SAPIENZA
UNIVERSITÀ EDITRICE

E-ISSN 2724-2943
ISSN 2723-973X

Psychology Hub (2023)
XL, 2, 17-24

Article info

Submitted: 06 February 2023
Accepted: 09 May 2023
DOI: 10.13133/2724-2943/17992

The effect of Sample Size on Differential Item Functioning and Differential Distractor Functioning in multiple choice items

Hassan Alomari[1*], Mutasem Mohammad Akour[2] and Jehad Al ajlouni[3]

[1]*Department of Educational Psychology, Faculty of Educational Sciences, The University of Jordan,Amman,Jordan; Ha.omari@ju.edu.jo*

[2]*Department of Educational Psychology, Faculty of Educational Sciences, The Hashemite University, Zarqa, Jordan; mutasem@hu.edu.jo*

[3]*Ministry of education; jehadalajlouni@gmail.com*

Abstract

The current study investigated the effect of sample size on the number of items that show differential functioning (DIF) and the number of distractors that also show differential functioning (DDF) using the Mantel-Haenszel procedure. Data came from a national 8th grade mathematics exam that is composed of 40 multiple-choice items that was administered to 40,000 examinees. Eight samples with 250, 500 1250, 2500, 5000, 10000, 15000, and 20000 examinees were randomly selected. The findings of the current study indicated that increasing sample size increased the number of items detected with DIF and DDF. In addition, larger sample sizes are needed to detect items with nonuniform DIF and with negligible magnitude of DIF. Moreover, detecting DDF requires larger sample sizes as compared to the detection of DIF. Finally, sample size of 2,500 provided adequate number of items flagged with DIF (both types, and different magnitudes) and with DDF

Keywords: Uniform DIF, Nonuniform DIF, DDF, Alternatives, Multiple-Choice Items

*Corresponding author.
Hassan Alomari
Department of Educational psychology.
University of Jordan, Amman. Jordan
Address: Aljubeiha, Amman 11942
Jordan
Phone: +962795476044
E-mail: Ha.omari@ju.edu.jo
(H. Alomari)

Introduction

Test developers and researchers are interested in developing tests with good psychometric properties that can be used in making better decisions related to individuals' performance, achievement, or classification. The most important consideration in developing tests is validity, which indicates that the interpretations of test scores for proposed uses of tests are supported by theory or by some evidence (AERA et al., 2014). Various sources of evidence can be used in evaluating validity such as evidence based on test content, evidence based on internal structure, and evidence based on relations to other variables.

Based on the internal structure of a test, evidence of validity can be collected to show whether some items may behave differently for different subgroups of test takers (e.g., males and females). Differential Item Functioning (DIF) occurs when the probability of the correct answer to a particular item is different for individuals who belong to two distinct groups but are similar in ability (Penfield & Camilli, 2007). If DIF exists for a given test item, this imply that one group of respondents (usually referred to as the reference group) may have an unfair advantage of obtaining the correct answer to this item as compared to another group of respondents (usually referred to as the focal group). In this case, this item would function in favor of the reference group or differently against the focal group (Walker, 2011).

Several procedures were proposed in the literature in conducting DIF. Millsap (2011) categorized these procedures into two broad statistical frameworks: observed variable analysis and latent variable analysis. Examples of the observed variable analysis include the Mantel-Haenszel method (Holland & Thayer, 1988) and the logistic regression method (Swaminathan & Rogers, 1990). These types of analyses test DIF using an observed variable (i.e., sum scores or total scores) as the conditional variable. However, in the latent variable analysis item response are conditioned on the latent variable (i.e., the ability parameter). Similarly, there are many methods for detecting DIF under this framework, such as IRT methods (Lord, 1980) and multiple-group confirmatory factor analysis (Meredith, 1993).

The Mantel-Haenszel procedure is a non-parametric approach in detecting DIF in multiple-choice items that has gain popularity due to its simplicity, yielding meaningful results even with small sample sizes as compared to other detection procedure, and in providing an effect size measure in addition to the test of significance (Clauser & Mazor, 1998). Therefore, the present study investigated the effect of sample size on DIF when detected using the Mantel-Haenszel procedure.

The Mantel-Haenszel (MH) procedure

In the MH procedure, observed total scores are used to match examinees from both groups, reference and focal groups. At each score level (j) of a dichotomous-scored item, a 2(group) x 2 (item response) contingency table is created for each item as shown in Table 1.

Tab. 1. A 2x2 contingency table for each test item under each level of the total test score (j) used in the calculation of the MH test statistic.

Group	Item Score		Total
	1 (correct)	0 (incorrect)	
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Total	M_{1j}	M_{0j}	T_j

The MH statistic (Mantel & Haenszel, 1959) is distributed as chi-square with one degree of freedom, which tests the null hypothesis of no DIF against the alternative hypothesis:

$$H_1: \frac{P_{rj}}{Q_{rj}} = \alpha \frac{P_{fj}}{Q_{fj}} \quad , \quad \text{for } \alpha \neq 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

The parameter α is called the common odds ratio, which is the ratio of the odds of correct response for the reference group over that of the focal group (Penfield, 2003). Its estimate is given by:

$$\acute{\alpha} = \frac{\frac{\sum AjDj}{Tj}}{\frac{\sum BjCj}{Tj}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Holland and Thayer (1988) proposed the use of the log of common odds ratio:

$$\Delta_{MH} = -2.35 \ln \acute{\alpha} \quad \ldots\ldots\ldots\ldots\ldots \quad (3)$$

This statistic is asymptotically normally distributed, with negative values indicating that the reference group found that item easier than did the focal group. In addition, Δ_{MH} can be used to interpret the practical significance of DIF. Absolute values of Δ_{MH} less than 1 correspond to no or negligible DIF, equal to or more than 1 and less than 1.5 correspond to moderate DIF, and 1.5 or more correspond to high DIF (Zieky, 1993). For example, a value of Δ_{MH}=-1 means that the item was found to be more difficult for members of the focal group than for the comparable members of the reference group by an average of one Δ_{MH} point. In other words, a value of Δ_{MH}=-1 corresponds to a value of $\acute{\alpha}$ =1.53 which means that the odds of answering the item correctly for the reference group are more than 50% higher than the odds of answering it correctly for the focal group after conditioning on ability (Zwick, 2012).

Even though the MH procedure has shown to have high power in detecting uniform DIF, it has been shown that it was less powerful in detecting nonuniform DIF (Swaminathan & Rogers, 1990). Uniform DIF results when there is no interaction between group membership and ability level. That is, the difference in the probabilities of a correct answer for the two groups of examinees is the same at all ability levels. However, when an interaction exists between group membership and ability level nonuniform DIF is said to occur. In other words, the difference in the probabilities of a correct answer for the two groups of examinees is not the same at all ability levels. In the odds ratio context, DIF exists when $\alpha \neq 1$. If α remains constant across all ability levels then uniform DIF is present, but nonuniform DIF is present if α varies across the ability continuum (Penfield, 2003).

However, Penfield (2003) applied the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. He proposed the use of both the Breslow-Day test and the MH test such that the null hypothesis of no nonuniform DIF is retained if both tests lead to decisions of retaining the null hypothesis, while the null hypothesis is rejected if either test rejected the null hypothesis.

Differential Distractor Functioning

Usually, DIF analysis are followed by conducting differential distractor functioning (DDF) in dichotomous items (Penfield, 2008) and by differential step functioning in polytomous items (e.g., Akour et al., 2015). The existence of DIF in a multiple-choice item does not indicate in which response option the DIF effect occurs. Therefore, past research (e.g., Green et al., 1989; Penfield, 2008) proposed the framework of DDF which refers to the difference on probabilities of selecting each of the distractors for individuals who belong to two distinct groups but are similar in ability.

Examining DDF can help in identifying the causes of DIF. If DDF effects were consistent across all distractors, this may signal that the cause of DIF resides in the correct response or the item stem. That is, there might be a biasing factor in the correct response which leads to a consistent difference between the two groups in the probability of selecting each of the distractors. This can guide the test reviewer in conducting a content review on the correct response option or the item stem. On the other hand, if substantial DDF effect was associated with only one distractor, this may signal that the cause of DIF resides in that distractor. This can guide the test reviewer in conducting a content review on the properties of the specific distractor that could make this distractor more attractive or unattractive to the disadvantaged group (Penfield, 2008).

To examine DDF, several approaches have been suggested. For example, a log-linear approach (Green et al., 1989), the standardization method (Dorans et al., 1992), a mixture item response model (Bolt et al., 2001), and an odds ratio approach (Penfield, 2008). In the odds ratio approach, ability is divided into k ability strata. The conditional odds ratio across all strata of ability can be estimated using:

$$\acute{\alpha} = \frac{\frac{\sum R_{0kF_{jk}}}{Tjk}}{\frac{\sum R_{jkF_{0k}}}{Tjk}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

Where R_{0k} is the number of reference group members in the kth stratum who selected the correct response, R_{jk} is the number of reference group members in the kth stratum who selected the jth distractor, F_{0k} is the number of focal group members in the kth stratum who selected the correct response, and F_{jk} is the number of focal group members in the kth stratum who selected the jth distractor.

The natural logarithm of $\acute{\alpha}$ denoted by $\hat{\lambda}$ is an estimator of the DDF effect associated with the jth contrast function. If the value of $\hat{\lambda}$ is zero for a given distractor, then no DDF exists in that distractor. Values of $\hat{\lambda}$ other than zero indicates the presence of DDF in that distractor, with positive values indicate that DDF is favoring the reference group, while negative values indicate that DDF is favoring the focal group. A Z test statistic can be formed by dividing $\hat{\lambda}$ by its estimated standard error, which is distributed approximately as standard normal. This test statistic can be used to test the null hypothesis of no DDF for each distractor.

Sample size and MH

Large sample sizes improve the detection rates of the MH procedure. However, the MH procedure in detecting DIF has an advantage of requiring smaller sample sizes to yield meaningful results as compared to other detection methods, e.g., IRT methods. Mazor et al., (1992) pointed out that some studies considered samples of size 250 as small, whereas they asserted that other studies suggested that the MH procedure would be well functioning for samples of size 100. However, in their study Mazor et al. (1992) found that when using samples of size 2000 the MH procedure missed 25 to 30% of the differentially functioning items. Moreover, more than 50% of the differentially functioning items were missed when using samples of size 500 or less.

Using eight different sample sizes, Acar (2011) examined the number of detected DIF items using the hierarchical linear modeling procedure. Sample sizes started from 100 to, 11000 examinees. This study revealed that the number of detected items increased as sample size increased.

In a recent study, Ukanda et al. (2017) conducted a simulation study to examine the effectiveness of the MH in detecting DIF under three conditions of sample size (20, 60, and 1000), ability distribution, and test length. The findings of this study revealed that sample size had a significant effect on the detection of the three types of DIF items (A, B, and C) under the MH procedure. More DIF items were detected when sample size increased, while detecting more type C DIF items as compared to types A and B.

Different studies detected DDF in different disciplines (e.g., Ozdemir & AlGhamdi ,2022; Deng, 2020; Terzi & Suh, 2015; Tsaousis, Sideridis & Al-Saawi, 2018; Koon, 2010; Wang, 2000; Green, Crone, & Folk, 1989). However, no study investigated the effect of sample size on DDF. Therefore, the purpose of the current study was to explore the effect of sample size on the detection of DIF and DDF in multiple-choice items using the MH procedure.

Purpose of the current study

A good test depends on good items. One of the criteria of a good item is to be fair for distinct groups who are taking the test. Test developers are not only concerned with determining the type of the items, but also are concerned with determining the characteristics of the items that will be included in the test, such as: difficulty, discrimination, and to be free of DIF. The presence of DIF threatens validity, and thus infected items are candidates for removal.

Since a good test is constructed from good items, a good multiple-choice item depends on good distractors. Therefore, it is a good practice to examine distractors for differential functioning. The characteristics of the items and distractors are affected by the size of the samples. Therefore, the aim of this study was to investigate the effect of the sample size on the number of items that show DIF, both nonuniform and uniform types, and the number of distractors that also show DDF using the MH procedure.

Numerous studies in the literature (e.g., Acar, 2011; Ukanda et al., 2017) have examined the effect of sample size on the number of detected items with DIF. Such studies focused on uniform DIF and utilized simulated data. It is a good practice to have multiple evidence of the effect of sample size on the detection of DIF and DDF. Thus, the current study investigated whether using empirical data would support the findings of previous studies that relied on simulated data. It is

hoped that the current study would provide researchers with guidelines regarding the recommended sample size that can yield an adequate number of items detected with DIF and DDF. Additionally, it is hoped that this study would reveal if the detection of DIF requires similar sample sizes as compared to the detection of DDF, and if comparable sample sizes are required to detect both types of DIF, uniform and nonuniform.

Method

Participants

The data for this study came from a national 8th grade mathematics exam that is composed of 40 multiple-choice items. This test was administered by the Ministry of Education in Jordan to, approximately, 40,000 examinees. Mazor et al., (1992) pointed out that some studies considered samples of size 250 as small. Therefore, in the current study we selected different sample sizes from the population of examines starting with a sample of size 250. Another sample sizes of 500, 1250, 2500, 5000, 10000, 15000, and 20000 were also randomly selected from the same population.

Data Analysis

In the current study, we examined gender related DIF and DDF. Female students were considered as the focal group and male students were considered as the reference group. DIF and DDF analyses were conducted using the MH procedure. DIF analyses were conducted using the DIFAS computer program (Penfield, 2005), while the DDF analyses were conducted using the DDFS computer program (Penfield, 2010).

For DIF detection, statistical significance tests are not considered satisfactory for the interpretation of the practical significance of DIF (Camilli, 2007). Therefore, $\Delta_{MH} = -2.35 \ln \acute{\alpha}$ is an effect size measure that was used to supplement the chi-square test of statistical significance to test the null hypothesis of no DIF. Positive values of Δ_{MH} indicate that the item is favoring the focal group (i.e., females), whereas negative values indicate that the item is favoring the reference group (i.e., males). For the classification of the size of the effect size, we followed the three-category scheme proposed by Zieky (1993) and by Dorans and Holland (1993):

- Type A items: Negligible DIF, where the MH chi-square test is not significant or where Δ_{MH} is less than 1 in absolute value.
- Type B items: Moderate DIF, where the MH chi-square test is significant, and the effect size is between 1 and 1.5 in absolute values.
- Type C items: Large DIF, where the MH chi-square test is significant, and the effect size is greater than 1.5 in absolute value.

Dorans (as cited in Zwick, 2012) clarified the reasoning behind selecting the cutoffs of 1 and 1.5. He stated that a tolerated and a minimum undesirable difference in delta is 1 point. However, a difference in delts of 2 points or more should be avoided. The value of 1.5 represents the lower limit of the delta difference of 2 (1.5 to 2.5).

For the detection of nonuniform DIF, the Breslow-Day test of trend in odds ratio heterogeneity was used in addition to the MH test. The Breslow-Day test is a chi-square test with one degree of freedom. The null hypothesis of no nonuniform DIF is retained if both tests lead to decisions of retaining the null hypothesis, while the null hypothesis is rejected if either test rejected the null hypothesis.

To examine DDF, the natural logarithm of the odds ratio was utilized. If the value of this natural logarithm is zero for a given distractor, then no DDF exists in that distractor. Values of other than zero indicates the presence of DDF in that distractor. Positive values indicate that DDF is favoring the reference group, while negative values indicate that DDF is favoring the focal group. The nominal level of 0.05 was used for all analyses.

Results

The purpose of the current study was to examine the influence of sample size on the number of detected items with DIF and DDF according to students' gender. The MH Common Log-Odds Ratio method was utilized on a multiple-choice national math test for eight different sample sizes ranging between 250 and 20,000. The results of the current study are presented below.

The effect of sample size on DIF

The number of items detected with DIF, number of items detected with each type of DIF, and the number of items detected with DIF with different magnitudes at different sample sizes are presented in Table 2 and figures 1 to 3.

Tab. 2. Number of items showing DIF according to type and magnitude of DIF, and the number of items that show DDF at different sample sizes

Sample size	Number of DIF items	Type of DIF		Magnitude of DIF			Number of items with DDF
		uniform	nonuniform	Negligible	moderate	large	
250	7	6	1	0	0	7	8
500	12	12	0	0	6	6	15
1250	12	12	0	0	6	6	16
2500	23	13	10	14	5	4	41
5000	26	14	12	15	8	3	52
10000	30	18	12	24	3	3	63
15000	34	14	20	26	4	4	70
20000	34	10	24	27	4	3	73

Table 2 and Figure 1 shows that as sample size increased, the number of detected items with DIF increased significantly ($\chi 2 = 35.7$, $df = 7$, $p < 0.001$). The number of detected DIF items at sample size 20,000 was five times the number detected at the smallest sample size of 250. However, no gain was obtained in the number of detected items when sample size increased from 15,000 to 20,000.

Fig. 1. Number of detected items with DIF across different sample sizes

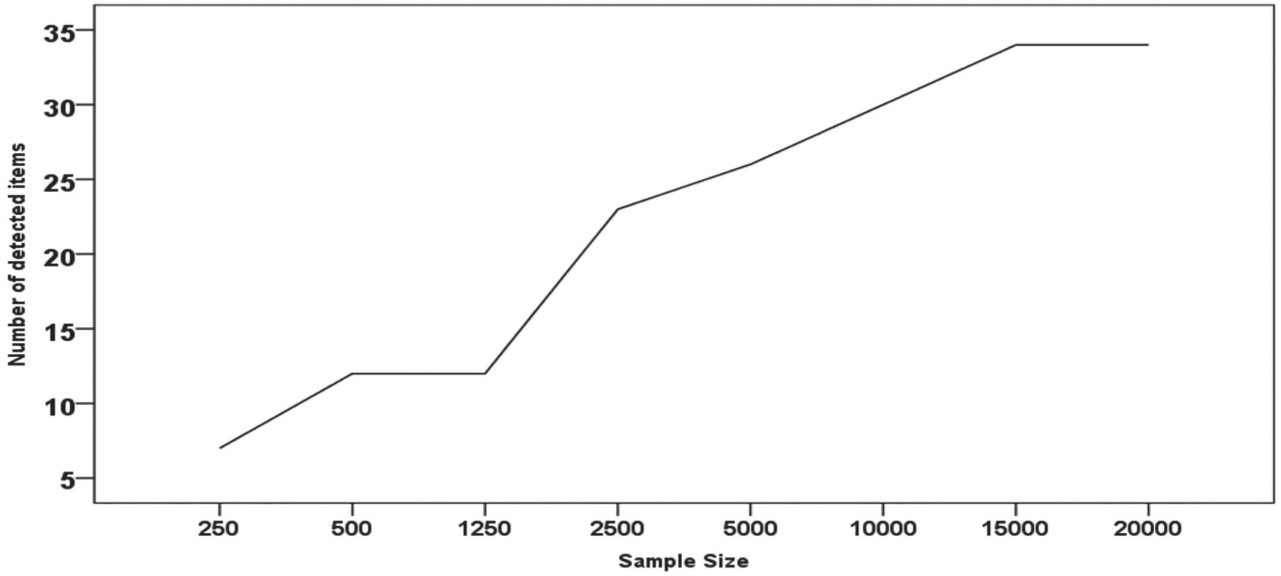


Fig. 2. Number of detected items with the two types of DIF across different sample sizes

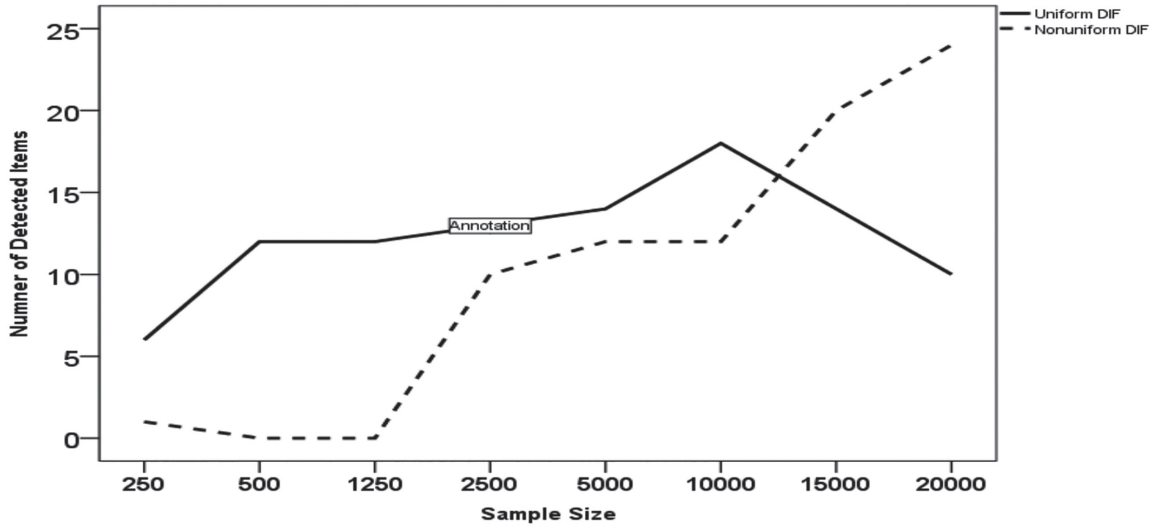
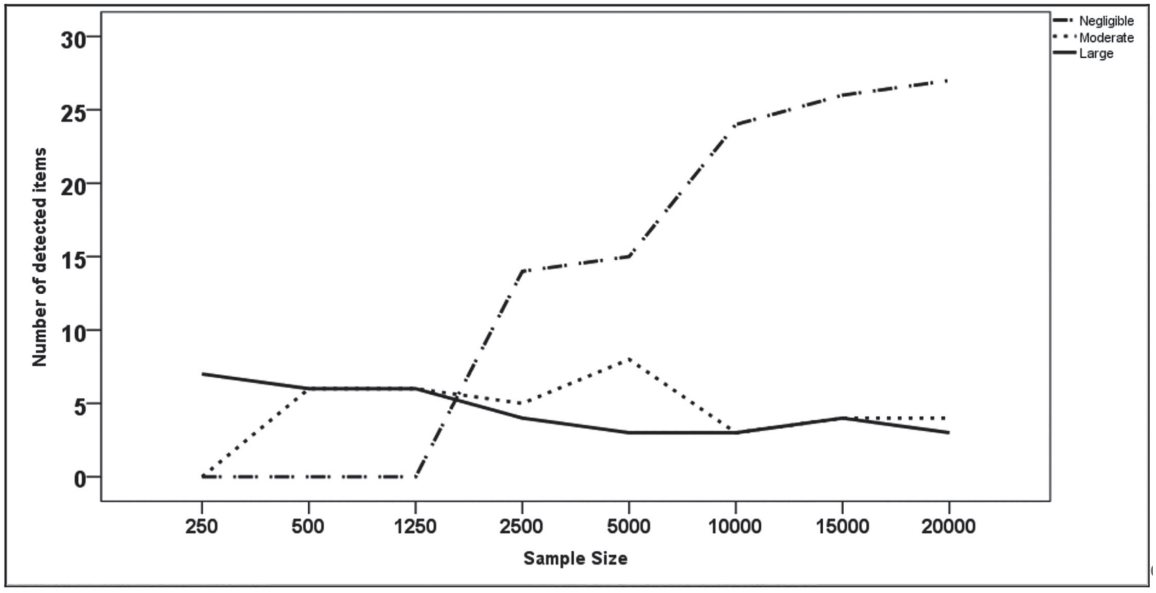


Fig. 3. Number of detected items with different magnitudes of DIF across different sample sizes



Regarding the number of items that were flagged with nonuniform DIF, Table 2 and Figure 2 indicate a significant increase as sample size increased from 250 to 20,000 ($\chi 2$ = 24.7, *df*=5, $p < 0.001$). However, the number of items detected with uniform DIF did not increase significantly as sample size increased ($\chi 2$ = 6.8, *df*=7, p = 0.45). Therefore, larger samples sizes are needed to detect items with nonuniform DIF.

Concerning the impact of sample size on the number of DIF detected items according to DIF magnitude (negligible, moderate, and large), Table 2 and Figure 3 show that larger sample sizes are needed to detect items with negligible DIF. Increasing sample size did not significantly affected the number of detected items with moderate DIF ($\chi 2$ = 3.8, *df* = 6, p = 0.8) or those with large DIF ($\chi 2$ = 4, *df* = 7, p = 0.8).

The effect of sample size on DDF

The number of items detected with DDF are presented in Table 2 and Figure 4 which show that the number of items flagged with DDF increased significantly as sample size increased ($\chi 2$ = 114.7, *df* = 7, $p < 0.001$).

Even though larger sample sizes detected more items flagged with DIF and DDF, Figure 4 shows that the effect of sample size on the detection rate was more evident for DDF than for DIF. Therefore, detection of DDF requires larger sample sizes as compared to the detection of DIF.

Discussion

The current study examined the relationship between sample size and the detection rate of multiple-choice items that exhibit DIF and DDF. Eight different sample sizes ranged from small sample with size of 250 to large one with size of 20,000 were utilized in this study. DIF and DDF analyses were conducted via the MH common log odds ratio test statistic.

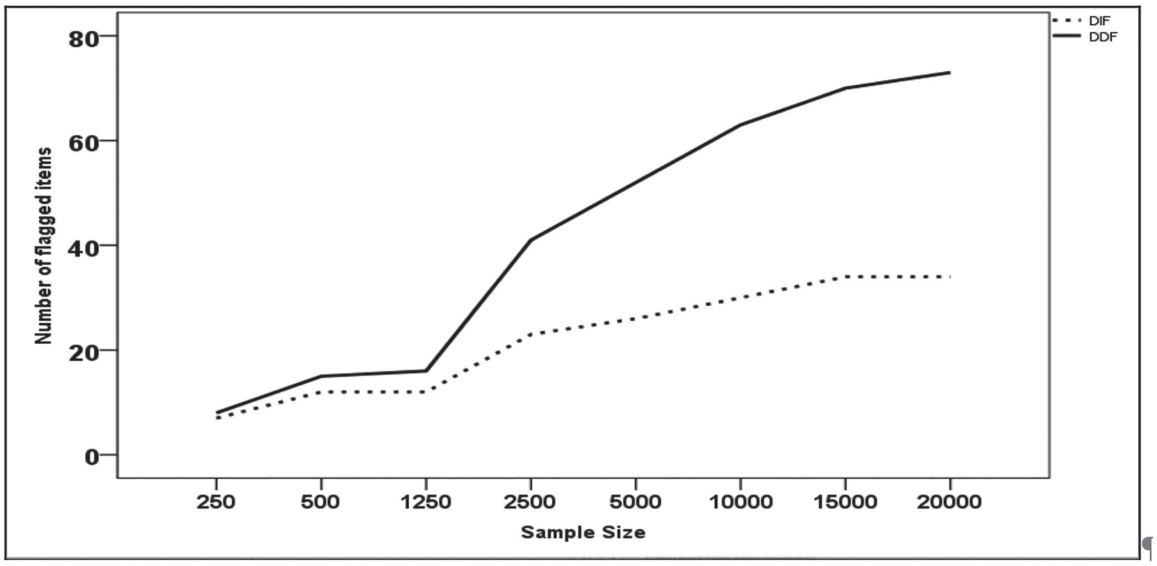
One of the main findings of the current study is that there was a positive relationship between sample size and the detection of items with DIF. Increasing sample size resulted in an increase in the number of items flagged with DIF. Larger sample sizes yielded more powerful test statistic. This finding aligns with the findings of previous research (e.g., Mazor et al., 1992 and Ukanda et al., 2017). The number of detected items with DIF doubled when the sample size also doubled (when sample size increased from 1,250 to 2,500 examinees). Therefore, the sample size of 2,500 examinees may be considered adequate for detecting most items with DIF. Those items which were not flagged with DIF at this sample size may be difficult or poorly discriminating items.

However, the finding concerning the number of items detected with nonuniform DIF was questionable for samples with size less than 2,500. Larger sample sizes are needed to detect non uniform DIF as compared to uniform DIF. It seems that sample size of 2,500 was also adequate in detecting uniform and nonuniform DIF. There were not much gain in the number of items detected with nonuniform DIF for samples with size larger than 2,500.

Moreover, items with negligible magnitude of DIF started to be detected at sample size 2,500. Larger sample sizes resulted in more items detected with negligible amount of DIF. This is inconsistent with the findings of Ukanda et al. (2017) where larger sample sizes detected more items with large amount of DIF. According to the findings of the current study, there were not much gain in the number of items detected with moderate and large magnitudes of DIF for samples with size larger than 2,500.

In addition, more items flagged with DDF were also detected as sample size increase. However, as sample size increased the number of items detected with DDF was larger than that detected with DIF. The effect of sample size on DDF detection was more obvious than its effect on the detection of DIF. The number of items flagged with DDF increased by more than the double when the sample size increased from 1,250 to 2,500. It seems that sample size of 2,500 is the smallest sample

Fig. 4. Number of items detected with DIF and DDF across different sample sizes



size that enables test developers and practitioners to flag most items with DIF (uniform and nonuniform, and with different magnitudes of DIF), and with DDF.

The widespread use of all kinds of achievement and psychological tests in making decisions related to the classification of individuals and their success or failure, prompts test developers to pay attention to the fairness and invariance issues (Wiberg, 2007). Unfairness exists when there are differential performance between two groups on a given test item in the conditional probability of the correct answer (Penfield, 2008). Accordingly, the use of large samples would help test developers in detecting those items that exhibit DIF and DDF, and thus removing such items to enhance the validity of the interpretation made based on test scores.

One limitation to the current study is that samples of size less than 250 was not used. In addition, only one method for DIF and DDF detection was used, the MH procedure. Therefore, the findings of the current study may not be generalizable to sample sizes smaller than 250 and/or to other methods used in DIF and DDF detection. It is recommended for future research to study the effect of smaller sample sizes on the detection of DDF, and using other methods of detection, such as the regression-based approaches. Furthermore, research needed to examine the effect of sample size on DIF and DDF according to some factors such as: type of test, types of groups used to detect DIF (other than male vs. female focus), item difficulty, item discrimination, and ability distributions.

Finally, the current study analyzed real data, but the generalizability of the results in terms of sample sizes needed to increase sensitivity need to be supplemented by some simulation results that consider specified levels of DIF and DDF and investigate sensitivity across simulated samples of the same sample sizes used for the actual data.

Author Contributions
All authors contributed equally to this manuscript.

Conflict of interest:
The authors declare no conflict of interests.

Funding:
The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval:
The current study used an archival data that came from a test administered by the Ministry of Education in Jordan. The Ministry agreed to provide the authors with the needed data, and the authors did not make any modifications to it.

References

Acar, T. (2011). Sample Size in Differential Item Functioning: An Application of Hierarchical Linear Modeling. *Educational Sciences: Theory & Practice*, *11*(1), 284-288.

AERA, APA, & NCME. (2014). Standards for Educational and Psychological Testing: National Council on Measurement in Education. American Educational Research Association.

Akour, Mutasem, Sabah, S., & Hammouri, H. (2015). Net and Global Differential Item Functioning in PISA Polytomously Scored Science Items: Application of the Differential Step Functioning Framework. *Journal of Psychoeducational Assessment, 33*(2), 166-176. https://doi.org/10.1177/0734282914541337

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381–409. https://doi.org/10.3102/10769986026004381

Camilli, G. (2007). Test fairness. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed., Vol. 4, pp. 221-256). Westport: American Council on Education& Praeger Publishers.

Clauser, B.E. and Mazor, K.M. (1998), Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*, *17*: 31-44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Deng, J. (2020). The Relationship between Differential Distractor Function (DDF) and Differential Item Functioning (DIF): If DDF Occurs, Must DIF Occur? Lawrence, KS: Doctoral dissertation, University of Kansas. docs/OnlinePubs/PARA/examining/examiningDDFreport.pdf

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309–319. https://doi.org/10.1111/j.1745-3984.1992.tb00379.x

Green, B. F., Crone, C. R., and Folk, V. G. (1989). A Method for Studying Differential Distractor Functioning. J. Educ. Meas. 26 (2), 147–160. doi:10.1111/j.1745-3984.1989.tb00325.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Koon, S. (2010). *A Comparison of Methods for Detecting Differential Distractor Functioning* (Publication No. FSU_migr_etd-2840) [Doctoral Dissertation, Florida State University]. ProQuest. http://purl.flvc.org/fsu/fd/FSU_migr_etd-2840

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* 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, 719-748.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*(2), 443–451. https://do.org/10.1177/0013164492052002020

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org//10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Ozdemir B and AlGhamdi HM (2022) Investigating the Distractors to Explain DIF Effects Across Gender in Large Scale Tests with Non-Linear Logistic Regression Models. Front. Educ. 6:748884. https://doi.org/10.3389/feduc.2021.748884

Penfield, R. D. (2003). Applying the Breslow-Day Test of Trend in Odds Ratio Heterogeneity to the Analysis of Nonuniform DIP. *Alberta Journal of Educational Research, 49*(3), 231–243.

Penfield, R. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, *29*(2), 150–151. https://doi.org/10.1177/0146621603260686

Penfield, R. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45*(3), 247-269.

Penfield, R. (2010). DDFS: Differential Distractor Functioning Software. Applied Psychological Measurement, 34(8), 646–647. https://doi.org/10.1177/0146621610375690

Penfield, R., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay, & C. Rao (Eds.), *Handbook of statistics, Volume 26: Psychometrics* (pp. 125-167). Elsevier.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tsaousis, I., Sideridis, G., and Al-Saawi, F. (2018). Differential Distractor Functioning as a Method for Explaining DIF: The Case of a National Admissions Test in Saudi Arabia. Int. J. Test. 18 (1), 1–26. https://doi.org/10.1080/15305058.2017.1345914

Terzi, R. and Suh, Y. (2015), An Odds Ratio Approach for Detecting DDF Under the Nested Logit Modeling Framework. Journal of Educational Measurement, 52: 376-398. https://doi.org/10.1111/jedm.12091

Ukanda, F., Othuon, L., Agak, J., & Oleche, P. (2017). Effect of sample size, ability distribution and test length on detection of differential item functioning using Mantel-Haenszel statistic. *International Journal of Education and Research*, *5*(5), 91-104.

Walker, C. (2011). What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment*, *29*(4) 364–376. https://doi.org//10.1177/0734282911406666

Wang, W. C. (2000). Factorial Modeling of Differential Distractor Functioning in Multiple-Choice Items. J. Appl. Meas. 1, 238–256

Wiberg, Marie. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretical comparison of methods*. EM No 60. Retreived August 20, 2017, from www.edusci.umuse/digital Assets/159/59534-em-no-60.

Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P.W. Holland ve H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Erlbaum.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, 2012(1), i–30. doi:10.1002/j.2333-8504.2012.tb02290.x