Calderón *et al. Large-scale Assessments in Education* (2026) 14:23
https://doi.org/10.1186/s40536-026-00294-x

Large-scale Assessments in Education

RESEARCH Open Access

Differential item functioning analysis in large scale assessments: a case study for DIF in SABER 11

John Alexander Calderón[1], Nelson Andrés Rodríguez[1] and Víctor H. Cervantes[2*]

*Correspondence:
Víctor H. Cervantes
victorhc@illinois.edu
[1]Instituto Colombiano para la Evaluación de la Educación - Icfes, Bogotá, Colombia
[2]Department of Psychology, University of Illinois at Urbana-Champaign, 603, E. Daniel Str., Champaign, IL 61820, USA

Abstract

Background Absence of differential item functioning (DIF) is an important piece of evidence to support inferences based on group comparisons of test results. We illustrate how to tailor the DIF identification process following the guiding questions proposed by Sireci and Rios for large scale assessment's (LSA) specific characteristics by examining DIF between two test forms of the Mathematics test of a Colombian LSA (SABER 11).

Methods We investigate the performance of the non-compensatory DIF (NCDIF) index and the Mantel–Haenszel (MH) DIF procedure under large sample sizes and sample size ratios (up to 1:25), and the performance of effect size guidelines under these conditions. These simulations were needed to adequately address the guiding questions for DIF analyses of SABER 11. DIF analyses of SABER 11 test forms were conducted in light of the results of these simulations.

Results Type I error is affected, for both procedures, by both the sample size and sample size ratio, as well as by the magnitude of impact between the groups. The joint use of the effect size guidelines helps mitigate this issue without much loss of power given the large sample sizes involved. The DIF analyses of the Mathematics test forms of SABER 11 provide robust evidence that the inferences derived from score comparisons are fair.

Conclusions Beyond the immediate implications for the use of SABER 11 tests, the presented case study may help guide practitioners in the assessment of DIF by illustrating how to perform several of the steps involved. Moreover, the simulation studies shed new insights into the frequentist behavior of the two DIF indices under conditions that had not been previously explored but which are applicable to many LSA. Additionally, the results indicate that simulation studies examining the performance of NCDIF, MH, and possibly any DIF statistic, should implement realistic item parameter pools and not only sanitized well-distributed sets of item parameters.

Keywords Differential item functioning, DIF, NCDIF, Mantel–Haenszel, Validity, Large scale assessments

Springer

Introduction

Results across different test forms ought to be comparable; that is, any differences in scores should be ascribable only to the unreliability of the measurements. However, even in such an ideal case, there will be a need to address those differences. For example, in large scale assessments (LSAs) of educational achievement, there may be slight differences in difficulty between two test forms in spite of the care given during test development to equally cover the test specifications on both forms. And in general, there may be additional factors affecting the test scores.

Test fairness is one of the most important practical and ethical considerations related to the comparability of scores on different test forms. At its core, a test would be fair when the interpretations of the test scores obtained by examinees in different groups are comparable given equal underlying abilities. The use of separate test forms naturally splits the population of interest and when these forms are used on different occasions, this separation can hardly be taken as random. This is regularly the case in LSAs; different test forms are employed on different dates. The question of whether the results of an assessment are fair towards particular groups goes back to at least the 1960s (Gómez Benito et al., 2018; Teresi et al., 2013); although, most notably after the "Golden rule" settlement of 1984 (see e.g. Linn & Drasgow, 1987), there was an increased interest in properly and technically approaching test fairness, and especially in being able to distinguish results that showed real differences between groups of those where results are biased against a particular group. As a consequence, the concept of *Differential Item Functioning* (DIF) was introduced.

DIF occurs when, conditioned upon equal levels of the measured trait, the statistical properties of an item differ across two or more groups. Most frequently, DIF is considered when comparing only two groups: a group that can be taken as the standard, generally a "majority" group that might be thought of as being favored, known as the *reference* group; and the special group of interest, denoted as the *focal* group, which would be comprised of examinees from the "minority" or the possibly disfavored group (Sireci & Rios, 2013).

We examined the Mathematics test of an LSA that utilizes Item Response Theory (IRT) for its scoring: Colombia's end of high-school examination SABER 11. This examination takes place twice a year for different subpopulations of end of high-school students and, consequently, different test forms are used for each of the two testing dates. Assessing whether DIF is present across the two test forms constitutes an important way of gathering evidence of the validity of the results from this LSA. The investigation of DIF in specific contexts encompasses several practical decisions, as outlined by Hambleton (2006) and Sireci and Rios (2013). Although the decisions that need to be made are more or less shared by all testing endeavors, which choices and how to make them must be bespoken to the specific testing circumstances. Although we present the process for only SABER 11 Math tests, this process can be applied to make these decisions on other LSAs. Along the way, we identify some choices for which there is a lack of sufficient empirical support for making decisions regarding the very large sample sizes usual in LSAs and the potentially large differences in the sample size of the compared groups. Thus, we also present a set of simulation studies that help guide these decisions in our analysis of SABER 11 and that may be useful to guide the analyses of DIF for similar LSAs.

Test scoring and DIF definition for IRT models

In IRT, the probabilities of the responses to an item are assumed to be a function of a latent trait (ability) that is measured by the test and item specific characteristics, such as its difficulty. For items scored dichotomously (e.g., correct/incorrect), an IRT model frequently used in LSAs is the two-parameter logistic model (2PL). This model is used for scaling and scoring of prominent LSAs, such as National Assessment for Educational Progress (NAEP; Mislevy et al., 1992), Trends in International Mathematics and Science Study (TIMSS; von Davier, 2020), and Progress in International Reading Literacy Study (PIRLS; Foy & Yin, 2017; TIMSS & PIRLS International Study Center, 2017).

Under a 2PL IRT model, the probability that an examinee whose ability is θ_i gives a correct response (encoded as 1) to item k of a test is modeled as

$$\Pr(X_{ik} = 1|\theta_i, a_k, b_k) = \frac{1}{1 + \exp\left[-Da_k(\theta_i - b_k)\right]}, \tag{1}$$

where X_{ik} denotes the coded response of examinee i to item k, D is a scaling constant which can equal 1 or 1.702 to define the θ ability scale on the logistic or normal metric,[1] a_k is the discrimination or slope parameter of item k, and b_k is the difficulty or threshold parameter of item k. The curve of this function is known as the Item Characteristic Curve (ICC) for the item.

Under IRT, the definition of DIF specifies as a difference in the item parameters of the groups. That is, if for each of the groups the IRT model appropriately describes the relation between the latent trait and the probability with which an examinee answers the item correctly, as in for example Eq. (1), and DIF occurs on item k, then at least one of the item parameters (a_k, b_k) will have a different value in some group. Within IRT models, the differences in item parameters map directly into what is now the traditional classification of DIF types (see e.g., Gómez Benito et al., 2018). Uniform DIF corresponds to differences in the item difficulties across groups and it is described as a "constant" difference along the latent trait (see panel a of Fig. 1). Non-Uniform DIF is associated with differences in the item discrimination, this type of DIF may be described in terms of an interaction between the latent trait levels and the group to which the examinee belongs

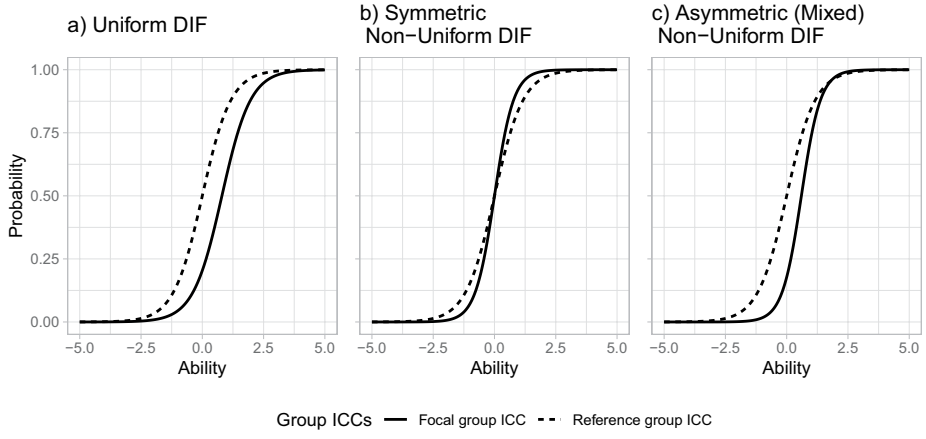

Fig. 1 Examples of DIF in 2PL IRT with varying item parameters across groups

[1] The value 1.702 constrains the metric to approximate the parameters that would be obtained on a probit IRT model rather than the logistic model written in Eq. (1).

(panel b of Fig. 1). When both item difficulty and discrimination differ across groups, it may known as mixed DIF or as asymmetric non-uniform DIF; this is illustrated in panel c of Fig. 1.

SABER 11

SABER 11 is a Colombian national examination whose target population comprises students who finish high school. It consists of a standardized battery of tests that aims to serve as an indicator of secondary education quality and an exam for higher education admission. An examinee sits in two sessions during which they respond to five tests: *mathematics, critical reading, science, social studies, and English*. All tests consist of between 41 and 58 multiple-choice items. The National Institute for Educational Evaluation (Icfes) is in charge of the design and administration of this assessment. Starting in the year 2000, the scaling and scoring of these tests has used IRT models.

The academic year for most schools in Colombia starts in late January or early February (AY1). However, a percentage of schools start their academic year in August (AY2). While most students complete high school under the AY1 schedule, about 3.2% of students complete high school under an AY2 schedule. In order to meet the testing requirements of all students finishing their secondary studies, there are two mass testing dates each year; the first in March, catering mostly to students in AY2, and the second in September, for students in AY1. In addition to the large disparity in the number of examinees who take the tests on each occasion, there are substantial contextual differences between the two examined subpopulations. Schools that follow AY2 are almost entirely (at least 97%) private schools and include all schools offering an International Baccalaureate curriculum; whereas more than 98.5% of public schools in the country follow AY1, which constitutes about 80% of schools in this AY. Hence, there is a concomitant difference in the distribution of important sociodemographic variables that impact learning outcomes among examinees of these two AYs, such as the expected socioeconomic status, the teacher remuneration, and the curricular flexibility in the schools. The average scores of students testing in the AY2 test date are thus higher than those of students tested in the AY1 test date (see e.g., Instituto Colombiano para la Evaluación de la Educación [Icfes], 2020). Clearly, when contrasting these two subpopulations, students in AY1 constitute the majoritarian group. It is, therefore, crucial to guarantee that scores across these two testing dates are comparable. To achieve this goal, the test forms need to be shown free of DIF by appropriate analyses of the test forms.

The remainder of this paper is organized as follows: we first present the set of guiding decisions proposed by Sireci and Rios (2013) for conducting DIF analysis. For each of these decisions, we introduce results from the literature to guide the decision-making, the relevant aspects of the SABER 11 examination pertaining to the decision, and if a decision can be reached, the choice that was made. For two of the decisions, there was no sufficient empirical evidence to make a choice. A set of simulation studies, presented in the subsequent section, were conducted to help make the decisions which befit the large sample sizes and sample size ratios of the two AYs in SABER 11, as well as the use of effect sizes as part of the DIF identification procedures. The results of the simulation studies are presented next, and afterward, we present the results of the DIF analysis on two forms of a SABER 11 test following the choices made in the previous sections. We

close this paper with a discussion of the lessons learnt from these analyses for the detection of DIF in LSAs.

What decisions to make when conducting DIF analysis

Hambleton (2006) argued that there are multiple practical consideration when conducting DIF studies that go beyond the selection of an appropriate statistical procedure. More recently, Sireci and Rios (2013) expanded upon these recommendations and suggested a set of decisions that ought to be pondered in the process of identifying DIF items. The analysis of DIF as part of the overarching gathering of evidence of validity takes place in the context of the particular test, its uses, and how its scores are interpreted. We will follow the set of five guiding questions from Sireci and Rios (2013) and illustrate how we have answered them for studying DIF in a Colombian LSA—SABER 11. The five questions involve making decisions about which DIF method(s) to use, how to define the comparison groups along with sampling (including sample sizes) from each group, the finding of the conditioning or matching variable, the criteria for employing effect sizes, and the level of analysis at which differences will be explored.

Choice of DIF method

The first decision in performing DIF analyses of a test is the choice of which method or methods to employ in its detection. There are several reviews of different methods for identifying DIF, such as Jodoin and Girel (2001), Kim et al. (2007), Khoeruroh and Retnawati (2020), Sireci and Rios (2013), Teresi et al. (2013) and Wu and Tsai (2010), among others. Considerations about the nature of the data and the psychometric model ought to play an important role in deciding what DIF method to apply (Sireci & Rios, 2013; Teresi et al., 2013). For instance, a test whose results are produced using a IRT model, may be suitably analyzed for DIF using an IRT-based DIF method. In the analyses below, we will employ the non-compensaory DIF (NCDIF) index from Raju's Differential Functioning of Items and Tests (DFIT) framework (Raju et al., 1995) which peruses the IRT model used in test scoring.

The most commonly cited disadvantage of using IRT-based DIF methods is the need of large sample sizes (Martinková et al., 2017; Sireci & Rios, 2013). This limitation is almost absent in LSAs which by definition have a large number of examinees; thus, facilitating the use of IRT-based procedures with LSAs. Among the IRT-based procedures, the DFIT framework (Oshima & Morris, 2008; Raju et al., 1995) has kept a steady development since its inception. The main DIF statistic within the framework is the NCDIF index which generalizes Raju's unsigned area measure (Raju, 1988). Moreover, unlike other IRT-based methods (Sireci & Rios, 2013), there are effect size guidelines for the values of the NCDIF index (Wright & Oshima, 2015).

At its core, the NCDIF index quantifies the area between the ICC of the focal and the reference groups (see Fig. 1) but weights different regions according to the distribution of the focal group. Formally, it is defined as the following expected value,

$$\mathrm{NCDIF}_i = \mathrm{E_F}\left([\mathrm{P}_{i\mathrm{F}}(\theta) - \mathrm{P}_{i\mathrm{R}}(\theta)]^2\right), \tag{2}$$

where $\mathrm{E_F}$ denotes the expectation operator with respect to the distribution of the latent trait for the focal group, $\mathrm{P}_{i\mathrm{F}}$ is the probability that given the parameters of item i for the focal group (F) an examinee with ability θ answers correctly item i, and $\mathrm{P}_{i\mathrm{R}}$

denotes the analogous probability using the parameters for the reference group (R). Figure 2 illustrates how the regions of greater density in the focal group receive a greater weight in the comparison of the two ICCs to compute the NCDIF index. It is easy to see that under usual IRT models, the difference $\mathrm{P}_{i\mathrm{F}}(\theta) - \mathrm{P}_{i\mathrm{R}}(\theta)$ equals zero only when the item parameters for both the focal and reference groups are equal. Thus, NCDIF will be positive whenever item parameters differ between both groups. This property is expected for any DIF statistic but, in addition, it should be noted that the magnitude of the NCDIF index is also influenced by the shape and location of the density function of the focal group. Figure 2 illustrates how the weighting of different ability regions, and hence the magnitude of the NCDIF index, is affected across focal groups with different distributions.

This is an interesting property in that it highlights the cases in which the differences in item parameters between the two groups have a larger (smaller) effect on the test results of the focal group. Inasmuch as the focal group is an actual group of interest, this property incorporates into the parameter of interest the concerns about what evidence there is on how the interpretations of the test results may or may not be affected by the presence of DIF for that group. However, when none of the two compared groups can be singled out as of special interest, this property might not be as desirable.

The test for detecting DIF with the NCDIF index is performed by approximating the statistic distribution by parametric bootstrap (see Cervantes, 2006; Oshima et al., 2017). This procedure is denoted as the Item Parameter Replication (IPR) method in the DFIT framework. For brevity, we do not discuss further procedures, although we would remind the reader of an additional recommendation regarding the choice of DIF method: that one should not limit DIF analyses to using a single procedure (Hambleton, 2006; Sireci & Rios, 2013). To compare the performance of the NCDIF method, we also report the results of the Mantel–Haenszel (MH) procedure for the main simulation as well as for the real data analysis. Computations for MH are carried out in R using the *difMH* function from the difR package (Magis et al., 2010). This function calculates the

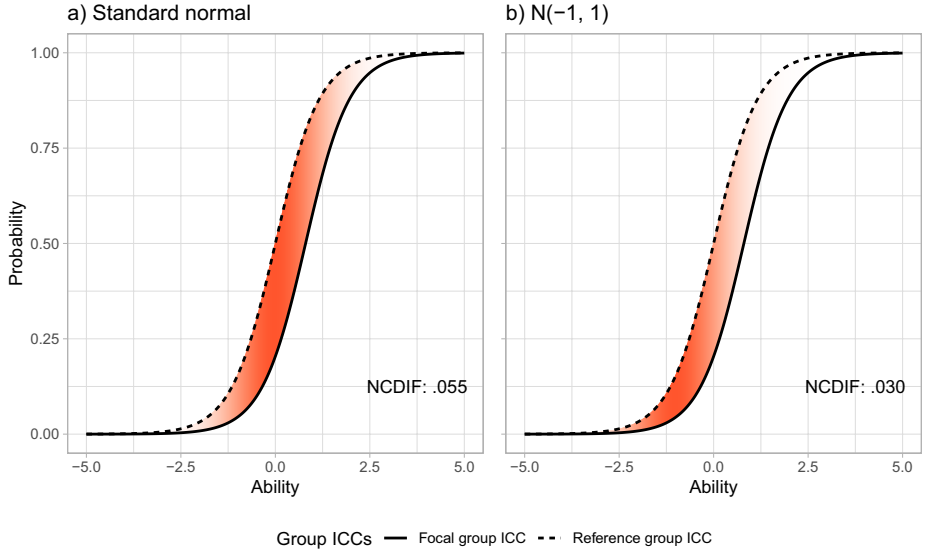

Fig. 2 Effect of different densities on NCDIF. *Note*. In (**a**), a larger area between the two ICCs receives more weight than in (**b**). Consequently, in this example item, the NCDIF magnitude will be larger for a focal group with a standard normal ability distribution than for a group with a normal distribution shifted by -1

MH log-odds parameter based on the response contingency tables at several levels of the raw total score.

Choice of comparison groups and sample sizes

Inherent to the decision to analyse DIF for some test is the selection of the groups that will be compared. Many DIF detection procedures may be affected both by the sample size of each group and by the relative sizes of the groups. Most studies find better statistical behavior of DIF procedures when the sample sizes of the groups are similar to each other (Cohen & Kim, 1993; Herrera & Gómez, 2008; Kim & Nering, 2007; Oshima et al., 2006, 2015).

For SABER 11, we need to consider both the sample size and relative sizes of the groups in order to equate the scores of two different testing dates. As noted above, there are two separate AY schedules followed by schools in Colombia. Given the differing testing dates when examinees from each AY sit for the examination, parallel test forms, with the same structure and a set of overlapping common items, are used. However, only a small percentage of examinees complete their studies under AY2. If we look at a single test form of one area subject from the testing dates for AY1 and for AY2, we see that there are approximately 40,000 and 1500 student examinees, respectively. Thus, if all examinees are included into DIF analyses, then both group sample size and sample size ratio for our data will be much more extreme than what is found in any study about the statistical properties of the NCDIF detection procedures we have seen.[2] Additionally, the ability distributions of both groups may considerably differ as explained above in the SABER 11 section. Given the lack of information about the performance of NCDIF under comparable conditions, we defer the specifics of sample sizes to the results of the simulation studies presented below.

Choice of attribute matching criterion

All DIF detection procedures require that some matching criterion is found upon which conditional comparisons of the groups is made. In general, for DIF procedures based on IRT, such as NCDIF, the matching variable is the latent ability of examinees on a common scale. However, the inclusion of DIF items in the computation of the latent ability compromises the matching. Hence, the recommendation is to follow some purification procedure that excludes likely DIF items from the equating of the latent trait across the groups (Hambleton, 2006; Hidalgo & López-Pina, 2004; Sireci & Rios, 2013). Such purification may be performed either as a 2-stage procedure, as an iterative one, or as a stepwise process. In a 2-stage purification, the first stage consists of linking the scales using all items to detect possible DIF items. In the second stage, the researcher removes those just identified items from the set of common items and repeats the linking procedure with the purified set. The items (from the full set of items) that are detected after the second linking are those deemed to have DIF (Hambleton, 2006). The iterative purification procedure repeats the elimination of items to link the scales for as long as the set of items identified with DIF varies between consecutive stages (Fidalgo, 2000; Hidalgo & López-Pina, 2004). In a stepwise purification (Khalid & Glas, 2014), only k items, $k = 1, 2$, are

[2] Few studies consider large sample size ratios. Herrera and Gómez (2008) look into ratios up to 1:10 but relatively small sample sizes, and Cuevas and Cervantes (2012) consider ratios up to 1:250. Both of these studies analyzed only the performance of the logistic regression procedure for DIF detection.

no longer considered common between consecutive stages; Khalid and Glas (2014) suggested stopping the iterations when linking results do not vary considerably between stages. Seybert and Stark (2012) found that 2-stage and iterative purification improve the performance of DIF detection using the NCDIF index, and that 2-stage purification produced similar results to the iterative procedure. Similar results were found on the performance of MH with respect to the various purification approaches (Fidalgo, 2000). Considering these results, we will use 2-stage purification for the analysis contrasting test forms of SABER 11 for the two AYs.

Concomitant with the purification procedure that is used (or not), the researcher needs to decide upon the linking procedure to use for matching the criterion. Stocking and Lord (1983) proposed an estimating procedure where the necessary linking constants are found by minimizing the squared difference between the sum of the re-scaled ICCs. The other procedure based on the comparison of the ICCs, proposed by Haebara (1980) can be defined as finding the constants that will minimize the sum of the NCDIF indices of the items given the re-scaled item parameters. Comparisons of the two procedures have shown that they perform similarly, and that they outperform other procedures to estimate the linking constants (Kim & Kolen, 2006; 2007; Kolen & Brennan, 2014). Considering this similarity and that previous studies on the performance of NCDIF have used Stocking and Lord's method (Kim & Nering, 2007; Oshima & Davey, 2000; Oshima et al., 2015; Seybert & Stark, 2012), we will use the Stocking and Lord linking procedure in our study. For MH, we employed the raw total score as the attribute matching variable (Magis et al., 2010).

Choice of effect size measures

As it is the case in general, DIF detection procedures are more sensitive to smaller differences between items across groups as sample sizes increase. To avoid undue detection of items that present too small differences, it is recommended that some effect size measure is used to complement the results of the statistical tests (Hambleton, 2006; Sireci & Rios, 2013). The traditional classification of effect size measures for MH comes from the Educational Testing Service (ETS) Delta measure (Δ_{MH}, see e.g., Dorans & Holland, 1992; Holland & Thayer, 1988). Using these effect size guidelines, items are classified as: (A) No DIF or negligible DIF if the MH statistical test is nonsignificant or $|\Delta_{\mathrm{MH}}| \leq 1$, (B) Moderate DIF if the MH test is significant and $1 < |\Delta_{\mathrm{MH}}| \leq 1.5$; and (C) Large DIF if the MH test is significant and $|\Delta_{\mathrm{MH}}| > 1.5$. Also, as we will compute the NCDIF index and there are proposed guidelines to use the statistic's value as its own effect size measure (Wright & Oshima, 2015), we will use them in our study.

The guidelines that Wright and Oshima (2015) proposed include a set of cut points for the magnitude of NCDIF that relate it with the classification used for Δ_{MH} as item parameters vary. We shall look at the effect of using this effect size classification together with the IPR detection procedure across varying sample sizes and ratios to decide whether to use it for the analysis of SABER 11 tests.

Choice of level of analysis

The last decision to be made, following Sireci and Rios (2013), is to choose the level of analysis. This refers to whether one wishes to identify differences at the item level or at the aggregate effect on the full test scores. The latter level of analysis refers to Differential

Test Functioning (DTF) and no DIF procedure targets it directly. Even when some items favor the focal group and other items favor the reference group, the differences may not necessarily compensate on the total score. Nor is it the case that because no items show significant DIF, there is no aggregate effect. Within the DFIT framework, there are indexes that aim explicitly to identify both DTF and how much each item within a given set contributes to that DTF. In this paper we focus on the item level analysis of DIF without looking at how DIF may aggregate. We agree with the spirit of what Sireci and Rios (2013) recommend; that researchers ought to "consider aggregate effects of item-level DIF results and eliminate items when there seems to be an imbalance of direction across items flagged for DIF" (p. 184). However, we also think that some more nuance is required for LSAs. LSAs use an item pool from which several different forms are created, rather than consisting of one or few fixed forms. These item pools are in constant renovation. Therefore, it is important to identify which characteristics of an item may determine the presence of DIF between specific groups, so that future item writers know how to avoid them when writing new items.

Research goals

Our main goal is to tailor the DIF analysis protocol for the tests of SABER 11 in order to answer whether there is evidence of items with DIF in those tests.

In the review of each of the decisions to be made for DIF analyses with respect to the characteristics of SABER 11 tests, there were two for which the choice was not clear. The goals of this study, in support of our main goal, are to answer these two questions:

1. What sample sizes are more appropriate to use to analyze SABER 11?
2. How to incorporate the effect size measure in the identification process?

To answer these questions, we conducted a series of simulation studies. Two preliminary simulations are described in the supplementary material A and the third simulation study is reported in the next section. With these simulation studies, we specifically aim to define whether to use full data samples in spite of unequal sample sizes or to achieve equal groups via subsampling, and whether to include the effect size classification at the DIF identification stage or not.

Simulation study

Test characteristics

The item parameters for two test forms (40 items each) with an anchor of half of their items were specified and held constant for the simulation. Such test design coincides with the item overlap found between consecutive test dates in SABER 11. The values of the item parameters were sampled from the pool of operational Mathematics items used in SABER 11 and are presented in Table 1. The common items between both forms are those numbered 1 to 20. Items numbered 21 to 40 are unique to Form A and used to simulate responses for the reference group; and the item parameters of the focal group test form are shown in the Form B columns and numbered 41 to 60. For this study, there were three DIF conditions (NDIF): 0, 2, and 4 DIF items. Changes in the item parameters for the DIF conditions are shown in Table 2. The magnitude of the changes in item parameters match those of the preliminary simulations presented in the supplementary material A. An additional ad hoc manipulation of the number of items with DIF and the

Table 1 Item parameters for Study 3

Item	Common Items		Unique Items					
	Form A and B		Form A			Form B		
	a	b	Item	a	b	Item	a	b
1	.62	− .71	21	.57	.05	41	.60	− .66
2	.58	− 1.30	22	.71	− .31	42	.88	− .73
3	1.09	− .88	23	.58	− .48	43	.49	.92
4	.37	− 1.87	24	.87	.65	44	.54	.63
5	.46	.53	25	.52	.43	45	.42	.48
6	.57	− .75	26	1.16	− .39	46	.58	− .33
7	.45	− 2.00	27	.33	1.43	47	.79	− .82
8	.61	.17	28	.45	1.36	48	.93	− 1.10
9	.64	.04	29	.35	− .86	49	.81	− .86
10	.44	1.30	30	.23	1.95	50	.75	− 1.41
11	.52	.44	31	.80	− .28	51	.15	1.25
12	.45	− .35	32	.58	.05	52	.20	2.88
13	.76	.10	33	.25	2.84	53	1.01	− .77
14	.53	− .24	34	.43	.54	54	.58	− .27
15	.72	− 1.02	35	.66	.14	55	.91	− .71
16	.46	− .85	36	.33	3.25	56	.86	− 1.87
17	.42	.60	37	.74	− 1.60	57	.27	.55
18	.22	2.93	38	.42	.21	58	.26	− 1.33
19	.54	.38	39	.96	− .51	59	1.49	− 1.55
20	.84	− 1.47	40	.87	− .24	60	.69	− .20

Table 2 Item parameters for DIF conditions

Item	Number of DIF items					
	0 items		2 items		4 items	
	a	b	a	b	a	b
5	.46	.53	.53	1.38	.53	1.38
6	.57	− .75	.66	− 1.60	.66	− 1.60
15	.72	− 1.02			1.09	− .62
16	.46	− .85			.69	− 1.25

magnitude of its differences was performed at the request of an anonymous reviewer. Its results are reported with the label of '10 DIF items' and the item parameters and NCDIF values are reported in the supplementary material B, Table B2. The number of items with DIF and their magnitude closely matches what was effectively found in the data analysis of the Mathematics test presented in Section "SABER 11 analysis".

Design

The four factors manipulated in all the simulation studies were (a) sample size and sample size ratio, (b) impact, (c) purification, and (d) use of effect size measure. Sample size and ratio, and use of effect size are needed to guide the decisions to perform DIF analyses for the SABER 11 test, as previously described. The presence of impact, that is true differences in the ability distributions between the focal and reference group, and the purification of the matching variable, are known to affect the performance of most, if not all, DIF procedures.

Sample size and sample size ratio (SSR)

For all simulations, we need to manipulate the sample size and sample size ratio to levels similar to those of SABER 11. We took five pairs of sample sizes of the reference and focal group. In the first one, we generated a population of 40,000 for both the reference and focal groups; these sample sizes reflect what would be expected for two AY1 testing dates of, say, two consecutive years. In the second, a focal group of 1500 was used instead; this condition reflects the expected sizes of two consecutive testing dates, one of AY1 and one of AY2. The third and fourth levels of sample sizes consisted of equal sized groups with either 1500 or 3000 individuals each. The fifth pair consisted of a reference group of 3000 individuals and a focal group of 1500. Several studies report better results for a variety of DIF statistics when sample sizes are similar than when they differ (Herrera & Gómez, 2008; Huggins, 2014; Oshima & Morris, 2008; Wright & Oshima, 2015); albeit, the ratios considered were not as large as those seen in our application. To compare with these studies and obtain smaller sample size ratios, we have included the conditions with pairs of a focal group of 1500 and a smaller than 40,000 sample for the reference group. The condition with 3000 individuals in both groups would reflect the approach of subsampling the condition with 40,000 subjects in each group. These conditions seek to assess whether it would be preferable to have less extreme sample size ratios, perhaps even equal sample sizes, which could lead to obtaining subsamples from the reference group for the analysis of real SABER 11 data. Hence, it is most relevant to our problem to examine how the index performs under the 1500:40000 condition compared with the other conditions with 1500 examinees on the focal group, most importantly, the 1500:1500 condition.

Impact (Focal group mean ability, FGM)

It is usually expected that the focal group is disadvantaged with respect to the reference group. However, as described in Section "SABER 11", the opposite is expected in SABER 11 when comparing AY2 against AY1. To consider this situation, but also compare the results from these simulations to previous studies, where the focal group may be disadvantaged, we included both positive and negative shifts in the mean ability of the focal group. The distributions of the latent trait were manipulated as follows. A vector of true ability values for the reference groups of each sample size (40,000, 3000, and 1500) was generated from a standard normal distribution. For the focal groups, several vectors were generated to include conditions without impact (same distribution as the reference group) and with it (distributions shifted by some amount). For each sample size 40,000, 3000, and 1500, a vector of true ability values for the focal group was generated from seven possible distributions. A standard normal, two normal distributions with a small shift on the mean (either .25 or – .25), two with a medium shift ($\pm.5$), and two normal distributions with a large shift ($\pm.8$). The historic differences observed across subject tests between AY1 and AY2 are around medium-sized differences ($\sim .5$) favoring AY2. Given these manipulations and the item parameters used in each study, the population DIF values are depicted in Table 3.

Purification and use of effect size measure (PES)

The matching criterion was defined by the latent trait estimated using a 2PL model (with constant $D = 1.702$) at each step of a 2-stage purification procedure. That is, we

Table 3 Population values of NCDIF index and Δ_{MH}

Item	NCDIF per Focal group mean							
	– .80	**– .50**	**– .25**	**.00**	**.25**	**.50**	**.80**	Δ_{MH}
5	.01886	.02143	.02325	.02465	.02553	.02580	.02529	*-1.12893*
6	.03518	.03417	.03226	.02958	.02635	.02283	.01854	*1.43945*
15	.01700	.01412	.01151	*.00899*	*.00673*	**.00485**	**.00315**	**-0.31089**
16	.01214	.01397	.01512	.01581	.01600	.01567	.01469	*1.02685*

When the true population value is classified as negligible (category A), its value is presented in bold text. Similarly, italics are used for items classified as having moderate DIF (category B); and no special format when they are classified in category C (large DIF)

examined the performance before and after conducting the purification of the matching variable.

Final DIF identification was conducted with and without the use of the effect sizes described in Section "Choice of effect size measures." Considering that not all items were classified with at least moderate DIF in all conditions, we will report both cases in which items in category A (negligible DIF) are taken as DIF items (as they strictly are) and in which they are taken as no DIF items (for not being practically important). The effect size guidelines were used to inform the detection only after the purification step.

These two factors were jointly manipulated to obtain three levels: without purification nor use of the effect size measure, after purification without using the effect size, and after purification and using the effect size.

Experimental replications and outcomes

For each of the simulation studies, we computed the Type I Error and the power rates, estimated across 200 replications per condition.[3] We also computed for each replication the correlation between the estimated discrimination and difficulty parameters with the generating ones in Tables A1, A2, and 1. Lastly, for each condition of each study, we estimated the bias and the mean squared error of the focal group mean obtained after equating[4] removing the items identified using both the Item Parameter Replication test and the effect size measures.

Simulation Results

Parameter recovery

The correlation between estimated and original item difficulties were larger than .97 for all groups, in all conditions. For the item discrimination parameters, these correlations were always larger than .72. Values below .8 were only found in some conditions with a sample size of 1500. The bias and mean square errors of focal group mean ability were small for all conditions and with little variability across conditions. The highest value of the bias was.001 in absolute value, and the MSE ranged between .00001 and .00121.

Type I error rates and power

Figures 3 and 4 show the Type I error rates for each condition for the NCDIF and MH methods, respectively. In some panels, error rates for the corresponding conditions were greater than 0.3 and are therefore not shown. Figures 5 and 6 present correct detection

[3] Atar (2007) reports that there is little gain in the precision of the estimates of these rates beyond 200 replications.

[4] The scale of ability parameters set the reference group mean to 0 makes this an estimate of the bias and of the mean square error of the group differences.

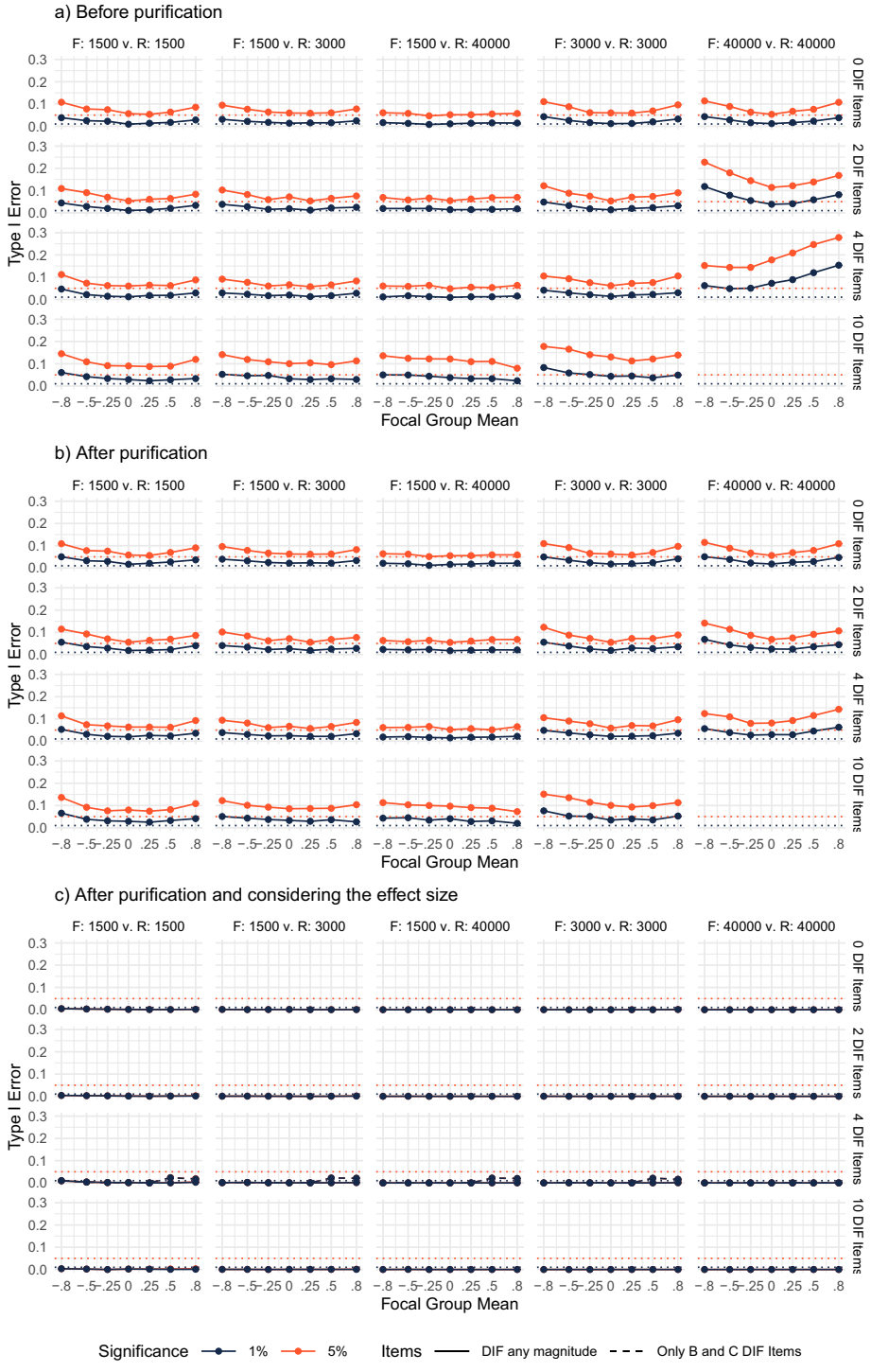

Fig. 3 NCDIF Type I Error. *Note*. Each small multiple presents the profile of the Type I Error rate of NCDIF for an experimental condition across the levels of impact at two significance levels. The profiles in (**c**) coincide for both significance levels. In (**a**) and (**b**), the error rates for the '10 DIF Items' conditions with sample sizes of 40,000 in both groups are larger than 0.3

rates. Note that in the respective subfigures c) of Figs. 3, 4, 5 and 6, the solid and dashed lines represent rates computed over different sets of items. The solid lines represent rates where any difference in the item parameters was taken as DIF, whereas the dashed lines show rates where only items classified by the respective population values (cf. Table A3)

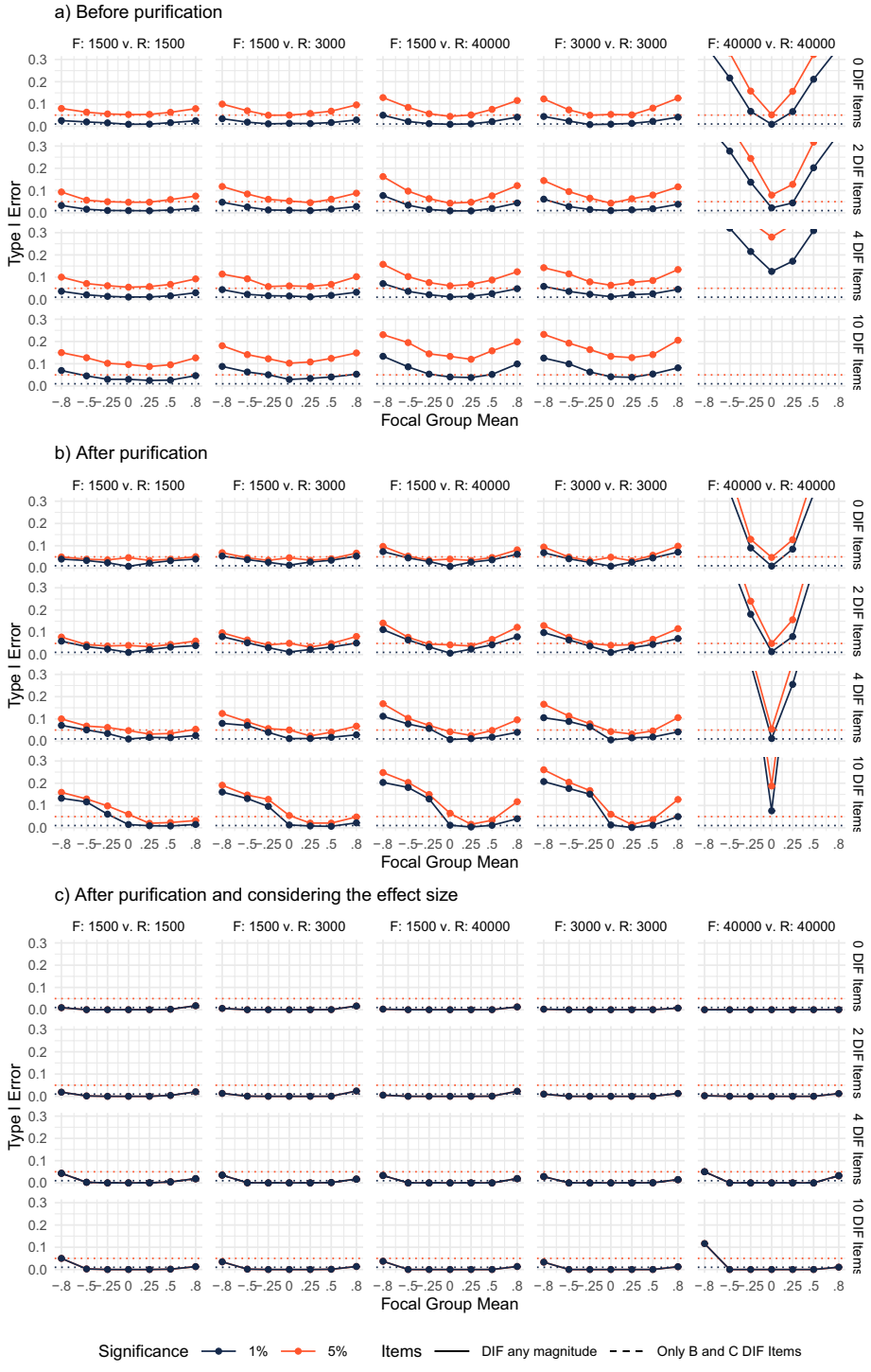

Fig. 4 MH Type I Error. *Note*. All elements as in Fig. 3 except that profiles depict the Type I Error rates for the MH procedure

as B or C items. Consequently, rates computed for only B and C DIF items may be greater than rates computed for all items. This increment on the rates arises because items classified with negligible DIF are more likely to be classified on the B or C categories than items with no DIF, increasing the Type I Error rate when they are no longer considered as part of the set of items with (large enough) DIF. Similarly, the items with

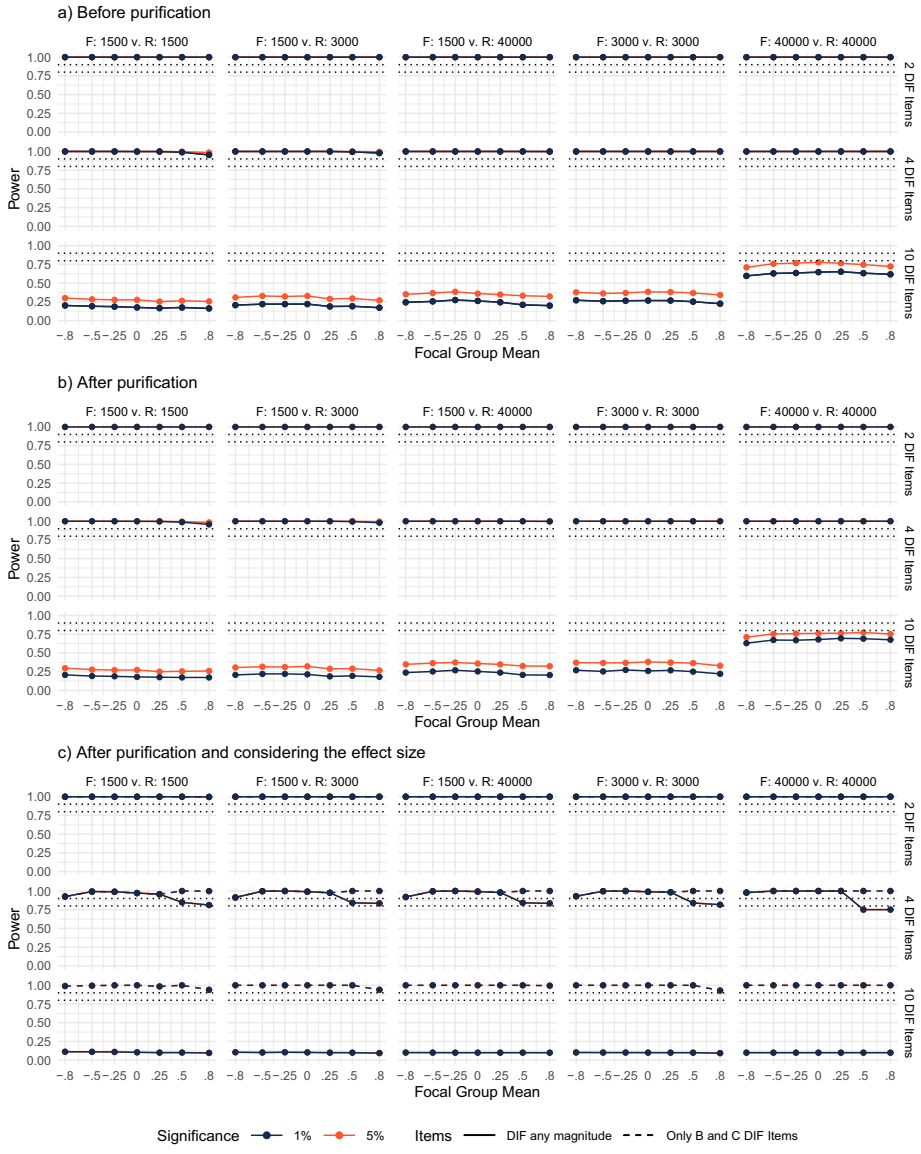

Fig. 5 NCDIF Power. *Note*. All elements as in Fig. 3 except that profiles depict the NCDIF correct detection rates

negligible are less likely to be classified as B or C items than items with larger true differences, hence improving the computed power when they are excluded.

The rates of incorrect detection before purification are much larger for the conditions where sample sizes were 40,000 for both focal and reference groups. They were also generally larger for the MH procedure than for NCDIF. For both statistics, Type I Error was larger for conditions with impact, although its effect appears to be asymmetric with respect to the sign of the group differences. Additionally, when error rates were above the nominal significance value, they were usually larger for the MH procedure.

After purification, the Type I Error rates of the NCDIF were closer to the nominal values, although still higher for the 40000:40000 conditions. The rates for the MH procedure were also better controlled after purification but less so than the NCDIF rates, and showed a marked asymmetry with respect to the sign of impact—with larger rates for negative focal group means.

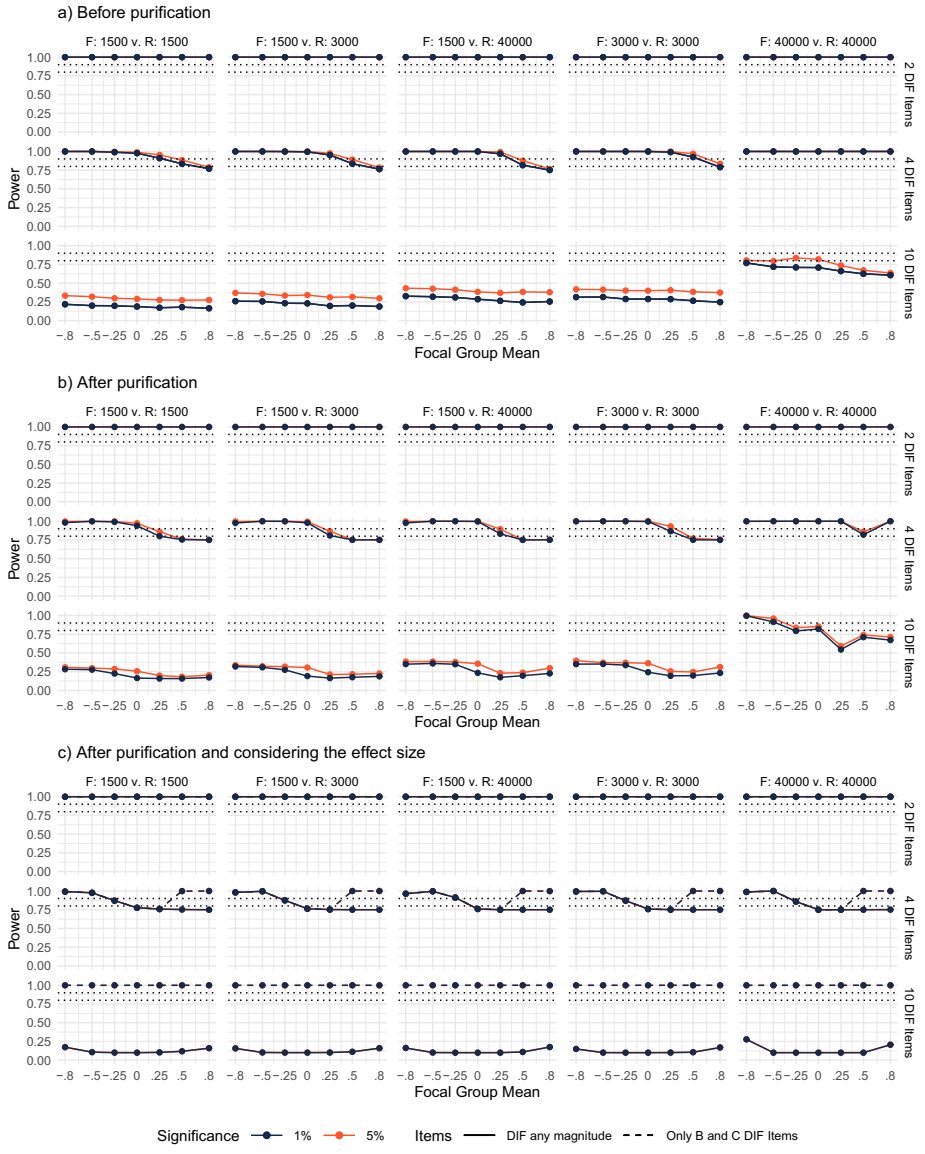

Fig. 6 MH Power. *Note.* All elements as in Fig. 3 except that profiles depict the correct detection rates for the MH procedure

In almost all cases, the use of the effect size measure together with the statistical test rendered zero or near zero rates of incorrectly detected items. The exceptions were the error rates for the MH procedure for some of the conditions with large impact ($\pm.8$).

These results indicate interactions among the manipulated factors. An analysis of variance of the Type I Error rates confirms that almost all interactions among the factors are statistically significant.[5] For the NCDIF index, the three-way interaction between sample size and sample size ratio (SSR), impact (FGM), and use of purification and effect size measure (PES), as well as the interaction between number of items with DIF on the test (NDIF), impact (FGM), and use of purification and effect size measure (PES), were not significant ($F_{84,144} = .8649, p = .7654$). All other interaction terms were statistically significant ($F_{176,228} = 102.33, p \approx 0$). Regarding the MH procedure, all three-way

[5] Note that there is only one rate per cell; hence, the interaction of all terms is confounded with the residual term.

Table 4 NCDIF Type I Error rates ANOVA

Source	Sum. Sq	df	F	*p*	η^2_G (LB)
SSR	0.8751	4	1310.8	0.0000	0.958 (0.951)
FGM	0.0260	6	25.9	0.0000	0.406 (0.319)
NDIF	0.6503	3	1298.6	0.0000	0.945 (0.935)
PES	1.2198	2	3654.1	0.0000	0.970 (0.964)
SSR:FGM	0.0080	24	2.0	0.0054	0.173 (0.029)
SSR:NDIF	1.3981	12	698.0	0.0000	0.974 (0.969)
SSR:PES	0.4866	8	364.4	0.0000	0.927 (0.915)
FGM:NDIF	0.0182	18	6.1	0.0000	0.324 (0.204)
FGM:PES	0.0128	12	6.4	0.0000	0.252 (0.145)
NDIF:PES	0.3393	6	338.8	0.0000	0.899 (0.881)
SSR:FGM:NDIF	0.0347	72	2.9	0.0000	0.477 (0.276)
SSR:NDIF:PES	0.7083	24	176.8	0.0000	0.949 (0.940)
Residuals	0.0381	228			

Sum. Sq.: Type II sums of squares; df: degrees of freedom; F: Value of the F ratio; p: p-value; η^2_G Generalized η^2 (eta-squared) effect size measure; LB: 95% confidence lower bound of the η^2

Table 5 MH Type I Error rates ANOVA

Source	Sum. Sq	df	F	*p*	η^2_G (LB)
SSR	4.6383	4	774.5	0.0000	0.956 (0.945)
FGM	0.5952	6	66.3	0.0000	0.734 (0.674)
NDIF	0.5807	3	129.3	0.0000	0.729 (0.670)
PES	2.2157	2	739.9	0.0000	0.911 (0.891)
SSR:FGM	0.5749	24	16.0	0.0000	0.727 (0.651)
SSR:NDIF	0.6775	12	37.7	0.0000	0.759 (0.700)
SSR:PES	2.3492	8	196.1	0.0000	0.916 (0.896)
FGM:NDIF	0.0366	18	1.4	0.1610	0.145 (0.000)
FGM:PES	0.2186	12	12.2	0.0000	0.503 (0.386)
NDIF:PES	0.2241	6	24.9	0.0000	0.510 (0.408)
SSR:FGM:NDIF	0.1722	72	1.6	0.0090	0.444 (0.072)
SSR:FGM:PES	0.4722	48	6.6	0.0000	0.687 (0.570)
SSR:NDIF:PES	0.2944	24	8.2	0.0000	0.577 (0.453)
FGM:NDIF:PES	0.1015	36	1.9	0.0048	0.320 (0.062)
Residuals	0.2156	144			

Sum. Sq.: Type II sums of squares; df: degrees of freedom; F: Value of the F ratio; p: p-value; η^2_G Generalized η^2 (eta-squared) effect size measure; LB: 95% confidence lower bound of the η^2

interactions were statistically significant ($F_{260,144} = 13.156, p \approx 0$). Table 4 summarizes the effects of the experimental factors on the error rates of the NCDIF index and Table 5 those on the rates of MH.

The pattern of significant and non-significant interactions for NCDIF indicates that the interaction between impact (FGM) and the use of purification and effect size (PES) does not interact with the other factors, whereas the interaction between samples size and sample size ratio (SSR) with number of DIF items (NDIF) does. This pattern may be interpreted in Fig. 3 by considering each set of small multiples sharing the same sample size ratio and number of DIF items; when looking at those three panels, the shape of the profiles of incorrect detections is approximately the same, but the profiles are not flat, nor are they at the same height. The significant interactions for the MH Type I error rates, instead indicate that for the same set of panels, the profiles have different shapes. Among these interactions, the interaction of the sample size ratio, the number of DIF items, and the use of the effect size had the largest effect on the error rates associated

with the NCDIF ($\eta^2_G = .949$). For the MH, the largest effect was associated with the interaction of the sample size ratio, impact magnitude, and the use of the effect size measure ($\eta^2_G = .687$)

Regarding the specific contrast across the 1500:1500 and 1500:40000 sample size and sample size ratio levels when the focal group mean is .5, we found that the Type I Error does not statistically differ between the two conditions (NCDIF: – .004, SE = .008, $t_{12} = -.505$, $p = 0.623$; MH: .014, SE = .006, $t_{12} = 2.179$, $p = 0.0499$) across the conditions varying the number of DIF items for the 1500:40000 condition after purification and controlling for the use of the effect size measure.

Lastly, we should note that overall, NCDIF Type I error rates were more favorable than those for the MH procedure. Table 6 presents the cross-classification of the two DIF detection indices for items without DIF. Most items were correctly classified by the two approaches; however, it is clear that in the case of disagreement between the two classifications, if an item is detected using the MH but not by NCDIF, it is more likely that the item does not truly have DIF. Analogously, it is very unlikely that an item without DIF is classified by both procedures as presenting non-neglibible DIF (B or C).

The correct detection rates for conditions with uniform DIF items only (2 DIF items) were perfect for both procedures, regardless of purification or the use of the effect size measures. For the NCDIF index, when the effect size measure was included in the decision, that is, only when items identified by the hypothesis test after purification and classified with non-negligible DIF (B or C categories) were detected, power decreased to the proportion of items whose generating population statistics led to a B or C category. These rates became near perfect again if only the items with true non-negligible DIF were considered in its computation. For the MH statistic, the rates were not always near perfect regardless of the true classification into negligible or not; however, rates were always close to .80 or above. Power was generally higher for the MH procedure when the group differences favored the reference group. Using the effect size measure in the detection decision also improved the power of the MH procedure but not as consistently as it did for the NCDIF index. Recall from Table 3 that for NCDIF there were only items classified as negligible in the 4 DIF items conditions when the focal group mean was at least .5 and in all conditions with 10 DIF items; for MH that was the case for all conditions with 4 or 10 DIF items. Tables 7 and 8 present the ANOVA tables for the effects of the experimental factors on the power of the NCDIF and the MH detections, respectively. For both statistics, only the three-way interaction between sample size and sample size ratio (SSR), impact (FGM), and use of purification and effect size measure (PES) was not significant ($F_{48,96} = 1.0821, p = .3656$), all other interaction were significant ($F_{156,144} = 778.82, p \approx 0$). For both NCDIF and MH, the largest effect was due to the interaction of the sample size ratio, the number of DIF items, and the use of the effect

Table 6 Cross-classification of items free of DIF by NCDIF and MH

MH	NCDIF				
	No DIF	A	B	C	Total
No DIF	83.4	3.67	.0691	.0014	87.1
A	8.34	3.85	.0275	.0010	12.2
B	.427	.215	.0043	0	.646
C	.0082	.0160	.0007	0	.0248
Total	92.1	7.75	.102	.0024	

Each cell contains the percentage of times an item was classified in the corresponding categories by each procedure

Table 7 NCDIF Power ANOVA

Source	Sum. Sq	df	F	*p*	η^2_G (LB)
SSR	0.5072	4	2160.7	0.0000	0.984 (0.980)
FGM	0.0069	6	19.7	0.0000	0.451 (0.341)
NDIF	10.8484	2	92432.4	0.0000	0.999 (0.999)
PES	2.5341	2	21591.9	0.0000	0.997 (0.996)
SSR:FGM	0.0027	24	1.9	0.0104	0.242 (0.034)
SSR:NDIF	0.9530	8	2029.9	0.0000	0.991 (0.989)
SSR:PES	0.2136	8	455.0	0.0000	0.962 (0.953)
FGM:NDIF	0.0137	12	19.4	0.0000	0.618 (0.526)
FGM:PES	0.0036	12	5.2	0.0000	0.302 (0.157)
NDIF:PES	5.4526	4	23229.2	0.0000	0.998 (0.998)
SSR:FGM:NDIF	0.0070	48	2.5	0.0000	0.453 (0.202)
SSR:NDIF:PES	0.4736	16	504.4	0.0000	0.982 (0.978)
FGM:NDIF:PES	0.0099	24	7.0	0.0000	0.539 (0.402)
Residuals	0.0085	144			

Sum. Sq.: Type II sums of squares; df: degrees of freedom; F: Value of the F ratio; p: p-value; η^2_G Generalized η^2 (eta-squared) effect size measure; LB: 95% confidence lower bound of the η^2

Table 8 MH Power ANOVA

Source	Sum. Sq	df	F	*p*	η^2_G (LB)
SSR	0.6862	4	352.8	0.0000	0.907 (0.886)
FGM	0.2072	6	71.0	0.0000	0.747 (0.690)
NDIF	9.0510	2	9307.7	0.0000	0.992 (0.991)
PES	2.3681	2	2435.3	0.0000	0.971 (0.965)
SSR:FGM	0.0075	24	0.6	0.8940	0.097 (0.000)
SSR:NDIF	0.9771	8	251.2	0.0000	0.933 (0.918)
SSR:PES	0.3690	8	94.9	0.0000	0.841 (0.803)
FGM:NDIF	0.1261	12	21.6	0.0000	0.643 (0.556)
FGM:PES	0.2672	12	45.8	0.0000	0.792 (0.742)
NDIF:PES	5.6301	4	2894.9	0.0000	0.988 (0.985)
SSR:FGM:NDIF	0.0712	48	3.1	0.0000	0.504 (0.284)
SSR:NDIF:PES	0.5072	16	65.2	0.0000	0.879 (0.849)
FGM:NDIF:PES	0.3785	24	32.4	0.0000	0.844 (0.803)
Residuals	0.0700	144			

Sum. Sq.: Type II sums of squares; df: degrees of freedom; F: Value of the F ratio; p: p-value; η^2_G Generalized η^2 (eta-squared) effect size measure; LB: 95% confidence lower bound of the η^2

size measure ($\eta^2_G = .982$ and $\eta^2_G = .879$, respectively). The contrasts for the two levels of sample size ratio (1500:1500 vs 1500:40000) on the power of either NCDIF and MH were also non-significant (NCDIF: $.025$, SE $= .237$, $t_8 = .106$, $p = .918$; MH: $.015$, SE $= .234$, $t_8 = .064$, $p = .95$).

With respect to the classification agreement of DIF items between NCDIF and MH, Table 9 presents the corresponding cross-classification table for all DIF items. This table must be read with some caution, however, because of the differences of classification of the true DIF parameters for each procedure, as shown on Tables 3 and B2. Hence, the direct comparison between the classifications should be limited to the incorrect non-detection of DIF items versus correct detections. On average, both statistics achieve similar rates of correct detections and similar rates of disagreement.

Table 9 Cross-classification of DIF items by NCDIF and MH

MH	NCDIF				Total
	No DIF	A	B	C	
No DIF	28.8	4.03	.556	.0727	33.5
A	4.05	9.84	2.08	1.06	17.0
B	.262	.933	2.16	7.24	10.6
C	.0192	.0841	2.60	36.2	38.9
Total	33.2	14.9	7.39	44.6	

Each cell contains the percentage of times an item was classified in the corresponding categories by each procedure

Discussion of simulation results

The parameter recovery results both for items and abilities indicate excellent recovery, which supports the quality of the simulations. Hence, these results further provide support for the correctness of the findings regarding the frequentist properties of the indeces. The good recovery of the ability averages indicates that the DIF detection procedure led to a successful equating of the two groups.

The IPR test shows a good control of the type I error rate when impact is not large. Previous studies (Oshima et al., 2006) had only considered a medium impact value (focal group shifted by $-.5$) and had limited replications within each condition, so it was not clear that this effect could be systematic. Another interesting result about the effect upon the Type I error of large impact across these studies is its apparent interaction with the ratio of the two sample sizes. The effect largely disappears in the conditions where that ratio is roughly 1:25 (1500 individuals in the focal group and 40,000 in the reference group).

The most important consequence of the simulation results of Type I Error is that rates clearly remain less well controlled in conditions with equal sample sizes. In particular, the lack of significant differences between the 1500:40000 and the 40000:40000 conditions suggest that there is no need to select sub-samples for the DIF analyses of SABER 11 data.

With respect to power, the simulations show that items with moderate or large DIF can be consistently identified under all conditions. They also show that items with negligible DIF, while still identified under all conditions by the NCDIF index and at least at good rates by the MH, can be correctly classified by using the effect size guidelines. The differences in power and Type I Error between the identification using the NCDIF versus the MH procedure suggest that items identified by both procedures and classified as having non-negligible DIF by the respective effects size measures can be confidently considered as showing DIF; on the other hand, items only identified in this manner using the MH are more likely to be incorrect detections than items identified only by the NCDIF index.

Given these results from the simulation studies, we can now make choices that were deferred in Section "What decisions to make when conducting DIF analysis." Firstly, it is clear that one should *include the effect size measure* in the detection process because we expect that there is moderate impact between the groups—whenever there was impact, there were some marked increases in the Type I Error rates, especially for the MH. This decision is strengthened by the results for the ad hoc conditions (10 DIF items) which show and even larger error inflation which is appropriately tamed by the inclusion of the effect size measure. Moreover, the size of the inflation of Type I Error rates (for both procedures) even after purification, justifies the decision of not considering as showing DIF items that were only detected by the statistical tests. That means that such items

will not be flagged and further examined due to DIF in the data analysis.[6] Otherwise, resources would be routed to an in-depth search for possible sources of bias for items that are very likely not functioning differentially, or worse, if automatic removal of items flagged for DIF were to be performed, the quality of the measure would likely be unnecessarily reduced because it was not truly compromised by potential bias. Whereas it was not questioned whether to use purification as part of the procedure, the results illustrate the importance of purifying the matching variable.

Lastly, regarding the second question, about which groups to use given the use of the effect size measure, we can choose to keep the original samples; there would be *no gain in subsampling the data* from AY1 to match the size of sample from AY2. This decision contrasts with the suggestion found in the literature of using samples of equal size for DIF analyses (Oshima & Morris, 2008; Wright & Oshima, 2015).

SABER 11 analysis

We used data from the Mathematics test from SABER 11 exams on two consecutive testing occasions: (a) the second test date of 2018 (AY1) and (b) the first date of 2019 (AY2). Several test forms are utilized in the assessment of students in each examination. We illustrated the procedure with one test form from each of the two dates. Table 10 presents the item parameters for the two forms estimated separately from their

Table 10 AY1 and AY2 Common Item parameters SABER 11

Item	Rescaled*AY1		AY1		AY2		NCDIF	ES	Δ_{MH}	ES
	a	b	a	b	a	b				
1	.65	− .65	1.00	− 1.02	.78	− 1.08	.00109	A	− .03	A
2	.29	3.97	.45	1.99	.40	1.94	.00064	A	.16	A
3	.60	− 1.10	.92	− 1.31	.87	− 1.27	.00026	A	− .11	A
4	.50	.58	.77	− .22	.67	− .27	.00077	A	.25	A
5	.79	.10	1.22	− .53	1.04	− .61	.00127	A	.59	A
6	.83	− .47	1.27	− .90	1.16	− .91	.00017	A	.09	A
7	.97	− .49	1.50	− .91	1.21	− 1.00	.00121	A	.36	A
8	.27	2.32	.42	.92	.47	.86	.00024	A	− .30	A
9	1.01	− .80	1.56	− 1.11	1.41	− 1.15	.00016	A	.20	A
10	.52	.22	.80	− .45	.94	− .34	.00154	A	− .25	A
11	.42	−2.12	.65	− 1.97	.61	−2.09	.00006	A	.13	A
12	.80	− 1.47	1.23	− 1.55	1.05	− 1.58	.00031	A	− .20	A
13	.15	3.73	.23	1.83	.42	.91	.00515	A	− .22	A
14	.42	− 1.52	.64	− 1.58	.55	− 1.68	.00025	A	− .01	A
15	.55	.18	.84	− .48	.92	− .43	.00034	A	− .12	A
16	.82	.05	1.26	− .56	1.33	− .55	.00009	A	.20	A
17	.38	.69	.59	− .14	.76	− .21	.00192	A	.10	A
18	.25	4.01	.39	2.01	.76	.63	.03096	C	1.64	C
19	.52	.58	.80	− .21	.85	− .07	.00152	A	− .43	A
20	.76	− 1.67	1.18	− 1.68	.90	− 1.78	.00063	A	− .32	A
21	.44	.60	.68	− .20	.84	− .23	.00137	A	.16	A
22	1.03	− .75	1.58	− 1.08	1.33	− 1.09	.00050	A	− .05	A

The columns with the rescaled item parameters of the test form used in AY1 are transformed to the scale of the AY2 test form. The scaling constants are $s^* = .65$ and $m^* = -.59$ for the linear linking function where $a_k^* = s^* a_k$ and $b_k^* = s^* b_k + m^*$

[6] For completeness, the complementary decision would have been to flag and further examine the content, and perhaps additional studies, of items detected based on the statistical tests with perhaps some priority given by the effect size classification.

corresponding dates before scale linking and the rescaled item parameters from the test form fom AY1.

Based on the results of the simulation studies, we will look for DIF between items of these two test forms using the whole samples for the chosen test forms. That is 1508 examinees from AY2 and 39377 from AY1, which are, respectively, the focal and reference groups for this analysis. After purification, the IPR test for NCDIF with $\alpha = .01$ (or $\alpha = .05$) was significative for eight (ten, respectively) items, out of the 22 common items. For the MH procedure, the statistical test was significative for three (four, respectively) items. However, only one item was also classified as having non-negligible differences when considering the effect size. All other items were classified as having negligible DIF. Hence, for this test, item number 18 in Table 10 is the only item identified with DIF. Figure 7 shows the ICCs of this item for both AYs, and its NCDIF value is $.03096$ with a Δ_{MH} of 1.6386. After removing item 18 from the common set of items, the scaling constants, estimated using the Stocking-Lord procedure, are $s^* = .65$ and $m^* = -.59$. That is, the transformation $s^*\theta + m^*$ will take an ability estimate from the AY1 scale into the AY2 scale[7]. Once this item is removed from the common items, the difference in means between AY1 and AY2 is $.59$, favoring AY2.

Identified item

Item construction for SABER 11 follows the Evidence-Centered Design approach (Mislevy et al., 2003; Mislevy, 2006). The DIF identified item is presented below translated into English. This item was written to assess *'planning and plan execution'* to find solutions to quantitative problems requiring *'algebra or calculus'*. These correspond to the

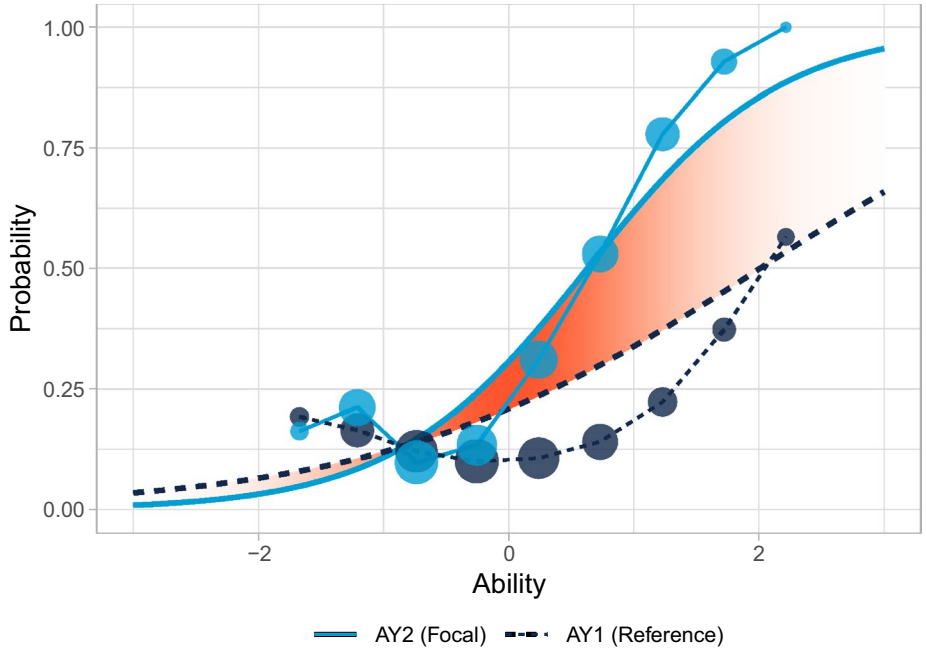

Fig. 7 Characteristic curves for Item 18. *Note.* Item parameters transformed to the scale where AY2 abilities are standardized (mean zero and unit variance). The dots in the figure for each AY are proportional to the relative frequency of examinees within the corresponding bin

[7] Equivalently, the transformation $s^*\theta + m^*$, with $s^* = 1.54$ and $m^* = .91$, will take an ability estimate from the AY2 scale into the AY1 scale.

specific *skill* and *subject area* that the item was designed to assess. The *claim* is that a student will *'plan and implement strategies that lead to adequate solutions to a problem involving quantitative information'*, which is *warranted* by the *'execution of such plan to solve the problem'*.

Translated Item

A Mathematics teacher asks their students to solve the following equation:

$$(x+2)(x+3) = 5(x+3)$$

María, Nelson, and Óscar show the work below, respectively:

María	**Nelson**	**Óscar**
$x^2 + 5x + 6 = 5x + 15$ $x^2 + 5x - 5x = 15 - 6$ $x^2 = 9$	$(x + 3)[(x + 2) - 5] = 0$ $(x + 3)(x - 3) = 0$	$2x + 5 = 5x + 3$ $2x - 5x = 3 - 5$ $-3x = -2$ $3x = 2$

Which student(s) have performed a correct procedure to solve the equation?

A	Only María and Óscar.	C	Only Óscar.
B	Only María and Nelson.	D	Only Nelson.

Item analysis

The differences in the item characteristic curves for the two academic years show that for examinees from AY2 the probability of correctly answering item 18 is larger than that for examinees from AY1. Moreover, that is the case for almost all examinees of AY2, as highlighted by the shaded region in Fig. 7. This difference is relatively small near the average ability for AY2 examinees[8] but it rapidly increases together with the ability magnitude. For an ability value of 1.0, the probability of a correct response is close to $.62$ when the examinee belongs to AY2 but is below $.35$ when they belong to AY1.

The item was reviewed in light of this behavior, as well as the other item analysis statistics,[9] by the Mathematics test team at Icfes. During this revision, it was pointed out that neither of the three procedures, but most particularly Nelson's procedure, reach a final answer. Although two of the procedures are correct and lead towards the correct answer to the problem, the lack of the last steps to reach a solution could be argued to lead to confusion for some examinees. Particularly, it was hypothesized that, due to this incompleteness, the step at which Nelson's procedure ended could appear as incorrect to examinees from AY1 more frequently than to examinees from AY2. Figure 8 shows that the option for which Nelson's procedure is rejected (option A) remains an attractive response for examinees from AY1 for a longer range of abilities. An analogous behavior is observed with respect to the option in which both María's and Nelson's procedures are rejected (option C). Based on these issues, this item has now been modified making the three procedures much more explicit and reaching a final answer for the value(s) of x.

[8] The two ICCs cross at $\theta = -.824$ and the mean for AY1 on the scale used in the figure is $-.59$.

[9] Such as the fit or lack thereof between the model predictions and the observed proportions.

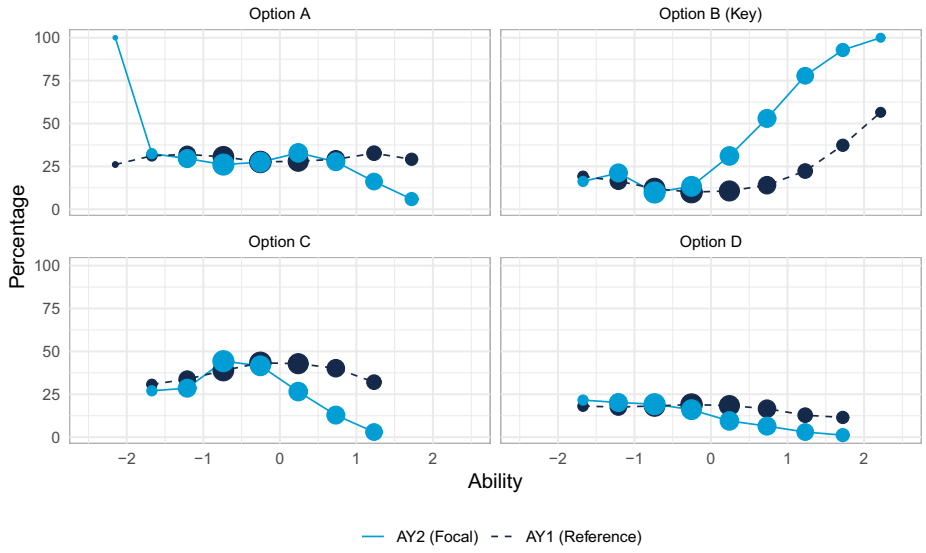

Fig. 8 Empirical option characteristic curves for Item 18. *Note*. Ability scales transformed to the scale where AY2 abilities are standardized (mean zero and unit variance). The ordinates for each dot indicate the percentage of examinees whose ability estimate lies within the bin centered at the dot's abscissa who select the option. The dots in the figure for each AY are proportional to the relative frequency of examinees within the corresponding bin

Summary and discussion

In this paper, we present DIF analyses of a test in SABER 11, a Colombian LSA used for college admissions and as an indicator for the quality of secondary education in the country, contrasting test forms from two different AYs. Through this case study, we have illustrated how to use and answer the guiding questions proposed by Sireci and Rios (2013) to help perform such analyses in light of the particular testing conditions of SABER 11. Additionally, in the process of making the relevant choices outlined by Sireci and Rios (2013), we identified a gap in the literature regarding the frequentist properties of the NCDIF index under the large sample sizes and group size ratios that can be found in LSAs such as SABER 11. To resolve this gap and be able to make the corresponding decisions for analyzing the tests in this LSA, we performed and reported in this manuscript a set of simulations. The results from these simulations aimed at providing information on the frequentist properties of the statistical test (Type I and Type II Error rates) based on the IPR procedure for the NCDIF statistic.

The simulations allowed us to make decisions with respect to the two questions in the analysis for which there was some uncertainty on the best approach: (a) What sample sizes are more appropriate to use to analyze SABER 11? and (b) How to incorporate the effect size measure in the identification process? In particular, for SABER 11, and likely for other similar LSAs, where impact between the two groups is expected to be at least moderate and the ratio between the groups sample size may be large, *one should make sure to use the effect size classification as part of the decision to classify an item as presenting DIF or not*, and it may be preferable to use the original full sample sizes despite the large difference between the group sample sizes. Both of these factors evidenced increased inflated Type I Errors and these decisions radically help control the Type I error rates—rendering them almost nil for most conditions and both indexes while not doing so allowed Type I Error to increase to unacceptable levels—and do not harm the identification power in the conditions of this LSA. One should note that the inclusion

of the effect size in the decision process is not a reiteration of the usual recommendation of considering an effect size measure to qualify the practical importance of a significant result due to large samples (as it is also the case for SABER 11 and our simulations) because that recommendation deals with small true non-null effects; in our results, we find that the use of the effect size measure is useful to tame false significant results even for true null cases. Since the costs of incorrectly flagging for DIF are particularly important in LSAs, such as increasing costs associated with the reduced precision of individual ability estimates, the additional resources for investigating possible sources of bias, and overfitting the item construction process to deal with spurious DIF sources, we recommend to include the effect size measure as part of the decision to prevent excessive chances of incorrect detections and their related costs when true differences between the groups are expected and samples are sufficiently large, such as it is typical in LSAs like SABER 11.

The choice of samples for the analysis had the greatest uncertainty prior to the simulation studies because previous studies examining the performance of NCDIF explored conditions very different to the LSA we consider. The simulation results indicated, however, that *it was not necessary to have equal samples in our context*. Hence, we decided not to create equal sample sizes by (possibly repeated) subsampling of the larger group.

Regarding the results of the DIF analyses for the two forms of the Mathematics test of SABER 11 reported, we expected appropriate rates for both Type I Error and power provided the outcomes from the simulation (see Fig. 3c, central panels for 1500:40000 samples and ratio, and moderate [$\sim$.5] impact favoring the focal group). Moreover, in light of the results of the ad hoc simulation conditions (see the corresponding panel of the '10 DIF items' conditions), and given the results and DIF classifications of each item, we can confidently assert that the SABER 11 tests do not have important DIF across the two AYs and that the identified item is very likely the only item in those test forms truly exhibiting DIF.

Limitations and future directions

Before purification, the increment in Type I Error was not symmetric with respect to the change of the mean of the focal group. Some of this asymmetry persists after purification, or even appears sharper for the MH. It may be noted that, although the partition between item parameters into common and not common for Forms A and B was random, the resulting average difficulty of the common items was shifted (the difference in average difficulty between non-common and common items was .66). There may be an interaction of detection with the distribution of common item parameters. From Fig. 3, one may intuit that the relationship is also affected by the number of DIF items, and probably their DIF magnitude. However, because this factor was not systematically controlled in our study, our results are not sufficient to understand the relation between the distribution of item parameters in the test and the rates of incorrect detections. Future studies could pursue a comprehensive study of these relationships. The main conclusion from this effect is clear and highlights the recommendations by Hambleton (2006), among others: when computing the matching variable for DIF analysis, it is crucial to purify it from the noise introduced by DIF items.

Additionally, we would like to highlight that identifying the items that present DIF between two or more groups is but the first step in the quest to ensure fairness in testing.

The content and item statistics analysis revealed a possible explanation of DIF for the identified item. Although this explanation seems reasonable, additional evidence for it ought to be gathered from other sources and methods. For instance, Thinking Aloud Protocols (Ercikan et al., 2010) could be used to understand how students from each AY interpret the question and each option, and teacher interviews and curricular analyses could be employed to understand how the material and instructional practices are effectively presented and implemented across these contexts.

Conclusion

In summary, the results from the present study provide evidence for the validity and fairness of the SABER 11 examination, particularly in its Mathematics subject test, for comparing students examined in the two AYs, who come from very different subpopulations of examinees. In all, this manuscript serves as a case study on how to use the guiding questions proposed by Sireci and Rios (2013) and how to tailor the process for identifying DIF to one's specific context. Lastly, the simulation studies both provide information on the performance of the NCDIF index with very large samples and sample size ratios, and suggest that simulation studies examining the performance of NCDIF, and possibly any DIF statistic, should implement realistic item parameter pools and not only sanitized well distributed sets of item parameters.

Abbreviations

2PL	2-Parameter logistic model
AY1	Academic year 1
AY2	Academic year 2
CDIF	Compensatory differential item functioning
DIF	Differential item functioning
DTF	Differential test functioning
ES	Effect size
ETS	Educational testing service
FGM	Focal group mean ability
ICC	Item characteristic curve
Icfes	Instituto Colombiano para la Evaluación de la Educación
IPR	Item parameter replication
IRT	Item response theory
LSA	Large scale assessment
MH	Mantel–Haenszel
NCDIF	Non-compensatory differential item
NDIF	Number of DIF items
PES	Purification and use of effect size measure
SSR	Sample size and sample size ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-026-00294-x.

Supplementary Material 1.

Acknowledgements

The authors thank Icfes for providing access to SABER 11 data and test production team. This project was supported by the Statistics Subdirectorate team at Icfes who, in pursuit of their institutional mission, develop research projects on psychometric and statistical methods. The authors would also like to thank Andrés Ricardo Rodríguez, Jenny Paola Martínez, and Sandra Camargo, as well as to the anonymous reviewers, for useful discussions and feedback on previous drafts of the manuscript.

Author contributions

JAC and VHC conceived the studies. JAC, NAR, and VHC planned the simulations and data analyses. JAC performed the simulation studies and produced the associated tables and figures. NAR cleaned the data from the two SABER 11 test forms, performed the data analysis, collaborated with the Mathematics test team at Icfes for the item content analysis, and produced the tables and figures for the case study. VHC supervised the project and took the lead in writing the manuscript. All authors discussed the results and commented on the manuscript.

Funding
The authors declare that no external funding supported this research.

Data availability
The data that support the findings of this study are available from Icfes but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available upon reasonable request and with permission of Icfes by contacting oficinadeinvestigaciones@icfes.gov.co.

Declarations

Ethics approval and consent to participate
Only secondary and simulated data were used in this study. The use of Icfes' secondary databases was authorized by Icfes' Research Office.

Competing interests
The authors declare that they have no conflict of interest.

Received: 22 July 2024 / Accepted: 11 March 2026
Published online: 12 April 2026

References

Atar, B. (2007). Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures. Doctoral dissertation. Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_etd-0248

Cervantes, V. H. (2017). DFIT: An R package for Raju's differential functioning of items and tests framework. *Journal of Statistical Software, 76*(5), 1–24. https://doi.org/10.18637/jss.v076.i05

Cohen, A. S., & Kim, S.-H. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39–52.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques et Sciences Humaines, 199*, 45–59.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel–Haenszel and standardization (Vol. 1992; Tech. Rep. No. 1). https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24–35. https://doi.org/10.1111/j.1745-3992.2010.00173.x

Fidalgo, A. M. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel–Haenszel procedures. *Methods of Psychological Research, 5*(3), 43–53.

Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In: M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), Methods and procedures in PIRLS 2016 (pp. 12.1–12.38). TIMSS & PIRLS International Study Center.

Gómez Benito, J., Sireci, S., & Padilla, J. L. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104–117.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144–149.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(Suppl 3), S182–S188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

Herrera, A.-N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739–755. https://doi.org/10.1007/s11135-006-9065-z

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential Item Functioning detection and effect size: A comparison between Logistic Regression and Mantel–Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903–915. https://doi.org/10.1177/0013164403261769

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. https://doi.org/10.4324/9780203056905-19

Huggins, A. C. (2014). The effect of Differential Item Functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement, 74*(4), 627–658. https://doi.org/10.1177/0013164413506222

Instituto Colombiano para la Evaluación de la Educación. (2020). *Informe nacional de resultados para colombia—pisa 2018* (1st ed.). Author.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates using an effect size measure with the Logistic Regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Khalid, M. N., & Glas, C. A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement, 50*, 186–197. https://doi.org/10.1016/j.measurement.2013.12.019

Khoeruroh, U., & Retnawati, H. (2020). Comparison sensitivity of the differential item function (DIF) detection method. *Journal of Physics: Conference Series, 1511*(1), Article 012042. https://doi.org/10.1088/1742-6596/1511/1/012042

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357–381.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics, 32*(4), 371–397.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

Kim, W., & Nering, M. (2007). *Evaluation of equating items using DFIT*. In Annual meeting of the National Council on Measurement in Education, Chicago, IL.
Kolen, M. J., & Brennan, R. L. (2014). Item response theory methods. In *Test equating, scaling, and linking* (pp. 171–245). Springer. https://doi.org/10.1007/978-1-4939-0317-7_6
Linn, R. L., & Drasgow, F. (1987). Implications of the golden rule settlement for test construction. *Educational Measurement: Issues and Practice, 6*(2), 13–17. https://doi.org/10.1111/j.1745-3992.1987.tb00405.x
Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862.
Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE–Life Sciences Education, 16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.
Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x
Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*(2), 131. https://doi.org/10.2307/1165166
Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*(4), 357–373.
Oshima, T. C., & Morris, S. B. (2008). Raju's Differential Functioning of Items and Tests (DFIT). *Educational Measurement: Issues and Practice, 27*(3), 43–50. https://doi.org/10.1111/j.1745-3992.2008.00127.x
Oshima, T. C., Raju, N. S., & Nanda, A. O. (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.
Oshima, T. C., Wright, K., & White, N. (2015). Multiple-group noncompensatory differential item functioning in Raju's differential functioning of items and tests. *International Journal of Testing, 15*(3), 254–273. https://doi.org/10.1080/15305058.2015.1009980
Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502. https://doi.org/10.1007/BF02294403
Raju, 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/10.1177/014662169501900405
Seybert, J., & Stark, S. (2012). Iterative linking with the Differential Functioning of Items and Tests (DFIT) method: Comparison of testwide and Item Parameter Replication (IPR) critical values. *Applied Psychological Measurement, 36*(6), 494–515. https://doi.org/10.1177/0146621612445182
Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting Differential Item Functioning. *Educational Research and Evaluation, 19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621
Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208
Teresi, J. A., & Jones, R. N., et al. (2013). Bias in psychological assessment and other measures. Test theory and testing and assessment in industrial and organizational psychologyIn K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 139–164). American Psychological Association.
TIMSS & PIRLS International Study Center. (2017). PIRLS 2016 achievement scaling methodology. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 11.1–11.9). Author.
von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 Technical Report* (pp. 11.1-11.25). TIMSS & PIRLS International Study Center.
Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's Differential Functioning for Items and Tests. *Educational and Psychological Measurement, 75*(2), 338–358. https://doi.org/10.1177/0013164414532944
Wu, L.-A., & Tsai, R.-C. (2010). A comparison of three polytomous DIF detection methods. *Annals of Measurement Statistics, 18*(2), 1–22.

Publisher's Note