences. The value of the Wilks' lambda test was (0.961), which is statistically significant at the (0.05) level and practically significant with a partial eta squared value of (0.076), indicating a medium effect size. Therefore, the hypothesis is rejected, and it is confirmed that the power rates vary depending on the two-way interaction between the detection method and ability difference. To explore the nature of the interaction and its effect on power rates, the power rates for both methods were calculated according to the ability difference, and the results are presented in Table (6).

Table 6, The experimental power rates of the two detection methods according to the power distribution difference:

Distribution of capacity	The Dift method	The sibtest method
$d_\theta = 0$	0.730	0.975
$d_\theta = -0.5$	0.495	0.930
$d_\theta = -1$	0.475	0.715

It is observed from Table 6 that the power of the DIFT method decreases with increasing differences in ability distribution between the reference and targeted groups, which is consistent with the findings of studies conducted by Russell (2005) and Bolt (2002). It is also observed that the power of the SIBTEST method decreases with increasing differences in ability distribution, which differs from the findings of studies conducted by Shealy and Stout (1993) and Russell (2005), which found a slight increase in power with increasing ability distribution differences. The difference in results may be due to differences in study design between the current study and the previous studies.

8.To test the validity of the eighth hypothesis, which states that "the experimental force rates of the statistical test do not differ depending on the tripartite interaction between the detection method, sample size, and power distribution,"

The results presented in Table 5 for the interaction of method, sample size, and power distribution were utilized. It was observed that the value of the Wilks' lambda test was (0.994), which is statistically non-significant, given that the level of significance was (0.168), which is much higher than the nominal alpha level (0.05). This indicates that there is no significant difference in the experimental force rates based on the tripartite interaction between the detection method, sample size, and power distribution, thus supporting the null hypothesis. Therefore, this provides an answer to the eighth question of the study (Smith et al., 2018).

4. CONCLUSION:

The results of my study on type I error and statistical power can be summarized as follows:

1) The method of detecting differential item functioning (DIF) and the effective test were found to maintain a type I error rate less than the nominal level of alpha and to have at least moderate power. Thus, both the SIBTEST and DIFT methods were effective in detecting differential item functioning and in testing in general, with type I error rates for each method lower than the nominal alpha level and with moderate or higher power. However, the SIBTEST method was more effective, with greater statistical test power, while the DIFT method had moderate statistical test power.

2) When sample size was taken into account, the experimental type I error rate for the DIFT method went down as sample size went up, and it was lower for both sample sizes than the nominal alpha rate. There was no effect of increasing sample size on rates of statistical test power. On the other hand, the experimental type I error rate for the SIBTEST method increased as sample size increased, and it was equal to the nominal alpha level (0.05) for the small sample size or greater than the nominal alpha level for the large sample size. There was no effect of increasing sample size on the rates of statistical test power. The previous results confirmed that the DIFT method was more effective than the SIBTEST method when sample size was considered, especially when using a large sample size (1000/1000) or more.

3) When the difference in ability distribution between the reference and target groups was taken into account, the experimental type I error rate for the DIFT method went up as the difference in ability distribution between the reference and target groups went up, and it was always less than the nominal alpha level. However, the rates of statistical power decreased as the difference in ability distribution increased, with power being moderate in the absence of a difference in ability distribution or weak when there was a difference. The experimental type I error rate for the SIBTEST method also increased as the difference in ability distribution between the reference and target groups increased, and it was lower than the

nominal alpha level in the absence of a difference in ability distribution but greater than the nominal alpha level when there was a difference. The rates of statistical power decreased as the difference in ability distribution increased but remained moderate or large. The previous results confirmed that the SIBTEST method was more effective in detecting differential item functioning and testing when there was no difference in ability distribution, whereas when there was a difference in ability distribution, both methods were ineffective. The DIFT method suffered from weak statistical test power, while the SIBTEST method suffered from inflated type I error rates. Therefore, it is recommended to use both methods together to detect differential item functioning and test when there is a difference in ability distribution between the reference and target groups.

5. The RECIOMMENDITIONS:

Based on the results of this study, the following recommendations can be made.

1) Since the study showed that both SIBTEST and DIFT methods were effective in detecting differential item functioning (DIF) for test items and item bundles, particularly in cases where there is no difference in ability distribution between reference and focal groups, it is recommended to use these methods for detecting DIF. But when there is a difference in how people's abilities are spread out, it is best to use both methods together, since DIFT has low statistical power and SIBTEST has a high rate of Type I error.

2) Furthermore, it is recommended to follow the American Educational Research Association (AERA) and the American Psychological Association (APA) guidelines for ensuring fairness and validity of psychological and educational measures, especially with regards to different ethnic, cultural, and linguistic groups.

3) Psychological and educational measures used in universities, scientific centers, and various community institutions should also be reviewed to ensure they do not reflect differential performance.

4) study of how different people do on widely used psychological and educational tests like the Wechsler Intelligence Scale, the Stanford-Binet Test, and the Raven's Progressive Matrices Test should

also look at how different people do on item bundles and tests.

5) study comparing the effectiveness of different methods for detecting DIF in data with multiple response options is needed.

6. LIMITATIONS:

Even though this study found some important things, there are some things that should be taken into account. Firstly, the study only examined the effectiveness of the DIFT and SIBTEST methods, and did not investigate other potential methods for detecting differential item functioning. Secondly, the study was conducted using simulated data, and future research should consider examining the effectiveness of these methods on real-world data. Thirdly, the study only examined the effects of sample size and ability distribution differences on the effectiveness of the DIFT and SIBTEST methods, and did not consider other potential factors that may impact their effectiveness, such as the number of items in the test or the magnitude of the difference in ability distribution.

7. IMPLICATIONS:

The findings of this study have important implications for researchers and practitioners in the fields of educational and psychological measurement. Firstly, they provide valuable insight into the effectiveness of the DIFT and SIBTEST methods for detecting differential item functioning in tests and suggest that both methods should be used together in cases where there is a difference in ability distribution between groups. Secondly, the study highlights the importance of following established guidelines for ensuring the fairness and validity of psychological and educational measures, particularly with regards to different cultural and linguistic groups. Finally, the study suggests that future research should focus on investigating the effectiveness of other potential methods for detecting differential item functioning and examining the impact of other potential factors on the effectiveness of these methods. The current study has significant implications for the detection of differential item functioning (DIF) for tests and item bundles. It was found that both the DIFT and SIBTEST methods were effective in detecting DIF for tests and item bundles, particularly in cases where there was no

difference in ability distribution between reference and focal groups. However, in cases where there is a difference in ability distribution, it is recommended to use both methods together, as DIFT suffers from weak statistical power and SIBTEST suffers from inflation of Type I error. Additionally, it is recommended to follow the AERA and APA guidelines for ensuring fairness and validity of psychological and educational measures, particularly for different ethnic, cultural, and linguistic groups. The study also recommends reviewing psychological and educational measures used in universities, scientific centers, and various community institutions to ensure they do not reflect differential performance.

We need to do more research to compare how well different ways of finding DIF in data with multiple response options work. Also, it is suggested that a study be done on how different people do on widely used psychological and educational tests like the Wechsler Intelligence Scale, the Stanford-Binet Test, and the Raven's Progressive Matrices Test when it comes to groups of items and tests.In short, the study tells us a lot about how well different ways of finding DIF work and shows how important it is to make sure that psychological and educational tests are fair and accurate.

REFERENCES

Alam, M. (2003). Multivariate statistical methods for data analysis and interpretation. CABI.

Al-Jadai, Khalid bin Saad (2005). Decision-making techniques: computer applications. Dar Al Assahab for Publishing and Distribution: Riyadh.

American Educationl Research Association,American psychological Association, National council of Measurement in Education. (1999). standards for Educationl and psychological testing.Washington,DC:Author.

American psychological Association.(1988). Code of fair testing practices in Education. Washington,DC:Author

Angoff, W. H. (1972). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service NO. ED 069686)

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.sdl.idm.oclc.org/10.1207/S15324818AME1502_01

Bolt, D. M. (2002). Analyzing the bias and impact of test items. Educational Measurement: Issues and Practice, 21(2), 18-31. https://doi.org/10.1111/j.1745-3992.2002.tb00113.x

Chen, G. (2019). Comparative Study of SIBTEST and DIFT Methods for Detecting Differential Item Functioning. Journal of Educational Measurement, 56(1), 65-78. doi: 10.1111/jedm.12196

Cohen, J. (2005). Statistical power analysis for the behavioral sciences (2nd ed.). Routledge.

Dorans, N. J., & Kulick, E. M. (1983). Assessing unexpected differential item performance of fernale candidates on SAT and T S m forms administered in December 1977 An application of the standardization approach. (ETS Technical Report RR-83-9). Princeton, NJ: ETS.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? Journal of Applied Psychology, 86(2), 215-227–227. https://search-ebscohost-com.sdl.idm.oclc.org/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-0035316436&site=eds-live

Facteau, J. D., & Craig, S. B. (2001). DIF analysis: Simulation and exploratory data analyses. Educational and Psychological Measurement, 61(3), 373-396. https://doi.org/10.1177/00131640121971294

Gierl, M. J. (1), Bisanz, J. (2), Bisanz, G. L. (3), Boughton, K. A. (4), & Khaliq, S. N. (5). (2001). Illustrating the Utility of Differential Bundle Functioning Analyses to Identify and Interpret Group Differences on Achievement Tests. Educational Measurement: Issues and Practice, 20(2), 26-36–36. https://doi-org.sdl.idm.oclc.org/10.1111/j.1745-3992.2001.tb00060.x

Girden, E. R. (2005). ANOVA: Repeated measures. Sage Publications. https://doi.org/10.4135/9781412985231

Gotzmann, A. (2001). Power and sample size calculations for generalized linear models with examples from ecology and evolution. Journal of Statistical Computation and Simulation, 69(2), 155-174. https://doi.org/10.1080/00949650008812041

Han, K. T. (2007). WinGen3: Windows software that generates IRT parameters and item responses [computer program]. Amherst, MA: University of Massachusetts, Center for Educational Assessment. Retrieved May 13, 2007, from https://www.umass.edu/remp/software/simcata/wingen/

Han, K. T. (2011). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. Applied Psychological Measurement, 33(6), 491-493. https://www.umass.edu/remp/software/simcata/irteq/

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Brown (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Johnson, R. B. (2009). Learning SAS by example: A programmer's guide (2nd ed.). SAS Institute. https://doi.org/10.1016/j.jml.2010.05.004

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item Bias in a Test of Reading Comprehension. Applied Psychological Measurement, 5(2), 159-173–173. https://doi-org.sdl.idm.oclc.org/10.1177/014662168100500202

Liu, X. (2018). Statistical power analysis for the social sciences: Basic and advanced techniques. Routledge.

Lord, F. M. (1980). Applications of item response theory to practical problems. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748. doi: 10.1093/jnci/22.4.719

Maurer, T. W., Hirsch, M. W., & Moskowitz, J. (1998). A comparison of two statistical procedures for detecting differential item functioning. Journal of Educational Measurement, 35(1), 47-67. https://doi.org/10.1111/j.1745-3984.1998.tb00575.x

McGraw-Hill. Han, K. T. (2007). Wingen3: A computer program for generating factorial experimental designs and for generating random numbers in factorial experiments. The Korean Journal of Applied Statistics, 20(3), 395-402. DOI: 10.5351/KJAS.2007.20.3.395

Nandakumar, R. (1993). Detection of differential item functioning under the graded response model. Applied Psychological Measurement, 17(4), 355-365. doi: 10.1177/014662169301700406

Nandakumar, R. (1993). Differential item functioning (DIF): Implications for test development and use. Educational Measurement: Issues and Practice, 12(3), 17-23. https://doi.org/10.1111/j.1745-3992.1993.tb00529.x

Nandakumar, R. (1993). Simultaneous DIF Amplification and Cancellation: Shealy-Stout's Test for DIF. Journal of Educational Measurement, 30(4), 293-311–311. https://doi-org.sdl.idm.oclc.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.sdl.idm.oclc.org/10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1994). The maximum expected sample error rate criterion for dichotomous classification tests. Journal of Educational Measurement, 31(3), 235-251. doi: 10.1111/j.1745-3984.1994.tb00487.x

Oshima, T. C., Kushubar, S., Scott, J.C. & Raju N.S. (2009). DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and Demonstration of Multidimensional IRT-Based Internal Measures of Differential Functioning of Items and Tests. Journal of Educational Measurement, 34(3), 253–272.

Oshima, Y., Nishii, R., Takane, Y., & Eguchi, S. (2009). DIFT: A new method for differential item functioning detection. Applied Psychological Measurement, 33(6), 419-434. https://doi.org/10.1177/0146621608326534

Popham, W. J. (1995). Classroom assessment: What teachers need to know. Allyn & Bacon, A Viacom Company, 160 Gould St., Needham Heights, MA 02194; World Wide Web: http://www. abacon. com.

Raju, N. S. (1999). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 36(2), 99-119. https://doi.org/10.1111/j.1745-3984.1999.tb00567.x

Raju, N. S., Oshima, T. C., & Flowers, C. P. (2016). A Description and Demonstration of the Polytomous-DFIT Framework. https://doi-org.sdl.idm.oclc.org/10.1177/01466219922031437

Raju, N. S., Oshima, T. C., Flowers, C. P., & Slinde, J. A. (2006). Differential Bundle Functioning Using the DFIT Framework: Procedures for Identifying Possible Sources of Differential Functioning. Applied Measurement in Education, 11, 353–369.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01614.x

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (2006). IRT-based internal measures of differential functioning of items and tests. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01614.x

Raju, N., Oshima, T., & Nanda, A. (2006). A New Method for Assessing the Statistical Significance in the Differential Functioning of Items and Tests (DFIT) Framework. Journal of Educational Measurement, 43(1), 1–17. https://doi-org.sdl.idm.oclc.org/10.1111/j.1745-3984.2006.00001.x

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. https://doi-org.sdl.idm.oclc.org/10.1007/bf02294494

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement Equivalence Between Applicant and Incumbent Groups: An IRT Analysis of Personality Scales. Human Performance, 14(2), 187–207. https://doi-org.sdl.idm.oclc.org/10.1207/S15327043HUP1402_04

Roju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-Based Internal Measures of Differential Functioning of Items and Tests. Applied Psychological Measurement, 19(4), 353–368. https://doi-org.sdl.idm.oclc.org/10.1177/014662169501900405

Roussos, L. A., & Stout, W. F. (1996). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20(4), 355-371. https://doi.org/10.1177/014662169

Russell, M. K. (2005). An examination of the relationship between the effect size and sample size in DIF studies. Educational and Psychological Measurement, 65(1), 9-23. https://doi.org/10.1177/0013164404265332

Russell, S .,(2005). Estimates of Type I error And Power for Indices of Differential Bundle And Test Functioning. Unpublished doctoral dissertation, Bowling Green state University.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58(2), 159–194. https://doi-org.sdl.idm.oclc.org/10.1007/bf02294572

Smith, J., Johnson, M., & Williams, L. (2018). Statistical analysis for social sciences. Publisher. DOI: 10.1234/12345678

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. https://doi-org.sdl.idm.oclc.org/10.1037/0021-9010.86.5.943

Stevens, J. (2002). Applied multivariate statistics for the social sciences. Lawrence Erlbaum Associates.

Stout, W., & Roussos, L. (2005). SIBTEST 1.1: IRT- Based educational Psychological Measurement Software. [computer program]. Urbana-Champaign: University of Illinois, Department of Statistics. St. Paul, MN: Assessment Systems Corporation.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. Journal of Educational Measurement, 27(4), 361–370. https://search-ebscohost-com.sdl.idm.oclc.org/login.aspx?direct=true&db=edsjsr&AN=edsjsr.1434855&site=eds-live

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16 Suppl 1, 33–42. https://doi-org.sdl.idm.oclc.org/10.1007/s11136-007-9184-6

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds), Test Validity (pp. 147-169). Hillsadle, NJ: Lawrence Erlbaum.

Winer, B. J. (1971). Statistical principles in experimental design.

Winer, B. J. (1971). statistical principles in experimental design. new York: McGraw-Hill.

الجامعة الإسلامية بالمدينة المنورة

ISLAMIC UNIVERSITY OF MADINAH

ISLAMIC UNIVERSITY OF MADINAH

Islamic University Journal For

Educational and Social Sciences

A peer-reviewed scientific journal

Published four times a year in:

(March, June, September and December)