Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Statistiche

Ciclo XXXI

Settore Concorsuale: 13/D1

Settore Scientifico Disciplinare: SECS-S/01

DETECTION OF DIFFERENTIAL ITEM FUNCTIONING IN IMBALANCED GROUPS. ARE INVALSI TESTS FAIR AMONG PUPILS FROM DIFFERENT ACADEMIC SCHOOLS?

Presentata da: Nicola Bazoli

Coordinatore Dottorato **Supervisore**

Prof.ssa Alessandra Luati **Prof.ssa Stefania Mignani**

Esame finale anno 2019

Abstract

Differential Item functioning (DIF) and bias measurement are often used as synonyms in standardized tests fairness evaluation between individuals belonging to different groups. Recently, Zumbo et al. (2016, 2017) have provided a redefinition of DIF/bias term and proposed a new methodology for DIF/bias detection analysis. The new definition of bias requires attributional reasoning; therefore, there is a need to find a way to control for possible confounding factors. Only by balancing groups with respect to covariates, it is possible to attribute DIF to group membership. Propensity score matching techniques allow to carry out groups balancing and bias is detected if item is flagged as DIF, after balancing groups. The conditional logistic regression is proposed for DIF detection analysis after matching because it allows to consider the data structure generated by matching.

The aim of this work is twofold. Firstly, we assess the efficacy and performance of the new methodology in imbalanced groups, comparing its performance to performance of traditional DIF detection methods (Mantel-Haenszel statistic, logistic regression and Lord's χ^2). Our research, through a simulation study, shows that the new methodology outperforms traditional DIF detection methods in imbalanced groups in situations of large sample and DIF items presence. Nevertheless, the new methodology suffers to I error inflation for large sample and simulation results suggest that the use of an effect size measure (ΔR^2) reduces significantly this issue. Secondly, the proposal methodology is applied to data coming from the large-scale standardized test administered by the National Evaluation Institute for the School System (INVALSI) to evaluate pupils' Italian language and mathematics competencies. The idea is to detect possible DIF items among pupils from different academic tracks. The results reveal that very few items are flagged as DIF, indicating the fairness of INVALSI tests.

Acknowledgments

First and foremost I want to thank my supervisor Stefania Mignani. I appreciate all her contributions of time, ideas and support to make my Ph.D. experience productive and stimulating. Secondly, I'd like to give special thanks to my Ph.D thesis reviewers, Prof. Guido Pellegrini and Dr. Roberto Ricci. Their useful suggestions and contributions have helped me improve my thesis, especially, in completing literature review on propensity score matching and in developing more adequate conclusions, reflecting on policy implications of the results. Thirdly, I want to thank IRVAPP (Research Institute for the Evaluation of Public Policies) which hosted me in the last Ph.D. year. Here, I understood what means to make research and the nearness of fellow researchers has teach me team working and the social implication of the research. Finally, I'd like to give special thanks to my parents and Serena for their continuous nearness and support in these doctoral years. Probably, without their support, my doctorate path would not have ended.

Table of Contents

List of Tables

List of Figures

1. Introduction

Educational standardized tests are useful tools for observing and measuring students' abilities and competences. Test users administer the same test, composed by several questions, to all pupils of a class or school in order to measure their competence in specific subject. Ability is usually estimated through a statistical approach: the latent trait analysis. This statistical approach allows to measure a latent variable (ability), assumed to be continuous, observing categorical variables (question responses). Test users assume that the test is comparable among different groups, but this is not always correct. It is possible that the test, or a part of the test, advantages some subgroups rather than others, therefore it results biased and unfair. Indeed, if standardized test presents this kind of issue than it does not measure pupils' ability in the same way for all subgroups.

When it comes to test inequity or bias among pupils allocated into different groups, psychometric literature refers to an item characteristic: the differential item functioning (DIF). DIF provides useful information about unfair items between groups. Indeed, DIF occurs when individuals with the same latent trait level but allocated into different groups present different probability of success to the item. Literature provides several DIF detection techniques, both parametric and non parametric. Recently, a new methodology has been proposed in the psychometric literature for detecting possible biased standardized test items. The new method is developed on a redefinition of bias concept in DIF detection analysis. From this redefinition, the new methodology allows to attribute DIF to group allocation, controlling for confounding variables. The new methodology applies matching techniques to DIF detection analysis and "*... the purpose of matching on covariates is to eliminate pre-test group differences to purify the sources of DIF and make a causal claim about DIF*" (Liu et al., 2016, p. 17).

The aim of this thesis is twofold. First of all, we want to assess the accuracy and performance of the new methodology in situations in which groups are imbalanced with respect to covariates. We use a simulation study in order to reach this goal. The assessment of effectiveness and accuracy is based on false alarm rate (type I error) and power (1 minus type II error), comparing traditional DIF detection methods and the new methodology.

Secondly, we want to assess if INVALSI tests are unfair among pupils from different (academic) school tracks. We apply the new methodology to maths and Italian language INVALSI tests 2016/2017 because it helps to reduce selection bias and attribute possible bias to group allocation. This kind of application allows to evaluate the test quality and fairness. In other words, we assess if INVALSI tests advantage some academic schools rather than others. If we find that some academic tracks present advantages in tests, the administration of the same test to different schools should be avoided. In addition, it is possible that test unfairness associated to different schools could be attributed to curriculum or teaching proposed by the schools. Therefore, if we find situations in which item content or format disadvantages some schools, it could be useful revise their curriculum and improve the teaching of specific content or accustom pupils to particular exercise format.

The new methodology was presented in the literature from applicative (Liu et al., 2016) and theoretical (Wu et al., 2017) point of view. Our contribution is to assess the efficacy and accuracy of this new methodological proposal. In other words, we assess how it performs in different situations, for example different test length, number of test takers and percentage of DIF items. If the methodology presents good performance, or better than traditional DIF methods, than it becomes an useful psychometric instrument in order to detect standardized test efficacy and fairness. Indeed, the attributional claim linked to the new methodology could help test users understand if a standardized test is fair with respect to different groups and assess the instrument validity. In addition, we apply the new methodology to a real dataset: INVALSI tests. The attributional claim allows us to evaluate the test equity, comparing pupils from different academic schools. This kind of analysis is useful for experts because it controls if is fair to administer the same test to different academic tracks. If so, test users should understand why this happens and they should modify and improve the test in order to make it fair.

The thesis structure is the following. The chapter 2 discusses the literature review. Firstly, it presents sections dedicated to test fairness and school tracking with particular focus on the school choice determinants and the educational Italian system. Subsequently, we introduce the concept of differential item functioning, its recent developments in the literature and the traditional methods for DIF detection analysis; finally, the new

methodology is presented. The chapter 3 illustrates data and methods used for our research objectives. Section about data describes INVALSI data, dependent and independent variables used for the simulation study. The simulation design involves covariates, latent trait, grouping variable and responses variables generation and the manipulated factors considered for constructing scenarios. Section about methods describes how we assess the performance of DIF detection techniques and details of the new methodology. Simulation results are presented in chapter 4. After checking if scenarios are consistent to real data, we provide propensity score matching analysis across different scenarios. Results about DIF detection methods performance involve analysis on type I error inflation and test power, with a deepening about the new methodology. We apply the new methodology to INVALSI data 2016/2017 in chapter 5. In the end, chapter 6 discusses conclusions about the main thesis results and its limits, useful for possible future developments.

2. Literature Review

The first chapter of the thesis presents the literature review. First of all, section 2.1 introduces standardized tests and fairness of measurement instrument (measurement invariance). Secondly, section 2.2 points out applicative context and motivations of research. Section 2.3 and 2.4 provide statistical methods and tests for assessing measurement invariance between two groups, with particular stress to a new methodological approach recently proposed in psychometric literature (section 2.4).

2.1 Test fairness

Standardized tests are useful tools for observing and measuring complex phenomena or constructs that are not directly measurable. For example, standardized tests can measure individuals' QI, well–being, anxiety or socio–economic status. You can think of studying the anxiety that individuals felt during an exam. It is not possible to directly ask how anxious people were while they were taking the exam. Anxiety is not directly observable, therefore we must find an instrument in order to take over the anxiety, for example a standardized test. Formally, tests measure latent constructs through something that is observable. Standardized tests have several advantages. The first advantage is the equanimity, namely, standardized tests are impartial in the judgment. In addition, standardized tests are easy to correct and compare among results and they should be independent among test takers.

The most famous standardized tests are administered in educational fields. In educational contexts, standardized tests are developed for different reasons as, for instance, mechanisms of school admission or educational system assessment. Most of the educational tests aim to measure pupils' ability and competence. Since these phenomena are not directly observable there is the need to find something observable. Tests administered to pupils present several questions or exercises (items) that refer to the ability and capacity which the tests want to measure. Therefore, these test items become observable variables. These variables can be dichotomous, which assume only two possible values, or polytomous (for multiple responses), which can assume more than two possible values. This work treats only the first case. The two possible values refer, respectively, to a correct response (usually

coded by 1) and to a wrong response (coded by 0).

Nowadays, different popular standardized tests exist in educational context. PISA tests (Programme for International Student Assessment), promoted by OECD, are the most popular. PISA tests aim to measure teenagers' learning level in maths, science and reading, in international comparison. Italy has developed *ad hoc* test for the Italian students, namely INVALSI (*Istituto Nazionale per la VALutazione del SIstema educativo di istruzione e di formazione–National Evaluation Institute for the School System*) tests. INVALSI tests are administered all years to pupils of four different levels of the educational system[1] and they aim to measure mathematical and Italian language abilities[2]. At the upper secondary schools, INVALSI administers the same tests to pupils from different schools. Therefore, INVALSI assumes that tests are unfair among the different school tracking. The starting point of this thesis is to find tools and methodology for answering to questions as: Do INVALSI tests measure the same ability among pupils from different school tracking? Is this instrument unfair or biased with respect to school tracking?

Standardized tests are very important in today's society. Although standardized tests are not perfect evaluation instruments, they provide useful information that other evaluation tools do not provide (Richard, 2008). The main feature of standardized tests is objectivity. The objectivity allows to make decisions consistent with respect to reality, without possible influence of personal opinions or subjective preconceptions. You can image a situation in which a college admits only the best students. The college chooses to select students through teachers judgment. It is possible that some worthy student will not attend the college because of teachers' bad personal opinions. This would not happen if students were selected by a standardized tests. Consequently, for standardized tests objectivity, policy makers often use these kinds of instruments in order to improve some public policies. These instruments might also support policy makers in addressing their interventions.

Since often standardized tests guide policy makers decisions, it is necessary that tests are fair among individuals and groups. In other words, if a test systematically advantages only some individual or group, a policy decision based on test results will not be efficient. Fairness is a complex concept and it has social rather than psychometric nature. Fairness can have

[1] II and V grade of primary school, III level of secondary school of I grade and II level of secondary school of II grade.

[2] Starting from 2018, English abilities are tested in INVALSI tests for students of V level of primary schools and III level of secondary schools of I grade.

different meanings and *Standards for Educational and Psychological Testing*[3] points out four possible fairness meanings. The first meaning refers on equity of group outcomes. Nevertheless, test literature agrees that differential outcomes do not unequivocally reflect fairness. For the second meaning, all test takers must have the same treatment. In other words, test takers equally have to enable to perform the test: same access to material, same environment, same test administration, etc. The third meaning concerns comparable opportunity to learn the subject matter covered by the test. The last meaning refers on predictive bias. Predictive bias involves a statistical approach. Multiple regression models are run, where the interested measure is regressed on the predictor score, grouping variable and an interaction term. Fairness is not present if subgroup does not differ in regression slopes or intercepts.

Despite the controversial nature of fairness concept, traditionally, in psychometric literature, when we talk about test fairness, we refer to some desirable properties that a measurement instrument should have. The main property is the measurement invariance. It occurs when a measure is *"...independent of the characteristics of the person being measured, apart from those characteristics that are the intended focus of the measure."* (Millsap, 2007). Measurement invariance is a statistical property of a measure and it guarantees that the instrument measures the same latent trait between individuals from different groups. A violation of the measurement invariance may prejudice the reliability of instruments used to measure latent trait.

As previously said, educational standardized tests are composed of different items. Therefore, it becomes focal to analyze item characteristics in order to evaluate test fairness and measurement invariance. It is possible to distinguish between item impact and bias (Dorans and Holland, 1992), when referring to test fairness assessment between individuals from different groups. Item impact refers to a situation in which pupils from different groups (e.g. gender or ethnic groups) present different probability to correctly answer to an item. For example, males outperform females, in average, in standardized maths total scores (INVALSI, 2016; OECD, 2015). Item impact may reflect the true existing difference between the groups. In other words, the differential occurs because individuals from different groups have different ability levels. Usually, test users assume test scores'

[3]Standards is a set of testing standards jointly developed by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). Its last version was released in 2014.

equivalence across various subgroups and they compare groups only through test scores. This assumption is not always correct.

Item impact may also occur because item could be biased. Bias refers to a situation in which individuals from different groups exhibit different probability to correctly answer to an item due to some item characteristics that are not relevant to the construct being measured. Item bias may occur because of the different meaning that individuals from different groups might give to items, or because item measures a second latent trait, as cultural or curricular latent trait (Martinková et al., 2017). This concept has a qualitative nature: contextualization and reconstruction of meaning are required. Differently, Differential Item Functioning (DIF) has a statistical nature. *"Differential item functioning refers to a psychometric difference in how an item functions for two groups"* (Osterlind and Everson, 2009). DIF can be considered as measurement invariance property applied to test items. DIF is a necessary, but not sufficient condition for bias. Traditionally, educational experts adopt a twofold analysis in test fairness assessment among groups. Firstly, they compute DIF detection analysis. If an item presents DIF, then it is subject to a qualitative analysis by a multidisciplinary equipe (sociologists, statisticians, educational experts, psychologists).

2.2 School tracking

All education systems group pupils together on age and grade. When this process happens separating pupils based on ability level, this is called formal differentiation or tracking. Despite this formal meaning, tracking can be adopted as a free students decision (it is the case of Italy). Education systems characterized by tracking present different tracks and schools in which different curricula and teaching contents are proposed to students. Nowadays, in western education systems tracking does not exist at primary school. Nevertheless, some systems group pupils into different school at (post–)secondary and tertiary school (OECD, 2010). For example, German and Austrian tracking happens at the age of ten, while in other countries tracking can happens later, at age of fourteen (Italy) or at age of sixteen (United States). Differently, systems without tracking, namely comprehensive schools, do not sort pupils in different school tracks. These schools propose an homogeneous curriculum for all students. British secondary school and public high school in the United

States and Canada are example of comprehensive schools.

2.2.1 The Italian system

Italy is a country in which tracking characterizes compulsory education, at secondary school. The first eight educational years occur at comprehensive schools. The fist level of education is characterized by an unique curriculum and teaching contents are equal for all children. The primary school (*scuola elementare*) includes children from at the age of six to ten[4]. The secondary school is divided in two levels. The first level (*scuola unica media*) lasts three years. First level secondary schools propose similar curricula for all pupils and only some teaching contents can differ[5]. At the end of the third year, student must take a state exam that allows the student to enter to second level of secondary schools.

At the age of fourteen Italian pupils must choose a track of second level of secondary school: academic, technical or vocational schools. Children and their parents are free to choose the type of schools. Academic schools (*licei*) provide academic and general curricula. These schools aim to prepare pupils for the next educational level, namely the tertiary school. Academic schools differ from each other in basis on the main school subjects: scientific, classical, social science contents, and so on. Differently, technical schools (*istituti tecnici*) aim to prepare students for labor market, especially for technical and economic positions. Finally, vocational schools (*istituti professionali*) transfer to pupils vocational skills oriented to two main sectors: industry and handicraft and services. All these tracks end with the final state exam, unique access channel for tertiary education. In addition to these three branches, Italian pupils can opt for vocational training courses (*formazione professionale*). These courses last from two years to five years and they exclusively form students for labor market. In addition, vocational training courses aim to finish compulsory education[6], but only state exam guarantees the access to tertiary education (Azzolini and Vergolini, 2014).

[4]Children born in the first four months of the year can anticipate the first year of primary school at the age of five, by school agreement.

[5]For example, some schools teach two foreign languages, while others teach only English.

[6]Compulsory education ends with state exam or professional qualification obtained within eighteen years old.

2.2.2 School tracking and educational performance

Literature shows how tracking may affect pupils' school performances. This can happen for a twofold reason (Gamoran, 1992). First of all, tracking may increase educational inequality, or the dispersion of achievement. For example, higher–status track (academic school) allows pupils to learn more than lower tracks: tracking produces gap among students' performance from different school branches. Researchers have shown that tracking increases social inequalities in education (Azzolini and Vergolini, 2014; Checchi and Flabbi, 2013; Hanushek, 2006): educational systems with tracking present more inequalities than those without tracking, which tend to be more fair. Secondly, "*the particular structure of tracking may influence a school's overall level of achievement, or educational productivity*"(Gamoran, 1992, pp. 812-813).

Previous studies about school tracking and pupils' performance have shown a systematically gap between different tracks (M. Becker et al., 2012; Opdenakker and Damme, 2006). Students form academic schools get better mean scores in standardized tests than students from other tracks, especially than students enrolled in vocationally-oriented tracks. This gap involves both language and mathematics achievement, but it is more pronounced for the second one.

Italian context also presents a gap in Italian language and mathematics competencies among pupils from different tracks of upper secondary school. Academic schools outperform technical schools in Italian language test. In turn, technical schools get higher average scores than vocational schools, both at national and macro–area level. Similar results emerge from mathematics test. At national level, academic schools outperform technical schools that, in turn, outperform vocational schools. At macro–area level, differently from Italian language test, north academic schools present significantly higher scores with respect to national average scores, while south academic schools present significantly lower scores. Similar results emerge for technical and vocational schools, but south schools do not deviate from national average scores (INVALSI, 2017c; INVALSI, 2016). To sum up, different Italian school tracks entail different competencies and abilities. It exists a gap in abilities among upper secondary school branches that are also differentiated on territorial level.

2.2.3 Determinants of school tracking

Literature, especially economical and sociological, presents several works that analyze and study determinants of school tracking. The choice of the secondary school track is extremely important because it represents a relevant mechanism that deeply affects the intergenerational persistence of educational attainment and labor markets returns across different social classes (Dustmann, 2004). Persons' education (in terms of achievement and attainment[7]) affects their future life. More qualified people have better working positions, higher wages, more satisfaction, their are healthier and so on. Consequently, researchers have payed attention about school tracking in optic of intergenerational mobility and social inequality. If pupils do not have the same chance to choose a determined track, social inequalities can be produced (or re–produced).

Gender is the first determinant of secondary school choice. European statistics show how youths opt for gender stereotyped working position. Mocetti (2012) shows that females have more propensity to choose academic tracks rather than technical and vocational schools. Academic schools prepare pupils for teaching, translation, secretarial duties that are traditionally feminised occupations. Immigrant status also affects the choice of secondary school track. If natives are enough homogeneous into different school tracks, immigrants tend to be segregated into technical and vocational schools, controlling for prior school outcomes. The segregation increases considering first generation, more present into vocational and training centers (Barban and White, 2011).

In addition, there exists a strong dependence between parental education and the children' choice of school track and this dependence is more accentuated for males rather than females (Checchi and Flabbi, 2013). More educated parents drive their sons and daughters to enroll academic schools. Conversely, less educated parents give more importance to work, guiding their children to technical and vocational schools. Connected to what has just been said, pupils of upper social classes are more represented in academic schools, while lower classes are are systematically under–represented (Azzolini and Vergolini, 2014) and over-represented in technical and vocational tracks. A strong dependence exists between social class of origin and secondary school choice. In addition, the relationship between social class and the choice of secondary school grew over time: absolute inequities in the

[7]Achievement refers to students' academic performance, development and cognitive skills. Attainment refers to qualifications and academic degrees obtained by individuals (Boudon, 1974).

probability of enrolling in academic schools decrease, but relative inequality persisted. Pupils from upper classes tend to attend academic schools, while pupils from working class tent to choose technical and vocational schools (Panichella and Triventi, 2014).

Aspiration is another important determinant of school tracking decision. Students and parents with high school aspiration tend to choose schools that provide academic and general curricula (R. Becker, 2003). A possible explanation is that pupils with high educational aspiration tend to enroll academic schools because are those that prepare better for post–secondary education. Conversely, pupils with low educational aspiration choose technical and vocational schools because are those that prepare better for the labor market. However, educational aspiration is affected by socio–economic background. Educated parents motivate their children to study and they transmit them the importance of scholastic success as a channel for future job carrier (Barone, 2006; Sewell and Shah, 1968).

2.3 Differential Item Functioning

Differential item functioning, as said in section 2.1, is the statistical instrument used to detect possible unfair tests between individuals from different groups (for example, gender and ethnic group). DIF occurs when the functional relationship between response variable and latent trait differs for the groups. Formally, Let Y as the response to a specific item, θ as the latent trait (for example maths capacity) and G as the grouping variable, DIF is present if

$$f(Y|\theta, G = R) \neq f(Y|\theta, G = F)$$

where R refers to reference group and F to focal group. Usually, the group that it is assumed to have some advantages is the reference group, while that it is assumed to have some disadvantages is the focal group. Nevertheless, this definition does not affect DIF detection. As an example, we are interested to study DIF in a mathematics test between males and females. From technical reports (INVALSI, 2016; OECD, 2015) emerge that males, in average, outperform females. Thus, in DIF analysis, males will be reference group, while females will be focal group.

The possible differential in the probability to give a correct answer to an item between

groups may reflect the true difference in ability between groups. Consequently, it needs to match individuals to the same level of ability in order to detect DIF items. It is essential that DIF analysis is developed only for matched groups in order to avoid Simpson's paradox (Simpson, 1951). Again a gender example, it is possible that an item results more difficult for females rather than males, while, if we control only for students with same ability, the same item may result less difficult (Osterlind and Everson, 2009). Literature provides two matching criteria for DIF detection analysis. The first directly involves the latent trait estimated by an IRT (Item Response Theory) model. The second uses the total test score as a proxy of ability, because observed score has high correlation with the IRT score (Tay, Huang, et al., 2016).

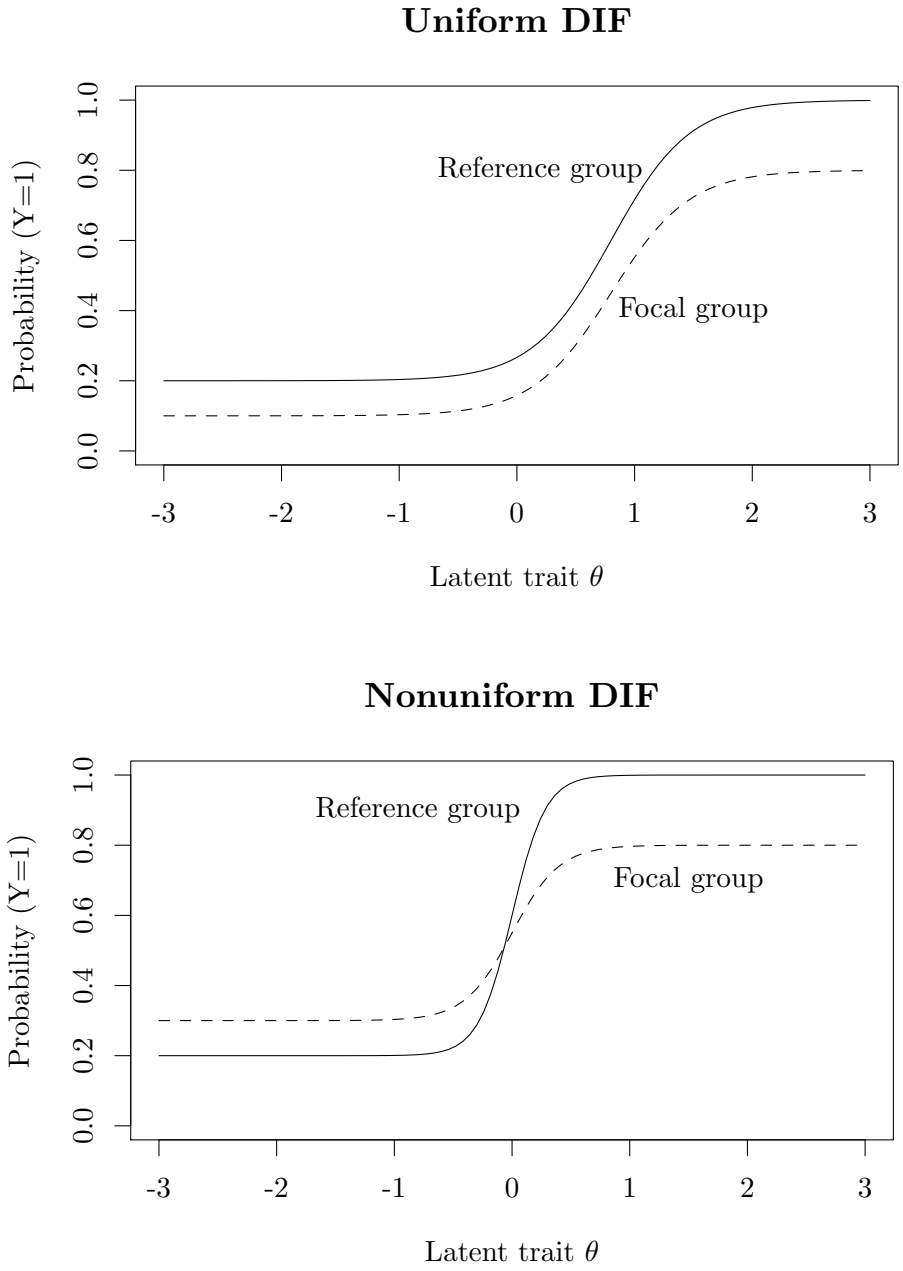
Differential item functioning can be uniform or nonuniform. The first one (represented at the top of figure 1), the simplest form of DIF, arises when the difference between reference and focal group remains constant across the continuum of latent trait θ. This means that one group (usually reference group) has the same amount of advantages throughout the underlying latent trait. Differently, nonuniform DIF (represented at the bottom of figure 1) occurs when the group's advantage changes in a certain point of the latent trait's distribution. It can be seen as an existing interaction between θ and group.

2.3.1 Developments and methods

Attention for detecting DIF starts from about the 60's in United States. In the last 30 years, the most important educational organizations have started to pay attention on this issue in order to analyze the tests validity. Zumbo (2007) distinguishes three generation of DIF analysis[8]. The first generation is characterized by the study of different performances in educational tests concerning demographical variables, such as gender and race. The purpose of this generation is to assess whether the difference in performance reflects the reality or are due to bias tests. The second generation refers the period in which the psychometric literature distinguishes between "impact" and "measurement equivalence" of a test. This period is marked by a development of new statistical tools and methods in order to detect DIF. In the last generation, scholars take into account new instruments for DIF detection. The DIF analyses involve test characteristics (item format and item

[8]Zumbo does not suggest a distinction in historical periods. He suggests a linear stepwise progression of knowledge and thinking.

Figure 1: Uniform and nonuniform differential item functioning.



content) and contextual variables (socioeconomic status, class size and so on) that may affect the test performance.

Especially from the second generation, described previously, several statistical methods about DIF detection were proposed. Nowadays, the majority of these methods is still used. The main methods for DIF detection are based on test score or on IRT modeling (Magis, Tuerlinckx, et al., 2015). The first approach uses a matching variable as a proxy

Table 1: Summarize of main methods for DIF detection.

METHOD	MATCHING CRITERION	Type OF DIF
Mantel–Heanszel *(Holland, D. T. Thayer, et al., 1988)*	Test score	Uniform
SIBTEST *(Shealy and Stout, 1993)*	Test score	Uniform
Logistic Regression *(Swaminathan and Rogers, 1990)*	Test score	Both
Lord's χ^2 *(Lord, 1980)*	Latent trait	Both
Raju's approach *(Raju, 1988)*	Latent trait	Both
Likelihood Ratio Test *(Thissen et al., 1988)*	Latent trait	Both

of the latent trait. Mantel Haenszel (MH) approach (Holland, D. T. Thayer, et al., 1988) compares proportion of correct response in focal and reference group. Simultaneously Item Bias Test (SIBTEST) (Shealy and Stout, 1993) is a comparison between weighted difference in the proportion of individuals in two groups which correctly answer an item conditioning on the underlying trait (French and Finch, 2015). Finally, Logistic Regression (LR) approach (Swaminathan and Rogers, 1990) is a generalized linear model which is able to identify both uniform and nonuniform DIF. LR regression compares, usually, three nested models: free from DIF, uniform DIF model and nonuniform DIF.

Conversely, the other models (based on IRT approach) involve differences in item parameters, estimated through IRT models. Lord's χ^2 (Lord, 1980) tests if statistical significance is present in the difference between estimated item parameters in the focal and reference group. Differently, Raju's approach (Raju, 1988) assesses significant differences in the focal group item characteristic curve and the focal group item characteristic curve. Likelihood Ratio Test (Thissen et al., 1988) involves two nested IRT models. It tests whether a model, constrained to present DIF on an item or multiple items, is significantly different from an other model, constrained to have no DIF. Table 1 summarizes the main statistical techniques for DIF detection analysis.

During the last few years, these methods for DIF detection have been developed and extended to various statistical features (Lee and Geisinger, 2014). Data for educational large-scale assessment are usually collected from a multilevel structure. Standard methods for DIF detection may underestimate standard errors whether it were ignored the multilevel

structure of the data. Thus it may lead to biased estimates, in statistical sense. Some researchers (French and Finch, 2015; French and Finch, 2013; Beretvas et al., 2012; French and Finch, 2010; Cho and A. Cohen, 2010) have developed new methods and extensions of pre-existing methods in order to control the multilevel data structure. Tay et al. (2015), Strobl et al. (2015) and Tutz & Berger (2015) have developed statistical methods to include several covariates in order to detect DIF. The firsts suggest Item Response Theory With Covariates (IRT-C), while the others propose partitioning recursive models. Svetina and Rutkowski (2014), Magis et al. (2012) and Woods et al. (2012) propose methods for DIF detection with multiple groups. Finally, mixture Item Response Theory models (Cho and A. Cohen, 2010) and approaches of multidimensional IRT (Walker and Sahin, 2016) have been developed.

2.3.2 Mantel–Heanszel statistic

In a seminal paper, Mantel and Haenszel (1959) have proposed a procedure for the study of matched groups (Dorans and Holland, 1992). In the 80s, this approach was developed for DIF detection by Holland (1985) and later by Holland and Thayer (1988). This procedure treats the DIF detection problem through three-way contingency tables, where the three dimensions are: whether one correctly (or incorrectly) responds to an item, the group membership and the total score. In this approach the sum score is used as a matching variable for the latent trait and the conditioning variable is categorized into several (j) bins. This procedure allows us to compare the item responses between the reference and the focal group conditioning on the various levels of matching variable.

The Mantel–Haenszel procedure can be understood by the sequent table:

Table 2: Mantel–Haenszel contingency table.

	Item score		
Groups	$Y_i = 1$	$Y_i = 0$	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{fj}
Total	M_{1j}	M_{0j}	T_j

It is possible to construct the Mantel–Haenszel statistic:

$$MH_{\chi^2} = \frac{[|\sum_j A_j - \sum_j E(A_j)| - 0.5]^2}{\sum_j Var(A_j)} \tag{1}$$

where $E(A_j) = \frac{N_{rj} M_{1j}}{T_j}$ and $Var(A_j) = \frac{N_{rj} M_{1j} N_{fj} M_{0j}}{T_j^2 (T_j - 1)}$. MH_{χ^2} follows a χ^2 distribution with one degree of freedom. This statistic assesses the null hypothesis (H_0) that there is no association between item responses and group membership. In this formulation we are interested to test a null hypothesis (H_0) versus an alternative hypothesis (H_1), where

$$H_0 : \frac{A_j/C_j}{B_j/D_j} = 1 \qquad versus \qquad H_1 : A_j/C_j = \alpha(B_j/D_j) \tag{2}$$

The Mantel–Haenszel statistic allows us to provide the DIF effect size (a linear association between the row and the column variables in table 2) through the common odds ratio. For an item and for $j$sm level of matching variable we can construct the odds ratio α_j:

$$\alpha_j = \frac{A_j D_j}{B_j C_j} \tag{3}$$

and for all levels of matching variable we have:

$$\hat{\alpha}_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \tag{4}$$

finally, the logarithm of common odds ratio $\hat{\alpha}_{MH}$ is normally distributed and is used as effect size measure:

$$\lambda_{MH} = log(\hat{\alpha}_{MH}) \tag{5}$$

Educational Test Services (ETS) uses a scheme in order to classify the DIF effect size. The scheme follows delta scale (Holland, D. T. Thayer, et al., 1988) and it is computed as follow:

$$\Delta_{MH} = -2.35\lambda_{MH} \tag{6}$$

with sequent cut-offs:

- Large DIF (class C) $|\Delta_{MH}| > 1.5$;
- Moderate DIF (class B) $1 < |\Delta_{MH}| \leq 1.5$;

- Small DIF (class A) $|\Delta_{MH}| \leq 1$.

To sum up, the Mantel–Haenszel approach is a three-step procedure. In the first step, it examines whether MH_{χ^2} is statistically significant. Secondly, it assesses the DIF effect size, through the size of common odds ratio. In the final step, it is possible to judge the significance of DIF using the ETS classification scheme. The main advantage of the Mantel–Haenszel approach resides in his powerful, that is in the capacity of detecting DIF items correctly. In addition, it provides both a statistical test and effect size. Finally, it is accessible through popular statistical software (SAS, SPSS, R). Nevertheless, this procedure has some limitations. First of all, it does not test for nonuniform DIF. Secondly, there is the need to choose bins or levels in which put the matching score that may be affect the statistical decision of DIF.

2.3.3 Logistic regression

Differently from Mantel-Haenszel procedure, the Logistic Regression (LR) is a parametric approach. LR for DIF detection has been proposed by Swaminathan & Rogers (1990). Like Mantel-Haenszel, LR assumes the total score as a proxy of the latent trait. The general idea of Logistic Regression for DIF detection is tested three nested logistic models for all items. The dependent variable is categorical and represents the likelihood of responding correctly or incorrectly to an item and it is conditioned on matching criterion, grouping variable and an interaction term (Osterlind and Everson, 2009).

The baseline model:

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \alpha + \beta_0 T \tag{7}$$

predicts the correct answer to an item conditioning only on the proxy of latent construct (T), where α is the model's intercept and β_0 is the parameter of total score. T is the sum of the scores of all test items without the considered item.

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \alpha + \beta_0(T) + \beta_1(group) \tag{8}$$

Adding β_1 at baseline model (model 8), the parameter of group membership, we can assess the presence of uniform DIF. A χ^2 test with one degree of freedom compares the improvement of model 8 with respect to baseline model. If adding the group membership

variable improves the fit, than uniform DIF is detected.

$$ln\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) = \alpha + \beta_0(T) + \beta_1(group) + \beta_2(T \cdot group) \quad (9)$$

Finally, it is possible to test a nonuniform DIF (model 9), adding the interaction between the group membership and the proxy of the latent trait (where β_2 is the interaction's parameter). Always through a χ^2 test with one degree of freedom it is possible to assess the fit improvement. If adding the interaction improves data fit, than nonuniform DIF might be present in the item.

As Mantel-Haenszel approach, there are some procedures in Logistic Regression in order to assess the size of the DIF. Zumbo and Thomas (1997) consider a large DIF whether the item displays a p-value ≤ 0.01 and $R^2 > 0.13$. This method is criticized for being too indulgent. More conservative, Gierl and McEwen (1998) proposed a different scheme:

- Large DIF (class C) $R^2 \geq 0.07$ and χ^2 test significant;
- Moderate DIF (class B) $0.035 \leq R^2 < 0.07$ and χ^2 test significant;
- Small DIF (class A) $R^2 < 0.035$ or χ^2 test non significant.

As said previously, the main advantage of LR for DIF detection is that it is able to identify both uniform and non uniform DIF. Another advantage of LR is its flexibility: "*LR model also allows for conditioning simultaneously on multiple abilities and can be extended to multiple test taker groups*" (Wiberg, 2007, p. 15).

2.3.4 IRT approach for DIF

Lord's χ^2 and Likelihood Ratio Test for DIF detection analysis are parametric methods which require Item Response Theory models. IRT is a paradigm which aims to specify information about test takers' latent construct and the characteristics of test items (Osterlind and Everson, 2009). When test responses are dichotomous, the conventional IRT models are logistic regressions with different parameters, that are the characteristics of test items (Özdemir, 2015). The characteristics of test items are difficulty, discrimination and guessing parameters and they identify three different IRT models. The simplest IRT model is the One–Parameter (1PL) model, also known as Rasch model (Rasch, 1960). Rasch

model estimates only difficulty parameter, fixing discrimination parameter equal to 1. The Rasch model is represented by:

$$P(Y_j = 1|\theta_i) = \frac{exp(\theta_i - \beta_j)}{1 + exp(\theta_i - \beta_j)} \tag{10}$$

where $P(Y_j = 1|\theta_i)$ is the probability to correctly answer to item j while θ_i and β_j represents, respectively, the latent trait of person i and the difficulty parameter for item j. Rasch model implies that test takers that correctly answer to the same number of items will have the same level of ability θ. The second IRT model, namely the Two–Parameter model (2PL) represented in equation 11, assumes a non–fixed discrimination parameter (Birnbaum, 1958; Birnbaum, 1968). In other words, 2PL estimates the discrimination parameter (a_j) which represents the j^{th} item capacity to distinguish people with different abilities.

$$P(Y_j = 1|\theta_i) = \frac{exp[a_j(\theta_i - \beta_j)]}{1 + exp[a_j(\theta_i - \beta_j)]} \tag{11}$$

The last IRT model, namely the Three–Parameter model (3PL), has less fortune in assessment program with respect to 1PL and 2PL. 3PL (equation 12) adds guessing parameter (c_i) to 2PL model. Guessing parameter represents the probability of a subject with very low ability to correctly answer to item j. In other words, 3PL allows to control for possible random answers.

$$P(Y_j = 1|\theta_i) = c_i + (1 - c_i)\frac{exp[a_j(\theta_i - \beta_j)]}{1 + exp[a_j(\theta_i - \beta_j)]} \tag{12}$$

All three models have two fundamental assumptions: unidimensionality and local independence. Unidimensionality refers to existence of only one latent trait underlying the test responses. For example, only one ability is associated with pupils' responses to a mathematical test. If this assumption is not verified it is possible to resort to multidimensional IRT (MIRT) models[9] which allow more than one latent trait underlying the test responses. Differently, local independence states that items responses are independent of each other given a level of latent trait. Multilevel IRT approach (French and Finch, 2010; Kamata, 2001) can be an optimal strategy in cases of local independence violation.

After briefly introducing IRT frameworks, now we focus on Lord's χ^2 for DIF detection.

[9]For more details, it is possible to consult Walker and Sahin, 2016.

Lord (1980) provided a simple method in which items parameters are compared between reference and focal groups. A statistic d tests, for each test item, the null hypothesis that difficulty parameter ($\hat{b}$) is equal for reference and focal groups. The statistic is given by

$$d = \frac{\hat{b}_R - \hat{b}_F}{SE(\hat{b}_R - \hat{b}_F)} \tag{13}$$

where

$$SE(\hat{b}_R - \hat{b}_F) = \sqrt{[SE(\hat{b}_R)]^2 + [SE(\hat{b}_R)]^2} \tag{14}$$

Under the null hypothesis $H_0 : b_R = b_F$ the statistic d is approximately distributed as standard normal distribution. In addition, Lord suggested an appropriate DIF test in situations where 2PL or 3PL models adapt better to data than Rasch model. The new statistic χ^2_L assesses simultaneously the differences of difficulty ($\hat{b}$) and discriminant ($\hat{a}$) parameters between focal and reference groups. The statistic is approximately distributed as a chi–square distribution with 2 degree of freedom and is given by

$$\chi^2_L = \hat{\nu}' S^{-1} \hat{\nu} \tag{15}$$

where $\hat{\nu}' = (\hat{a}_R - \hat{a}_F, \hat{b}_R - \hat{b}_F)$, while S represents the estimated variance–covariance matrix of $\hat{\nu}$. Lord's χ^2 has the advantage to use directly the latent trait θ as matching criterion with respect to Mantel–Haenszel approach and Logistic Regression.

As Mantel-Haenszel and Logistic Regression approach, Lord's χ^2 under Rasch model proposes an effect size measure. This measure, similar to Mantel-Haenszel effect size measure, is computed as -2.35 times the difference between item difficulties of the reference group and the focal group (Penfield and Camilli, 2006, p. 138). The effect size measure is classified with ETS delta scale (Holland and D. Thayer, 1985).

Likelihood Ratio Test (LRT) has been developed recently as IRT–based DIF method (Thissen et al., 1988). This approach compares likelihood of two models. In the first model (L(C)) parameters are constrained to be fixed between reference and focal group, while in the second model (L(A)) parameters are free to vary. The LRT is computed as

$$G^2 = 2ln\Big[\frac{L(A)}{L(C)}\Big] \tag{16}$$

G^2 is approximately distributed as a chi–square distribution. The degree of freedom correspond to the number of constraints associated to the IRT model. For example, one degree of freedom with Rasch model in which only item difficulty parameters are free to vary. Differently, two degree of freedom under 2PL model in which both item difficulty and discrimination parameters are free to vary.

2.4 A new bias approach

Psychometric literature usually refers to differential item functioning, bias and impact when it studies and analyzes test fairness or inequity. As said in section 2.1, DIF and bias definitions are very similar and they are usually used as interchangeable. It is possible to distinguish them through their statistical (DIF) and qualitative (bias) connotation. Differently, impact definition is clearer. Traditionally, impact refers to real difference in group performances (Wu et al., 2017). From a statistical point of view, impact involves groups' discrepancy in the measurement outcomes (Millsap and Everson, 1993). Outcomes averages among groups are usually used as instrument of impact detection.

Recently, Zumbo et al. have provided new developments in DIF detection analysis with two articles published on *Practical Assessment, Research and Evaluation (2016)* and on *Frontiers in Education (2017)*. In particular, they redefine DIF, bias and impact terms (the triplet DBI) and they provide a new methodology in order to detect bias and impact. Redefinition decouples the second and the third term from DIF and the methodology guarantees statistical techniques to detect them. The starting point is the redefinition of bias. *"It is biased to compare response outcomes among groups if the observed response difference is attributable to the groups that are equal in the measured construct"* (Wu et al., 2017, p. 4). From this definition four points emerge:

1) the group composition is the reason of the detected DIF;

2) the comparison, and not the item, is biased;

3) such as for DIF detection, bias can be detected only comparing individuals with same latent trait;

4) intention of attributional claim.

Bias is present if the DIF identification is due to group comparison. We can image an example in which test fairness is studied for a maths test between pupils from different schools. If an item presents DIF, it is possible to talk about bias only if the schools comparison is the reason for which the groups respond to the item in different way. Consequently, bias refers to the group comparison (in the example the comparison among schools) and not to item, as traditionally bias is conceived. Possible differential in the probability of correctly answering to the item between schools may reflect the true difference in ability. Therefore, such as for DIF detection (section 2.3), bias detection is possible only to compare individuals with the same latent trait in order to avoid Simpson's paradox. Referring to the new methodology, terms bias and DIF are interchangeably used in this work. The last point of bias definition is crucial. Due to attributional claim, control of possible factors that may confound this attributional process is necessary. In other words, it is not possible to directly compare the groups because they may be imbalanced with respect to covariates that can confound the attributional process. When groups are imbalanced it is possible that problems of selection bias are present.

2.4.1 Propensity score

Attributional claim requires randomized experiment studies where group assignment is random. In this kind of study groups are balanced and, consequently, they are comparable. Nevertheless, educational standardized tests concern observational studies in which group assignment mechanism is not random. In observational studies it is difficult to attribute whether group differences are due to group membership or to pre–existent group differences. Consequently, there is a need to find a way to balance groups; with imbalanced groups, statistical analysis can lead biased estimates. This bias, namely selection bias, is introduced by a non–randomization of group selection, thereby sample does not guarantee representativeness of the population. Rosenbaum & Rubin (1983) proposed the propensity score in order to reduce the selection bias. Propensity score was previously popular mainly in medical and economic research, but it has recently gained importance also in psychological, social and educational research and policy evaluation works (Austin, 2009a). The propensity score is defined as

$$e(X_i) = Pr(G_i = 1|X_i)$$

where G_i is an indicator for group membership. Usually, $G_i = 1$ refers to treatment group, while $G_i = 0$ refers to control group[10]. It is usually estimated by logistic regression:

$$Pr(G_i = 1|X_i) = \frac{exp(\beta_0 + \beta(X_i))}{1 + exp(\beta_0 + \beta(X_i))}$$

where β_0 is the intercept and β is the vector of coefficients related to covariates X_i. Observations with the same, or similar, propensity score have the same covariates distribution and they differ only for group membership. Propensity score matching follows a fundamental theorem: the balancing property:

$$G_i \perp X_i \mid pr(X_i) \tag{17}$$

From the formula 17 the distributions of group membership status G_i *"... and the observable control variables X_i are orthogonal to each other, once conditioning on the propensity score $p(X_i)$"* (Pellizzari, 2018, p. 1). In addition, propensity score matching allows to, as well as groups balancing, estimate the treatment impact in an observational study (Dehejia and Wahba, 2002; Dehejia and Wahba, 1999). Four main methods for balancing groups are common in literature: stratification, regression adjustment, weighting and matching.

In stratification subjects of a sample are divided into mutually exclusive subsets based on propensity score (Austin, 2011). Subjects into the same stratum will have the same, or very similar, propensity score; therefore, they will have same covariates distribution. First of all, all subjects are split into the subgroups according to propensity score. Here balanced property is checked and than quintiles of the estimated propensity score are used in order to reduce the confounding effect. This simple method guarantees an optimal selection bias reduction, approximately 90% (Rosenbaum and Rubin, 1984). Regression adjustment method adds to a regression model (the model choice depends on the nature of the outcome) a dummy variable referred to treatment status and the estimated propensity score (Austin, 2011). In this way it is possible to analyze the effect of grouping mechanism fixing, controlling, the estimated propensity score. Weighting, i.e. inverse probability of treatment weighting (IPTW), was proposed by Rosenbaum (1987) as a standardization

[10]In DIF context treatment group refers to focal group and control group to reference group.

based on a model. IPTW, in matching process, uses weights based on the propensity score in order generate a balanced sample in which the probability to group assignment is independent with respect to observed covariates. This method looks like survey sampling weights. This kind of studies weight survey samples in order to make representative a specific population (Morgan and Todd, 2008). Subject weight is defined as

$$w_i = \frac{G_i}{e_i} + \frac{(1 - G_i)}{1 - e_i}$$

where G_i is an indicator for group membership., while e_i refers to subject's propensity score.

Several methods belong to matching strategy. All matching methods aim to approximate random assignment mechanism. The main idea is to create matched sets/strata of treated and untreated subjects[11] using the estimated propensity score. Exact matching assigns subjects with the same value of propensity score to all strata. This method is not very common because many unmatched subjects can be present and, consequently, discharged, reducing sample size. Greedy matching (e.g., nearest neighbor) assigns subjects to strata in different way. A treated subject is randomly selected and the untreated with closest value of propensity score is assigned to him. The process carries on until all treated subjects are matched with untreated subjects.

Optimal matching works similar to greedy matching, but it does not match with closest value of propensity score. Optimal matching adopts an algorithm which minimizes the total differences in the estimated propensity score (Austin, 2009b) among treated and untreated subjects. Optimal pair matching and optimal full matching characterize this algorithm. The first one matches subjects in pair, discharging unmatched subjects. This involves in a sample size reduction, developing possible problem of under–representation and lower power for tests (Wu et al., 2017). Differently, the second one does not involve any type of discharging and it matches subjects using full data set. In particular, it is possible to match one treated with many untreated subjects (one–to–many) or to create matched sets with many treated subjects and one untreated (many–to–one). In both cases weights are used in order to adjust estimation of propensity score based on the number of subjects into all strata (Rosenbaum, 1991).

[11] In DIF context matched strata are formed by subjects from reference and focal group.

When you want to use propensity score matching you keep in mind an issue, called common support. The common support is an assumption of propensity score matching in which observations are matched according to their observed characteristics. Common support concerns situations in which the propensity score distribution between treated and controls overlaps. If groups (treated and controls) does not present common support, we are not able to match some treated to control observations and vice versa, because propensity score distribution does not perfectly overlap between group samples. Ignoring common support can produce a biased matching *"... because a comparison observation would be matched that is not sufficiently similar to the treatment observation it is matched to"* (Lechner, 2001, p. 3). Literature presents proposal strategies in order to overcome this problem estimating only a partial observed effect (Angrist and Imbens, 1995; Heckman and Robb, 1986; Rubin, 1974).

Propensity score matching is not a perfect method for reducing selection bias. King and Nielsen (2018) have shown that sometimes propensity score matching increases group imbalanced and statistical bias. They refer to PSM *paradox*: *"... if ones data are so imbalanced that making valid causal inferences from it without heavy modeling assumptions is impossible, then the paradox we identify is avoidable and PSM will reduce imbalance but then the data are not very useful for causal inference by any method."* (King and Nielsen, 2018, p. 1).

2.4.2 Summary of the new methodology

From the attributional claim emerged by the redefinition of bias term, the first step for bias detection is to balance groups with respect to covariates. Therefore, one of techniques explained in previous section (section 2.4.1) for balancing groups should be adopted. There is none better than others *a priori*, but in practice it is necessary to assess which of techniques guarantees the better balance of the covariates distribution (Liu et al., 2016) and, subsequently, to choose the best technique for the subsequent analysis. After balancing groups, bias can be detected with traditional DIF detection analysis[12]. To sum up, bias detection analysis can be summarized by the following points:

[12]Traditional DIF detection analysis does not consider the dependence structure of data, generated by matched sets (Liu et al., 2016). Section 3.3.3 provides more information and details in order to apply DIF detection analysis with this data structure.

- Balance of the covariate distribution between groups.
- Detection of DIF for balanced groups.

Besides bias detection, as said previously, Zumbo et al. provide a new methodology for impact detection. Studying group impact on the probability to correctly answer to an item is the ultimate aim of educational experts, sociologists and policy makers. Indeed, finding group impact may reflect possible group inequality or disparity in measured construct (ability or achievement). It is possible to detect group impact only if the comparison is unbiased, hence the first step involves checking for bias detection. Differently, with respect to bias detection, matching criterion is not required. Nevertheless, impact detection requires balanced groups due to control of confounding factors for the attributional process. Conversely, three steps characterize impact detection:

- Balance of the covariate distribution between groups.
- Check for bias detection.
- Detection of impact (group differences) for balanced groups, only for items in which the comparison is not biased.

In conclusion, this new analysis framework guarantees new perspective in test fairness analysis. In particular, it is possible to attribute to group composition possible items flagged as DIF.

3. Data and methods

This part of the work is dedicated to find a way for assessing Zumbo's methodology and traditional DIF detection analysis techniques. This chapter is divided into three parts. The first one (section 3.1) describes data and variables for the analyses, focusing on the covariates balancing of grouping variable, e.g., school tracks. The second part illustrates the simulation design (section 3.2). It describes different simulated scenarios, starting from real data in order to reflect similarly covariates distribution. Section 3.2 describes the generation of latent trait θ, grouping variable G, responses variables Y and the factors which are manipulated to obtain scenarios. The last part of the chapter (section 3.3) provides statistical techniques and methods for assessing accuracy of DIF detection methods.

3.1 Data

Empirical issues address the aims of this work, as anticipated in section 2.1. In particular, the main aim is to answers these questions: *Do INVALSI tests measure the same ability among pupils from different school tracking? Is this instrument unfair or biased with respect to school tracking?* Hence, INVALSI data are the starting point of this work. INVALSI (*Istituto Nazionale per la VALutazione del SIstema educativo di istruzione e di formazione–National Evaluation Institute for the School System*) is a research institute born in 1999 and its primary aim is to assess the Italian education system. The most famous instrument for the evaluation is the INVALSI standardized tests. INVALSI tests are administered all years, starting from 2009/2010, to all students of II and V grade of primary school, III level of secondary school of I grade and II level of secondary school of II grade. INVALSI tests include three parts[13]: Italian language test, mathematics test and a student questionnaire for V grade of primary school and II level of secondary school of II grade. The tests time differs according to educational level: from 45 minutes for II grade of primary school to 75 minutes for II level of secondary school of II grade. The student questionnaire lasts 30 minutes. The student questionnaire gathers information about student's background, family background, free time activities, opinions and behaviors about school. A sampling is carried out at school grade and pupils from sampled schools

[13]Starting from 2018, INVALSI administers English test for students of V grade of primary schools and III level of secondary schools of I grade.

perform test in the presence of external observers. The simulation part of this thesis refers to sample of INVALSI tests 2015/2016 that contains 33992 observations.

3.1.1 Dependent variables

Educational Italian system is characterized by a school tracking for the I level of upper secondary school. At the age of fourteen, Italian pupils must choose a track of second level of secondary school: academic, technical, vocational schools or vocational training courses. In INVALSI sample there are no information about vocational training courses. Therefore, the tracking variable can take as values *academic*, *technical* and *vocational*. Table A1, in appendix A, represents school tracking composition of INVALSI sample.

This thesis has the interest of assessing if INVALSI tests are unfair for pupils from different tracks. In chapter 2, a new methodology/framework has been presented, useful for assessing research question. The simulation analysis considers only academic and technical schools. Therefore, for the simulations a sample reduction (N=25058) is present because pupils from vocational track are not considered for simplifying the simulation study. Finally, in DIF context, academic track is the reference group, while technical track is the focal group. For simplicity, the simulation considers only maths test. At national level, academic track outperforms technical tracks, that, in turn, outperforms vocational both in Italian language and maths test (INVALSI, 2017c; INVALSI, 2016). Hence, here, two dependent variables are considered: tracking and mathematics proficiency.

3.1.2 Independent variables

Tracking is not randomized, but pupils have to make a choice that can depend on individual characteristics. Section 2.2.3 provided a review on possible determinants on school tracking. From the literature review, independent variables are selected in order to assess possible imbalanced covariates in school tracks and to create scenarios useful for the simulation.

First of all, gender is a determinant of upper secondary school choice. European statistics show how youths opt for gender stereotyped working position. Therefore, we expect that females are over–represented into academic track. Immigrant status (here better citizen) also affects the choice of secondary school track. If natives are enough homogeneous into different school tracks, immigrants tend to be segregated into technical

and vocational schools, controlling for prior school outcomes. For simplicity, this variable considers only natives *versus* non–natives. This is a limitation because three types of immigrant exist: I generations, II generations and mixed–parentage who present different behaviors. I and II generations have their own behaviors, while mixed–parentage pupils have similar behavior to natives (Azzolini, Schnell, et al., 2012).

Educational aspiration is another important determinant of school tracking decision. Students and parents with high school aspiration tend to choose schools that provide academic and general curricula (R. Becker, 2003). A possible explanation is that academic tracks are those that prepare better for post–secondary education. Aspiration affects also school achievement: pupils with high aspirations outperform those with low aspirations (Khattab, 2015). An INVALSI question (Q12), from student questionnaire, is used as a proxy of pupils' aspiration. The question asks students which is the qualification they intend to achieve. In this work, aspiration variable is treated as a dichotomous variable where 1 refers to university degree aspiration, while 0 refers to not university degree aspiration.

It is important to consider geographic area when you study Italian state: there are significant differences among north, middle and south (especially between north and south) in many spheres of individual life. In education context, northern pupils tend to outperform pupils from the middle, who themselves outperform southern pupils, both in Italian language and math test (INVALSI, 2017c; INVALSI, 2016).

Finally, there exists a strong dependence between parental education (Checchi and Flabbi, 2013), social class (Azzolini and Vergolini, 2014; Panichella and Triventi, 2014) and the students' educational track choice. INVALSI provides a synthetic index in order to simply this complex phenomenon. The continuous Economic, Social, and Cultural Status index (ESCS) is computed starting from discrete indicators like the parents occupational status and their education. In particular, this index is computed by a principal component analysis on three indexes, detected from the student questionnaire: the parent occupational status, the parent education (years of formal schooling) and a proxy of family wealth (the household possession). By construction, ESCS index has the mean equals to zero and the variance is set to one(Ricci, 2010). For INVALSI sample (without vocational track), this index varies between -3.376 and 2.048 with mean and variance, respectively, equal to 0.118

and 0.890. We expect that students from academic track exhibit higher ESCS index values than students from technical track (Azzolini and Vergolini, 2014; Panichella and Triventi, 2014; Checchi and Flabbi, 2013). For an easier computation, this index is considered as a discrete variable, despite of a lost of information. In particular, the first quartile (-0.522), the median (0.142) and the third quartile (0.806) are used as cutoff points. Figure A1, in appendix A, shows ESCS index box plot graph for INVALSI sample (without vocational schools). Hence ESCS index presents four categories: I quartile, II quartile, III quartile and IV quartile.

Table 3: Covariates balancing between groups in INVALSI sample.

	Academic	Technical	p	SMD
n	14185	10873		
Gender				
Male (%)	38.6	66.0	<0.001	0.571
Citizen				
Not Italian (%)	7.0	11.4	<0.001	0.154
Aspiration				
University degree (%)	79.7	34.1	<0.001	1.036
Area (%)			0.003	0.043
North	46.9	45.5		
Middle	19.7	19.1		
South	33.4	35.5		
ESCS (%)			<0.001	0.533
I quartile	21.9	39.5		
II quartile	22.5	26.8		
III quartile	24.3	20.7		
IV quartile	31.2	13.0		

Table 3 shows the covariates balancing between pupils from academic and technical tracks. It presents the percentage differences in average and no one formal test on the differences among covariates averages. This because we are not interested in testing if the means differ statistically between the two groups, but because we are interested if

covariates are over or under–represented between groups. Males tend to attend more technical than academic ones. Academic tracks tend to have fewer not Italians, while pupils with high aspiration are over–represented in these tracks. Finally, if the geographic area has no effect, pupils with high ESCS index are over–represented in academic tracks and under–represented in technical tracks. The results seem consistent with respect to literature review (Azzolini and Vergolini, 2014; Panichella and Triventi, 2014; Checchi and Flabbi, 2013; Mocetti, 2012; Barban and White, 2011; R. Becker, 2003).

The last column of the table presents the *standardized mean differences* (SMD), an information on imbalanced effect size.

$$SMD = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}+\hat{p}_{control}(1-\hat{p}_{control}))}{2}}}$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ denote the prevalence or mean of the dichotomous variable in treated and untreated subjects, respectively (Austin, 2009a, p. 3087). Herein, the reference group (academic track) is equivalent to treatment group, while the focal group (technical track) is equivalent to control group. Hence, treatment–control and reference–focal terms are used interchangeably.

The standardized mean differences were proposed in the psychological literature and Cohen (1988) suggested a SMD value of 0.2 as small, 0.5 medium and 0.8 large effect size. SMD is not influenced by sample size, therefore it can be used to compare balance in measured variables between subjects from different groups which can have different sample size (Austin, 2011; Austin, 2009a). From the table, it emerges that pupils' tracking allocation is not random, but it depends on the considered covariates, especially on gender, aspiration and ESCS index. INVALSI sample presents one large size imbalance (*aspiration*), two medium sizes (*gender* and *ESCS*), one small size (*citizen*) and one negligible (*geographic area*).

In addition, independent variables are correlated to maths proficiency. Table A2, in appendix A, presents means and standard deviations of INVALSI sample raw scores[14]. Males outperform females, Italians achieve better scores than not Italians and pupils with high aspiration outperform low aspiration pupils. North regions have the best performance, while the middle outperform the south. Finally, students at the top of ESCS

[14]Raw score is the sum of pupil's correct answer of entire test.

index distribution have better performance than students at the bottom of ESCS index distribution.

3.2 Simulation design

A stepwise simulation strategy was chosen in order to create scenarios similar to the real data. The first step (section 3.2.1) concerns the covariates generation. In the second step (sections 3.2.2 and 3.2.3) latent trait and grouping variable are simulated. Starting from previous steps, in the third step (section 3.2.4), responses variables are simulated. At this point, section 3.2.5 describes factors that are manipulated in order to create different scenarios.

3.2.1 Covariates generation

The first step of simulation strategy concerns the generation of covariates. The idea is to simulate covariates with high confidence similarity with the real data. Therefore, when a simulation scenario is generated, the proportions of main covariates reflect the distributions of them for INVALSI sample. The generation of the first four covariates (*gender*, *citizen*, *aspiration*, and *geographic area*) are based on the proportions in table 4. Note that results may not have an exact N value due to rounding process.

Table 4: Proportion of test takers by gender, citizen, aspiration and geographic area.

	University degree aspiration					
	Italian			Not Italian		
	North	Middle	South	North	Middle	South
Male	0.111	0.042	0.090	0.009	0.005	0.003
Female	0.146	0.059	0.108	0.015	0.007	0.004
	Not University degree aspiration					
	Italian			Not Italian		
	North	Middle	South	North	Middle	South
Male	0.093	0.042	0.085	0.013	0.009	0.003
Female	0.064	0.025	0.047	0.012	0.005	0.003

The Economic, Social, and Cultural Background index depends on *gender*, *citizen* and *geographic area*, whereas *aspiration* depends, in turn, on *ESCS*. *ESCS* is simulated after generating previous covariates in order to reflect as much as possible the real data. For each cell of table 4, the total simulated is multiplied by proportions of table A3, in appendix A. Suppose, for example, to having generated 100 Italian males from north with high aspiration, therefore 15 of them will belong to I quartile of *ESCS*, 21 to II quartile, 26 to III quartile and 38 to IV quartile. Also here, it is possible that results may not have an exact N value due to rounding process.

To sum up, five covariates are generated:

- Gender (dichotomous variable 1="Male", 0="Female");
- Citizen (dichotomous variable 1="Not Italian", 0="Italian");
- Aspiration (dichotomous variable 1="University degree", 0="Not University degree");
- Geographic Area (categorical variable 1="North", 2="Middle", 3="South");
- ESCS index (categorical variable 1="I quartile", 2="II quartile", 3="III quartile", 4="IV quartile").

This first simulation step is a sort of "*benchmark*" for the following steps. In other words, three datasets are created with three different numbers of observations[15]. These datasets are created maintaining the same covariates proportions of INVALSI sample. The next simulation steps are applied according to these three datasets.

3.2.2 Latent trait θ

The relationship between latent trait (θ), grouping variable (G) and covariates (X_p) must be high fidelity with the real data. Tay et al. (2016) developed a procedure to simulate latent trait[16], external covariates and their relationship high fidelity with real data. In particular, they predict standardized SAT[17] maths scores, used as a proxy of IRT latent trait, with a OLS regression. Subsequently, they estimate simulated ability using the previously predicted scores. Here, differently from Tay et al., an OLS model on real data

[15] See section 3.2.5.
[16] Maths proficiency.
[17] Scholastic Aptitude Test data.

(INVALSI maths test 2015/2016) predicts directly pupils' latent trait (θ_i), estimated from Rasch model. The linear model is:

$$\hat{\theta}_i = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + e \tag{18}$$

where X_p and β_p ($p = 1, \ldots, 5$) represent, respectively, predictor variables and parameters associated to them, while e represents error term. To simulate mathematics proficiency (latent trait) distribution coherent with respect to model 18, random normal distributions are simulated ($\theta_i \sim N(\theta_{mean}, 1)$) for each combination of simulated covariates ($\bar{X}_p$).

$$\theta_{mean}|\bar{X}_p = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p \bar{X}_p \tag{19}$$

In particular, each combination of covariates ($\bar{X}_p$) has its mean (θ_{mean} from model 19), while all standard errors are set to 1[18]. $\hat{\beta}_0$ and $\hat{\beta}_p$ are the OLS estimated parameters on sample, presented in table 5. In this way, θ depends on covariates which reflect the relationship among them in INVALSI sample.

3.2.3 Grouping variable G

Grouping variable generation follows similar approach to that for θ, in order to reflect the relationship between covariates and group membership highly close to reality. First of all, generalized linear model with logit link, (model 20) predicts the probability of belonging to one group, the reference group ($G_i = 1$), rather than another, the focal group ($G_i = 0$), in INVALSI sample. For the simulation, $G_i = 1$ refers to academic tracks and $G_i = 0$ refers to technical tracks. The predictor variables (X_p) are the same for the latent trait predictions.

$$P(\hat{G}_i = 1) = \frac{exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)}{1 + exp(\beta_0 + \sum_{p=1}^{P} \beta_p X_p)} \tag{20}$$

Once estimated β coefficients on real situation ($\hat{\beta}$), for each *simulèe* the grouping variable is drawn from a Bernoulli distribution ($P(G_i = 1) \sim Ber(P_i)$). The Bernoulli parameter

[18]The variance of mathematics competences for INVALSI sample 2015/2016 is approximately 1.55.

Table 5: OLS and GLM of total score and academic track.

	Dependent variable:	
	Total score	Academic track
	OLS	*GLM*
	(1)	(2)
Gender		
Male	0.548***	−1.108***
	(0.014)	(0.030)
Citizen		
Non Italian	−0.263***	−0.333***
	(0.025)	(0.053)
Aspiration		
University degree	0.698***	1.794***
	(0.015)	(0.031)
Area (ref. North)		
Middle	−0.345***	0.089**
	(0.019)	(0.041)
South	−0.699***	−0.007
	(0.016)	(0.034)
ESCS (ref. I quartile)		
II quartile	0.053***	0.283***
	(0.019)	(0.040)
III quartile	0.118***	0.569***
	(0.020)	(0.042)
IV quartile	0.238***	1.160***
	(0.020)	(0.044)
Constant	−0.167***	−0.633***
	(0.019)	(0.039)
Note:	*p<0.1; **p<0.05; ***p<0.01	

(P_i) depends on simulated covariates $(\bar{X}_p)$ and is computed in following way:

$$P_i|\bar{X}_p = \frac{exp(\hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p \bar{X}_p)}{1 + exp(\hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p \bar{X}_p)} \tag{21}$$

This simulation design allows to replicate scenarios similar to the real data. In addition, it guarantees to simulate imbalanced groups with respect to covariates, therefore it allows to assess new methodology presented in section 2.4.

3.2.4 Responses variables

From the previous steps three datasets are simulated with dependent and independent variables. The third step of simulation design concerns the generation of responses variables according to the three datasets. This work treats only dichotomous responses variables, where 1 refers to right answer and 0 to wrong answer. Consequently, the response of person i to item j, denoted by Y_{ij}, is drawn from a Bernoulli distribution ($Y_{ij} \sim Ber(P_j)$). Following Magis et al. (2015), the probability of success for the j^{th} item is computed under a Rasch model. More specifically, Magis et al. adopt a simulation design that guarantees easy DIF imputation. The probability of success for individuals from reference group (G_i=1) is computed by

$$P_j(\theta_i) = \frac{exp(\theta_i - \beta_j)}{1 + exp(\theta_i - \beta_j)} \tag{22}$$

while for individuals from focal group ($G_i = 0$) is computed by

$$P_j(\theta_i) = \frac{exp(\theta_i - \beta_j - \delta)}{1 + exp(\theta_i - \beta_j - \delta)} \tag{23}$$

where θ_i represents the latent trait for *simulèe* i and β_j is the difficulty parameter (or item location) of j^{th} item (see section 3.2.5 for more details). The parameter δ refers to DIF magnitude and it corresponds to the difference in item difficulty levels between the two group (Magis, Tuerlinckx, et al., 2015). With this simple parameter, it is possible to control and manipulate the proportion of DIF test items. It is clear that in presence of tests without DIF items, the value of δ is null. From this simulation design, it emerges that only uniform DIF is treated in this work, limiting results interpretations for situations in which DIF is constant for continuum of θ.

3.2.5 Manipulated factors

Four factors are manipulated in order to assess methods with respect to different possible situations. A different number of items can compose large–scale standardized test and researchers have found test lengths impact in DIF detection analysis (Tay, Huang, et al., 2016; Magis, Tuerlinckx, et al., 2015; Khalid and Glass, 2013; Uttaro and Millsap, 1994). Simulations are carried out for hypothetical standardized test composed by 20, 40 and 60 items.

Previous studies highlighted the effects of sample size in differential item functioning analysis (Magis, Tuerlinckx, et al., 2015; Khalid and Glass, 2013; Glas and Falcón, 2003; Uttaro and Millsap, 1994), therefore small sample (N=500), medium sample (N=1000) and large sample (N=2000) are taken into consideration.

Standardized tests can have a different amount of DIF items. Previous simulation studies considered small (5%. Berger and Tutz, 2016; Magis, Tuerlinckx, et al., 2015; Oliveri et al., 2014), moderate (10%. Berger and Tutz, 2016; Magis, Tuerlinckx, et al., 2015; Oliveri et al., 2014) and large (> 20%. Tay, Huang, et al., 2016; Oliveri et al., 2014; Woods et al., 2013; Cho and A. Cohen, 2010; Gómez-Benito and Navas-Ara, 2000) proportion of DIF items. This work considers three levels of proportion of DIF items: 0% (no DIF presence), 10% and 20%.

Moreover, previous studies found an influence of DIF magnitude/size (δ) on DIF detection analysis. Traditionally, three levels of DIF magnitude are considered: small ($\delta \leq 0.20$. Cho and A. Cohen, 2010; Gómez-Benito, Hidalgo, et al., 2009), moderate ($\delta = 0.40$. French and Finch, 2013; French and Finch, 2010) and large size ($\delta \geq 0.60$. Magis, Tuerlinckx, et al., 2015; French and Finch, 2013; French and Finch, 2010). The next simulations takes into account moderate ($\delta = 0.40$) and large ($\delta = 0.80$) DIF magnitude.

In addition, two different assumptions about difficulty parameters β are considered. Traditionally, these parameters are drawn from a normal distribution with null mean and variance sets to one (Tutz and Berger, 2016; Berger and Tutz, 2016; Magis, Tuerlinckx, et al., 2015; Khalid and Glass, 2013; Magis, Raîche, et al., 2011). Nevertheless, simulating βs from standard normal distribution makes the difficulty parameters focus around zero (the mean of standard normal distribution), excluding extreme values. This strategy precludes the presence of very easy or very difficult items. Simulating βs from an uniform

distribution (Weiss, 2014; Magis and Facon, 2012) allows to consider both very easy and very difficult items. Therefore, a second set of simulations takes into consideration difficulty parameters generated by an uniform distribution with parameters set to -2 and 2[19].

To sum up, the factors that compose scenarios of simulations are:

- Number of test takers (N = 500, 1000, 2000).
- Test length (J = 20, 40, 60).
- Percentage of items in which comparison is based (0%, 10%, 20%).
- DIF size (δ = 0.4, 0.8).
- β distribution assumption ($\beta \sim N(0,1), \beta \sim U(-2,+2)$).

In conclusion, methods for DIF detection analysis are assessed among 90 different situations/scenarios. In particular, 9 situations[20] with no DIF items and 36 scenarios[21] in which DIF is present, for a total of 45 scenarios. Finally, the assessment and comparison among methods consider the two different β assumptions, doubling scenarios for a total of 90 settings.

3.3 Methods

In each setting, described in section 3.2.5, 100 replications are generated. Twofold reason gives origin to this choice. First of all, the majority of simulation studies, that refer to DIF detection analysis, adopts the same choice (Berger and Tutz, 2016; Magis, Tuerlinckx, et al., 2015; Oliveri et al., 2014; Khalid and Glass, 2013; Jodoin and Gierl, 2001). Secondly, this choice allows to easily compute and interpret methods effectiveness and accuracy.

3.3.1 False alarm rate and power

The assessment of effectiveness and accuracy is based on false alarm rate (type I error) and power (1 minus type II error). On one hand, type I error concerns false positive detection.

[19] Traditionally β parameters vary between about -4 and 4. The choice of setting these parameters to -2 and 2 is consistent with respect to previous works and the estimations of difficulty parameters provided by INVALSI. For example, INVALSI 2015/2016 maths test for secondary schools presents difficulty parameters, estimated by Rasch model, with range -1.46 and 2.80 (INVALSI, 2016).

[20] 3 test lengths by 3 amount of test takers.

[21] 3 test lengths by 3 amount of test takers by 2 percentages of items in which comparison is based by 2 DIF sizes.

In DIF detection context, type I error refers to a situation in which a free DIF item is mistakenly identified as exhibiting DIF. Therefore, for each setting, false alarm rate is computed by the sum of items wrongly flagged as DIF divided for all items that should not present DIF. Acceptance level is set to 5%.

On the other hand, power concerns false negative detection, in particular it refers to the probability of making type II error. In DIF detection context, power refers to a situation in which DIF item is correctly identified as exhibiting DIF. Hence, power is computed by the sum of items correctly flagged as DIF divided for all items that should present DIF. Of course, power is not computed for settings with absence of DIF. Also for power analysis, nominal alpha level is set 0.05.

For example, you can image to compute false alarm and hit rate (power) for a test with 40 items, where 20% of items are DIF. Hence, for each replication, 8 items should be detected as DIF and 32 not. If the proportions of false positives are 3/32 in the 1st replication, 5/32 in the second, and so on, it is possible to compute the average of the test-wise type I error rate in the following way:

$$\text{False alarm rate} = \frac{1}{100} * (3/32 + 5/32 + \ldots)$$

Conversely, if the proportions of true positives are 7/8 in the 1st replication, 5/8 in the second, and so on, it is possible to compute the average of these proportions (power) in the following way:

$$\text{Power rate} = \frac{1}{100} * (7/8 + 5/8 + \ldots)$$

The incorrect identification of items flagged as DIF (inflation of type I error) is more problematic than the correct identification for two reasons. First of all, it could lead to remove items that are satisfactory, reducing the amount of items useful for the subsequent analyzes (Jodoin and Gierl, 2001). Secondly, it could get in the way "*... the development of a better understanding of the nature or underlying psychology associated with DIF*" (Jodoin and Gierl, 2001, p. 330). Therefore, we pay more attention to false alarm rate rather than power.

3.3.2 Matching analysis

Before assessing effectiveness and accuracy of DIF traditional methods and Zumbo's methodology, there is a need to keep in mind that groups are imbalanced with respect to covariates for construction. Selected covariates affect the group allocation mechanism. Therefore, without balancing statistical analysis may lead biased estimates and it is no possible to attribute DIF to group allocation, due to confounding variables.

Section 2.4.1 provided a brief review of statistical tools able to reduce selection bias and make comparable the groups. Zumbo et al. presented their works mainly with propensity score matching techniques. Hence, randomness of group allocation is implemented by propensity score matching techniques. In addition, propensity score matching creates a stratification useful for the methodology described in section 3.3.3.

Two different algorithms are available to create treatment–control (reference–focal) matches based on propensity scores: greedy and optimal matching. Greedy matching (e.g., nearest neighbor) assigns subjects to different strata. These strata contain one or many subjects from reference group and one or many subjects from focal group, which have equal or similar covariates distribution. A treated subject is randomly selected and the untreated with closest value of propensity score is assigned to him. The process carries on until all treated subjects are matched with untreated subjects.

Optimal matching works similar to greedy matching, but it does not match with closest value of propensity score. Optimal matching adopts an algorithm which minimizes the total differences in the estimated propensity score (Austin, 2009b) among treated and untreated subjects. Optimal matching algorithm is used by optimal pair matching and optimal full matching methods. The first one matches subjects in pair, discharging unmatched subjects. This involves in a sample size reduction, developing possible problem of under–representation and lower power for tests (Wu et al., 2017). Differently, the second one does not involve any type of discharging and it matches subjects using full data set. In particular, it is possible to match one treated with many untreated subjects (one–to–many) or to create matched sets with many treated subjects and one untreated (many–to–one) or to adopt a combination of one–to–many and many–to–one. Optimal algorithm performs sometimes better than greedy for producing closely matched pairs, only marginally better, but it is no better for producing balanced matched samples (Gu and Rosenbaum, 1993a).

Here, the matching analysis are carried out both with nearest neighbor matching and full matching with a combinations of one–to–many and many–to–one.

3.3.3 Conditional logistic regression

After matching analysis, traditional DIF detection analysis methods are inadequate because they are not able to treat and handle matched sets. In particular, they do not take account the dependence structure or the nested relationship of matched sets (Wu et al., 2017; Liu et al., 2016). Like for multilevel structure data, ignoring the dependence structure of data can lead biased estimation of standard errors (Raudenbush and Bryk, 2002). Consequently, biased estimations can compromise hypothesis test results, and, in DIF context, this can inflate type I error rates (French and Finch, 2013; French and Finch, 2010).

Zumbo et al. (2017, 2016) propose the conditional logistic regression models for DIF detection analysis in order to avoid these issues. Differently from conventional logistic models, parameters of conditional logistic regression are estimated using paired or clustered sample (Liu et al., 2016). In matched studies, conditional logistic regression can increase efficiency of estimations with respect to unconditional logistic regression[22] (Hosmer et al., 2013; Breslow et al., 1980). In addition, in matched studies, parameters estimated by traditional logistic regression could be biased (Breslow et al., 1980).

The simplest situation is conditional logistic regression for pair matching in which one unit of reference group is associated to one unit of focal group. The conditional likelihood function (Breslow et al., 1980) for the pair matching is:

$$l(\beta) = \prod_{k=1}^{K} \frac{1}{1 + exp[\beta^T(x_{1k} - x_{0k})]} \tag{24}$$

where k ($k = 1, 2, \ldots, K$) denotes the pairs; the vector β^T contains the covariates coefficients, whereas $(x_{1k} - x_{0k})$ is a data vector or matrix of covariate(s). In DIF context, for uniform DIF, the matrix $\beta^T(x_{1k} - x_{0k})$ can be split into $\beta_1(total_{1k} - total_{0k})$ and $\beta_2(group_{1k} - group_{0k})$ where *total* is the pupils' total score and *group* is the group membership variable[23]. "The constant term is summed to be equal to 0 and each pair corresponds to a positive outcome ($y = 1$)" (Breslow et al., 1980, p. 253), in order to fit conditional logistic regression

[22] Traditional logistic regression.

[23] It is possible to detect nonuniform DIF adding $\beta_3(total * group_{1k} - total * group_{0k})$.

with available statistical software.

It is possible to generalize equation 24 for more complex designs. Suppose that one unit of reference group is associated to M units of focal group (full matching one–to–many). Hence, each stratum k contains one unit of reference group and M_k units of focal group. The conditional likelihood function for this kind of design is:

$$l(\beta) = \prod_{k=1}^{K} \frac{1}{1 + \sum_{t=1}^{M_k} exp[\beta^T (x_{tk} - x_{0k})]} \tag{25}$$

where k $(k = 1, 2, \ldots, K)$ denotes the strata; the vector β^T contains the coefficients of covariates, whereas $(x_{tk} - x_{0k})$ is a data vector or matrix of covariate(s). t $(t = 1, 2, \ldots, M_k)$ denotes the t^{th} unit of focal group in the k^{th} stratum. In DIF context, for uniform DIF, the matrix $\beta^T(x_{tk} - x_{0k})$ can be split into $\beta_1(total_{tk} - total_{0k})$ and $\beta_2(group_{tk} - group_{0k})$ where *total* is the pupils' total score and *group* is the group membership variable[24].

Finally, suppose that M units of reference group are associated to one unit of focal group (full matching many–to–one). Hence, each stratum k contains one unit of focal group and M_k units of reference group. The conditional likelihood function for this kind of design is:

$$l(\beta) = \prod_{k=1}^{K} \frac{1}{1 + \sum_{t=1}^{M_k} exp[\beta^T (x_{1k} - x_{tk})]} \tag{26}$$

where k $(k = 1, 2, \ldots, K)$ denotes the strata; the vector β^T contains the coefficients of covariates, whereas $(x_{1k} - x_{tk})$ is a data vector or matrix of covariate(s). t $(t = 1, 2, \ldots, M_k)$ denotes the t^{th} unit of reference group in the k^{th} stratum. In DIF context, for uniform DIF, the matrix $\beta^T(x_{1k} - x_{tk})$ can be split into $\beta_1(total_{1k} - total_{tk})$ and $\beta_2(group_{1k} - group_{tk})$ where *total* is the pupils' total score and *group* is the group membership variable[25].

In DIF context, conditional logistic regressions are run for each item. Conditional logistic regression compares two nested models: the first one, baseline model, is a model with raw test score as only covariate and the second one, uniform DIF model, a model with grouping variable as additive covariate. Likelihood ratio test statistic is the test for assessing model significance: if the second one fits better data, then item is flagged as DIF. The statistic is computed by minus two times the difference between the log likelihoods of the two models. Likelihood ratio test statistic is asymptotically distributed as a chi-square

[24] It is possible to detect nonuniform DIF adding $\beta_3(total * group_{tk} - total * group_{0k})$.
[25] It is possible to detect nonuniform DIF adding $\beta_3(total * group_{1k} - total * group_{tk})$.

distribution, with the degrees of freedom equal to the difference in number of regression coefficients in the two models (Wu et al., 2017; Liu et al., 2016; Hosmer et al., 2013; Breslow et al., 1980).

3.3.4 Nagelkerke's R^2

As said in section 2.3, traditional DIF detection methods provide methods to compute DIF size measure. The effect size measure is a descriptive statistic that gives information about the magnitude or degree of DIF (Jodoin and Gierl, 2001). Using only null hypothesis significance testing for DIF detection has been criticized (Kirk, 1996; J. Cohen, 1994) because statistical test is sensitive to sample size. Therefore, using null hypothesis significance testing with an effect size measure could overcome this issue.

Logistic regression for DIF detection analysis uses R^2 for this kind of measure. Hosmer et al. (2013) indicate that there is no single measure in conditional logistic model similar to R^2 in multiple regression. Nevertheless, they suggest Nagelkerke's R^2 as adequate measure. Nagelkerke's R^2 for model m (Nagelkerke, 1991) is computed from the following formula:

$$\bar{R}^2_m = \frac{R^2_m}{max(R^2)} \tag{27}$$

where $max(R^2)$ is the R^2 for the baseline model (with no covariates) and

$$R^2_m = 1 - exp(l_0 - l_m)^{2/n} \tag{28}$$

where l_0 refers to the log–likelihood of no covariates model, while l_m refers to the log–likelihood of model with covariates.

Nagelkerke's R^2 allows to compare nested models for computing effect size measure. Thus, the effect size associated to uniform DIF is computed comparing R^2 for the model 0 and R^2 for the models 1, where model 0 and 1 are, respectively, the baseline conditional logistic model (with only total score as covariate) and uniform conditional logistic model (adding group variable). The effect size associated to nonuniform DIF is computed comparing R^2 for the model 1 and R^2 for model 2, where model 2 is the nonuniform conditional logistic model (adding interaction between group variable and total score). Finally, it is possible to compute effect size associated with simultaneously uniform and

nonuniform DIF as the difference between the R^2 of model 0 and the R^2 of model 2 (Gómez-Benito, Hidalgo, et al., 2009; Jodoin and Gierl, 2001). This work treats only case in which uniform DIF is present, so the effect size measure is computed as follows:

$$\Delta R^2 = |R_0^2 - R_1^2|$$

As for traditional DIF detection analysis, when you use DIF size measure it is a need to chose thresholds for interpreting the size of DIF: small, medium, and large effect sizes Cohen (1988). Here, the chosen criterion is the one proposed by Gierl and McEwen (1998) for traditional logistic regression, more conservative than that proposed by Zumbo and Thomas (1997):

- Large DIF (class C) $R^2 \geq 0.07$ and χ^2 test significant;
- Moderate DIF (class B) $0.035 \leq R^2 < 0.07$ and χ^2 test significant;
- Small (negligible) DIF (class A) $R^2 < 0.035$ or χ^2 test non significant.

4. A simulation study

This section presents simulations analysis results. Before presenting simulation analysis deeper, section 4.1 provides a check of simulated data. Especially, simulated scenarios must be constructed in such a way as to have reference and focal groups imbalanced with respect to covariates. Secondly, section 4.2 presents matching analysis for all generated scenarios. After matching, the core of results (section 4.3) presents analysis of false alarm rate and power for each scenario previously described. The last part of the chapter, section 4.4, involves a brief assessment to using effect size measure for conditional logistic regression methods applied to DIF detection analysis.

4.1 Checking scenarios

Before entering deeply into simulation results, here, some analysis are presented in order to check the simulated scenarios (e.g., the performance of simulation design). As said in section 3.2, the first step of simulation design has been to create scenarios with covariates distributions consistent with respect to INVALSI sample. As "*benchmark*" for simulations, three datasets are created with three different numbers of observations.

Table 6 represents the covariates distributions for INVALSI sample, scenarios with 500, 1000 and 2000 simulated test takers. Covariates distributions in simulated datasets are very close to covariates distributions of INVALSI sample. One half of observations (49%) are males and only about one in ten is not Italian (9%). Pupils and *simulèe* with university degree aspiration (high aspiration) are 60%. Almost half of them (46%) comes from north Italy, while 20% and 34% of them comes, respectively, from middle and south Italy. Reminding that ESCS index was coded into discrete variable using 3 cut off (e.g., I quartile, median and III quartile) we expect equidistribution in the new four categories of ESCS variable. Nevertheless, from the table it emerges that our expectations have not been perfectly reached: *III quantile* is under–represented while *IV quantile* is over–represented. This happens, probably, because of previous data cleaning: routine for DIF detection analysis in software R requires no missing data[26]. Another possible explanation could be

[26] difR package allows missing value for response variables but not for grouping variable (Magis, Beland, et al., 2010).

linked to rounding activities.

Table 6: Composition of variables among INVALSI sample and simulations.

		INVALSI sample	N=500	N=1000	N=2000
Gender					
	Male (%)	0.49	0.49	0.49	0.49
	Female (%)	0.51	0.51	0.51	0.51
Citizen					
	Italian (%)	0.91	0.91	0.91	0.91
	Not Italian (%)	0.09	0.09	0.09	0.09
Aspiration					
	University degree (%)	0.60	0.61	0.61	0.61
	Not University degree (%)	0.40	0.39	0.39	0.39
Geographic Area					
	North (%)	0.46	0.46	0.46	0.46
	Middle (%)	0.20	0.21	0.21	0.21
	South (%)	0.34	0.33	0.33	0.33
ESCS					
	I quantile (%)	0.26	0.24	0.25	0.25
	II quantile (%)	0.25	0.25	0.25	0.25
	III quantile (%)	0.22	0.22	0.21	0.22
	IV quantile (%)	0.27	0.29	0.29	0.28

Economic, social, and cultural background is affected by the other considered variables. If gender should not affect ESCS index, Italians and foreigners are characterized by a strong disadvantage for the latter in socio–economic–cultural status[27], especially in term of poverty and deprivation (Berti et al., 2014). This gap (in particular economic gap) not only is persistent but it seems to increase over the years (Gambacorta, 2017). Hence, we expect to find Italians at the top of ESCS index distribution and, conversely, not Italians at the bottom of ESCS index distribution. Traditionally, there exists gap between north–middle and south Italy in different life spheres. South is characterized by lower levels of principal socio–economic–cultural status indicators than north and middle: parental instruction, perceived wellbeing, health, income (D'Alessio, 2017). Our expectation is to find ESCS

[27]Keep in mind that this work does not consider mixed parentage pupils who have similar behavior to natives (Azzolini, Schnell, et al., 2012).

index equistribution among pupils from north and middle, while more southern students at the bottom of ESCS index distribution. Socio–economic–cultural status plays an important role in the formulation of educational aspirations of students. Low and high pupils socio–economic–cultural status impacts significantly in their academic aspiration (Salgotra and Roma, 2018). Parents' high social capital has an impact on academic aspiration of their children: these families transmit to their children high educational expectations which turn into academic aspiration (Shahidul et al., 2015). Therefore, we expect to find more pupils with high aspiration at the top of ESCS index distribution and, conversely, more pupils with low aspiration at the bottom of ESCS index distribution.

We provide an additional check in appendix A. Table A4, table A5, table A6 and table A7 show ESCS index distribution across other covariates for, respectively, INVALSI sample, simulations with 500, 1000 and 2000 sample size. Gender, aspiration and geographic area do not show problems: the distribution of ESCS index across other covariates is quite consistent with respect to INVALSI sample. Especially, pupils with higher values of ESCS index have higher aspiration; conversely, pupils with lower values of ESCS index have lower aspiration. In addition, southern pupils are more concentrated into lower quartiles of ESCS index, while students from north and middle are most equidistributed across quartiles. If citizen variable does not create problem for Italians, not Italians are over–represented into *IV quartile* and under–represented into *I quartile* among generated datasets, especially, for sample size of 500, probably due to the small number of not Italians. Nevertheless, it appears that the covariates were generated similar to the INVALSI sample, reflecting literature results.

The second step of simulation design concerns the generation of pupils' ability θ. Figure 2 represents kernel density of latent trait for INVALSI sample, sample size of 500, 1000 and 2000 units. Densities overlap well in the tails, while they does not overlap well around the mean. Latent trait for INVALSI sample has a mean of 0.293 and a standard deviation of 1.227. Generated θs have a mean of 0.254, 0.275, 0.272 and a standard deviation of 1.149, 1.119, 1.151, respectively, for sample size of 500, 1000 and 2000 units. In conclusion, simulated latent traits are satisfactory.

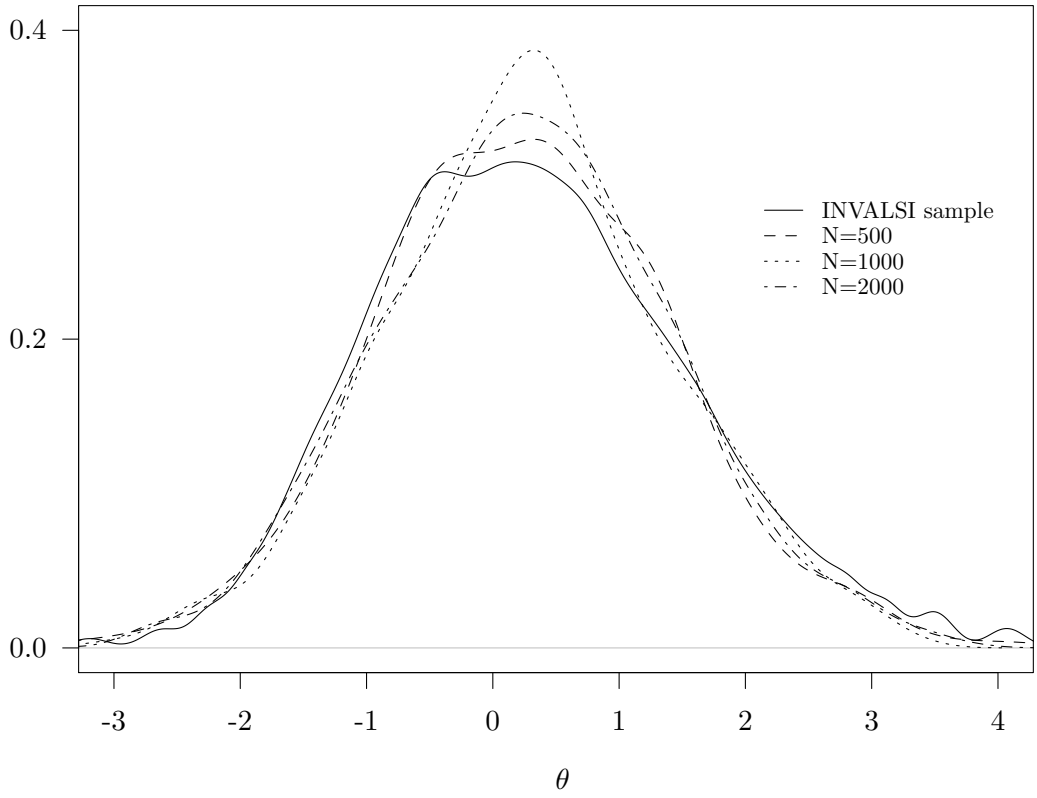


Figure 2: Kernel distributions of latent traits.

Section 3.1.2 has provided analysis for balancing check in INVALSI sample: it presents one large size (*aspiration*), two medium sizes (*gender* and *ESCS*), one small size (*citizen*) and one negligible (*geographic area*) imbalance. Generated grouping variables should be imbalanced by construction. Hence, we expect to find reference (academic track) and focal groups (technical track) imbalanced with respect to covariates, according to INVALSI sample. Here, we provide a check in order to assess whether this variables have been properly generated. Table 7 presents balancing analysis for INVALSI sample (already present in table 3) and simulation with 500, 1000 and 2000 test takers.

First of all, it is possible to verify the grouping variables distribution. INVALSI sample presents about 57% (14185/25058, where the denominator is the INVALSI sample size) of pupils enrolled in academic tracks and about 43% (10873/25058) of them enrolled in technical schools. If grouping variables are consistent with respect to INVALSI sample for

1000 and 2000 sample size, there is one percentage point of difference for 500 sample size. Therefore, proportions of grouping variables are respected in simulated datasets.

All situations present same patterns of INVALSI sample as regards group imbalance analysis. Scores of standardized mean differences (SMD) in simulated datasets are very close to SDM of INVALSI sample. Academic track tends to have more males, fewer not Italians, pupils with high aspiration and with high socio–economic–cultural status. All differences are statistically significant, except for geographic area. SMD columns show that simulated datasets present one large size imbalance (*aspiration*), two medium sizes (*gender* and *ESCS*), one small size (*citizen*) as for INVALSI sample. SMD scores vary across datasets and better INVALSI sample reproduction grows up with sample size.

4.2 Propensity score matching

In previous section we checked whether groups are imbalanced with respect to covariates in all simulated datasets. Groups are imbalanced by construction and SDM analysis confirmed it. The new methodology suggests (section 2.4) to match groups in situations of imbalanced groups. Matching reduces selection bias and it should guarantee the attribution of DIF identification to group allocation mechanism, controlling for other confounding variables. Therefore, now, we provide propensity score matching analysis in order to balance groups for the next DIF detection analysis.

As said in section 3.3.2, we use both greedy and optimal full matching. All matching analysis are carried out with *MatchIt* package of software R (Ho et al., 2011). Greedy (nearest neighbor) matching is an algorithm that matches one treated and one control unit with closest value of propensity score. Here, we opt for using replacement option because it outperforms without replacement option. Unfortunately, nearest neighbor matching could lead a sample reduction, due to discarding of unmatched units. Differently, optimal matching adopts an algorithm which minimizes the total differences in the estimated propensity score (Austin, 2009b) among treated and untreated subjects. Here, we opt for full option that allows to avoid sample reduction. In addition, we try both one–to–many option and a combination with one–to–many and many–to–one option. Nevertheless, this section presents only tables and figures referred to optimal full matching with a combination of one–to–many and many–to–one because it performs better than nearest

Table 7: Covariates balancing between groups in INVALSI sample, N=500, N=1000, N=2000.

	INVALSI sample				N=500			
	Technical	Academic	p	SMD	Technical	Academic	p	SMD
n	10873	14185			208	292		
Gender								
Male (%)	66.0	38.6	<0.001	0.571	62.5	39.4	<0.001	0.475
Citizen								
Not Italian (%)	11.4	7.0	<0.001	0.154	12.5	6.5	0.032	0.205
Aspiration								
University degree (%)	34.1	79.7	<0.001	1.036	35.1	79.5	<0.001	1.003
Area (%)			0.003	0.043			0.589	0.094
North	45.5	46.9			43.3	47.9		
Middle	19.1	19.7			22.1	20.2		
South	35.5	33.4			34.6	31.8		
Escs (%)			<0.001	0.533			<0.001	0.626
I quartile	39.5	21.9			40.9	17.5		
II quartile	26.8	22.5			26.0	24.0		
III quartile	20.7	24.3			18.8	25.3		
IV quartile	13.0	31.2			14.4	33.2		
	N=1000				N=2000			
	Technical	Academic	p	SMD	Technical	Academic	p	SMD
n	427	573			854	1146		
Gender								
Male (%)	66.0	36.3	<0.001	0.623	62.8	38.7	<0.001	0.495
Citizen								
Not Italian (%)	11.2	7.3	0.043	0.135	12.1	6.7	<0.001	0.184
Aspiration								
University degree (%)	34.2	81.0	<0.001	1.075	34.8	80.5	<0.001	1.045
Area (%)			0.726	0.051			0.558	0.049
North	44.7	46.9			47.3	45.0		
Middle	20.0	20.2			20.8	21.1		
South	33.3	32.8			31.9	33.9		
Escs (%)			<0.001	0.444			<0.001	0.529
I quartile	37.7	21.6			40.0	20.9		
II quartile	26.0	23.2			25.5	23.6		
III quartile	20.1	24.6			20.4	24.2		
IV quartile	16.2	30.5			14.1	31.4		

neighbor matching and full matching[28].

Tables 8, 9 and 10 represent full matching analysis with a combination of one–to–many and many–to–one for all three generated samples. In particular, they present mean covariates distributions before matching for reference and focal groups (column 2 and 3), mean differences between them (column 4), mean covariates distributions after matching for focal group (column 5) and mean differences between columns 2 and 5. The last column is the most interesting for matching analysis: the Percentage of Bias Reduction (PBR). It represents how much bias reduction is driven by the matching.

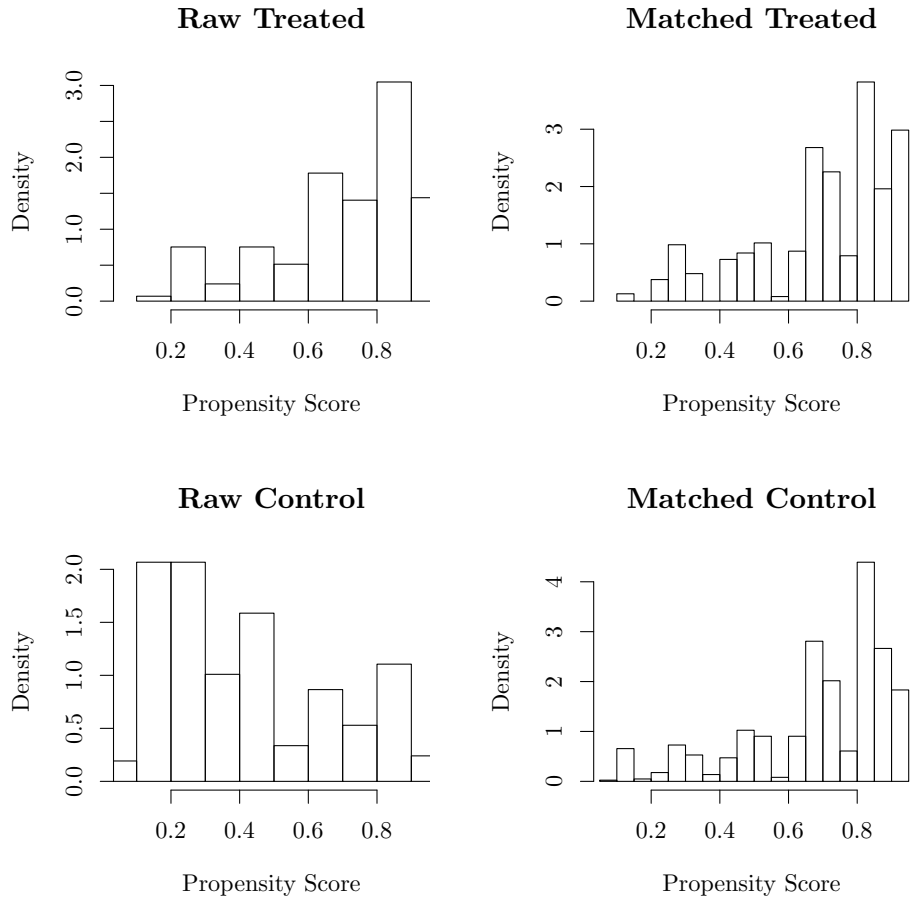


Figure 3: Propensity score distributions before and after matching using full matching with a combination of one–to–many and many–to–one (N=500).

For sample with 500 units we set an upper restriction of 5 on the maximum units for reference and focal group. From the table 8, it emerges that matching leads a good bias reduction, especially, comparing it to values of PBR with nearest neighbor and full

[28]In appendix B you can find nearest neighbor matching and full matching analysis.

Table 8: Percentage of Bias Reduction (PBR) using full matching with a combination of one–to–many and many–to–one (N=500).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.701	0.420	0.281	0.701	0.000	97.0
Gender (Male)	0.394	0.625	-0.231	0.426	-0.033	80.2
Citizen (Not Italian)	0.065	0.125	0.075	0.038	-0.001	83.0
Aspiration (High)	0.794	0.351	0.444	0.797	-0.003	99.4
Middle	0.202	0.221	-0.019	0.164	0.056	94.3
South	0.318	0.346	0.028	0.359	-0.041	36.2
II quartile	0.239	0.259	-0.020	0.221	0.019	83.6
III quartile	0.253	0.187	0.066	0.265	-0.011	38.4
IV quartile	0.332	0.144	0.188	0.356	0.024	73.0

matching, respectively, in table B1 and B2 in appendix B. Most of the variables shows high bias reduction that swings between 73.0% (*IV quartile*) and 99.4% (*aspiration*). Only *middle* and *III quartile* present a low bias reduction, 36.2% and 38.4%. In addition, figure 3 presents propensity score distributions before and after matching between reference and focal groups. In the absence of bias selection, the two distributions overlap. We can see that propensity score distribution of focal (control) group are very close to propensity score distribution of reference (treated) group after matching.

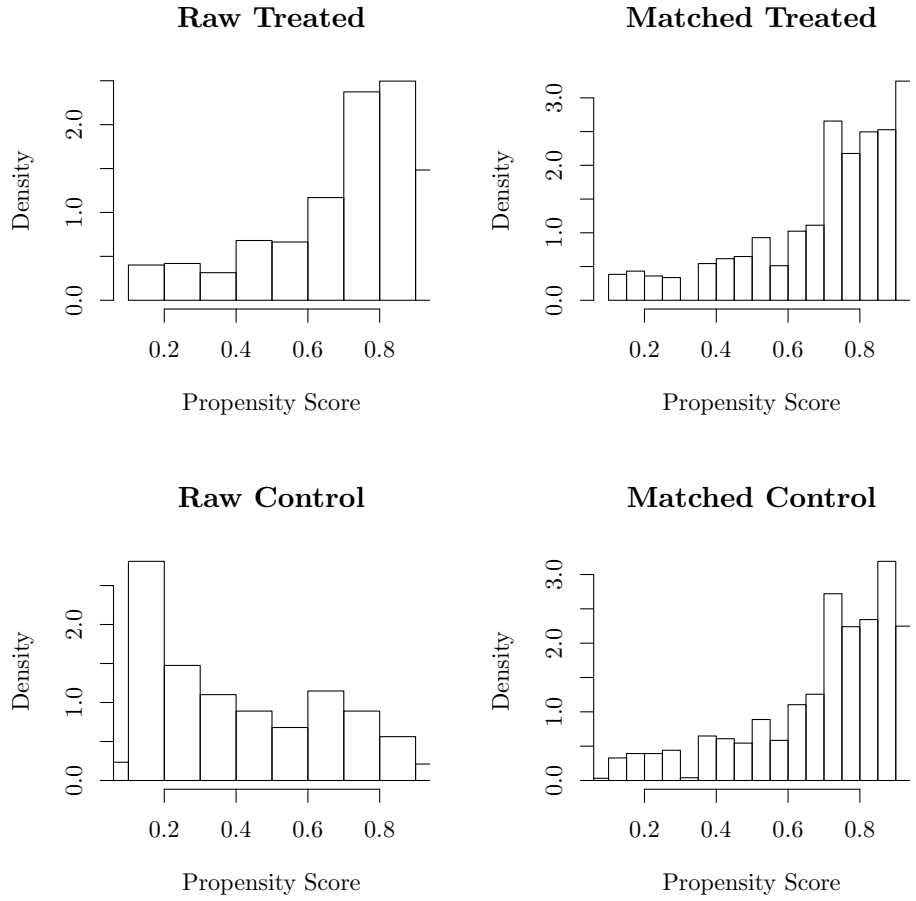


Figure 4: Propensity score distributions before and after matching using full matching with a combination of one–to–many and many–to–one (N=1000).

Full matching with a combination of one–to–many and many–to–one for sample of 1000 units is restricted to having the maximum of 7 treated and 7 control units. Here, the percentage of bias reduction (table 9) is better than bias reduction of greedy matching and full matching, respectively, in table B3 and B4 in appendix B. PBR swings between 56.5% (*III quartile*) and 100.0% (*aspiration*), suggesting the goodness of matching. One

Table 9: Percentage of Bias Reduction (PBR) using full matching with a combination of one–to–many and many–to–one (N=1000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.698	0.406	0.292	0.696	0.002	99.4
Gender (Male)	0.363	0.660	-0.297	0.372	0.009	97.1
Citizen (Not Italian)	0.073	0.112	-0.039	0.068	0.005	87.7
Aspiration (High)	0.809	0.342	0.468	0.809	0.000	100.0
Middle	0.202	0.220	-0.018	0.203	-0.001	97.7
South	0.328	0.333	-0.004	0.333	-0.004	-1.3
II quartile	0.232	0.260	-0.028	0.235	-0.003	88.2
III quartile	0.246	0.201	0.045	0.265	-0.019	56.5
IV quartile	0.305	0.162	0.144	0.283	0.023	84.1

covariate, *south*, presents PBR value close to zero, indicating no difference after matching. Furthermore, figure 4 analysis allows to judge positively the bias reduction driven by the matching.

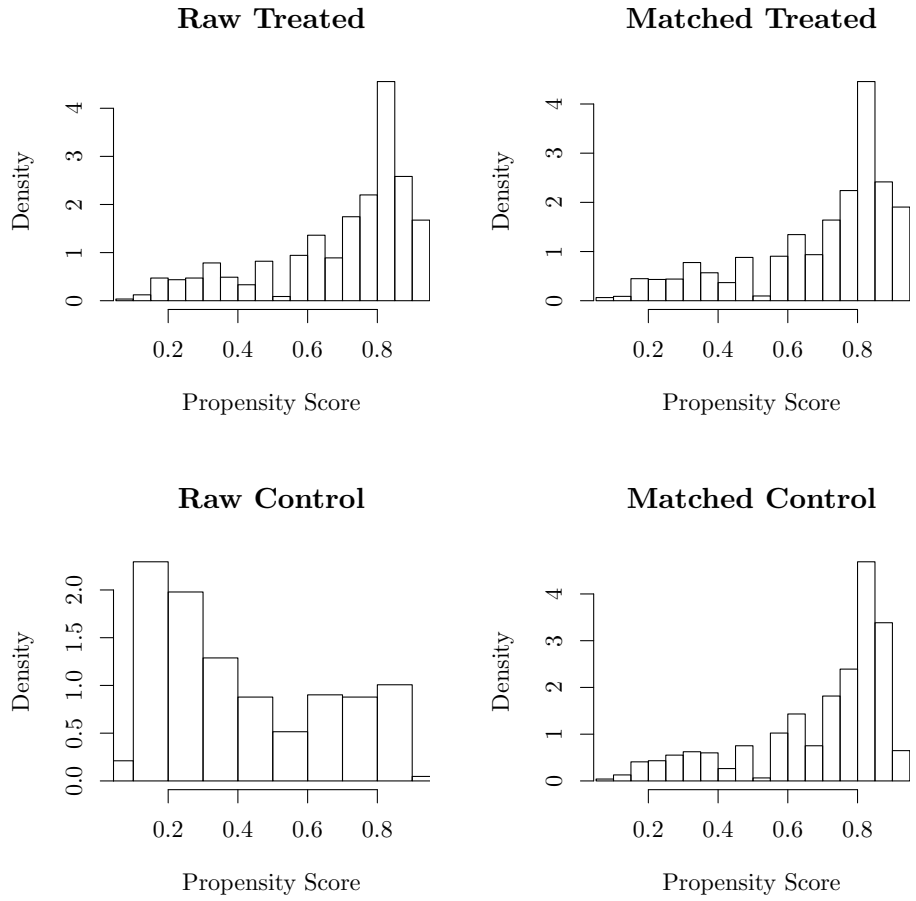


Figure 5: Propensity score distributions before and after matching using full many–to–one and one–to–many matching (N=2000).

Table 10 presents PBR in dataset with 2000 units. Here, we set an upper restriction of 10 on the maximum treated and control units. Bias reduction is satisfactory and it swings between 99.0% (*aspiration*) and 78.8% (*citizen*). It is not satisfactory only for *geographic area* variable. If for *south* we find low bias reduction (10.9%), *middle* presents high and negative value of PBR. This indicates that matching produces larger differences in *middle* variable between the two groups. Nevertheless, this value is smaller than value for other considered matching techniques (table B5 and B6 in appendix B). Once again, figure 5 confirms a good matching. To sum up, matching leads high bias reduction for all generated datasets: after matching reference and focal groups are balanced with respect to covariates.

Table 10: Percentage of Bias Reduction (PBR) using full matching with a combination of one–to–many and many–to–one (N=2000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.690	0.416	0.274	0.690	0.000	99.9
Gender (Male)	0.387	0.628	-0.240	0.387	0.009	96.2
Citizen (Not Italian)	0.067	0.121	-0.053	0.057	0.011	78.8
Aspiration (High)	0.805	0.348	0.458	0.801	0.005	99.0
Middle	0.211	0.208	0.002	0.191	0.020	-623.0
South	0.339	0.318	0.020	0.321	0.018	10.9
II quartile	0.236	0.255	-0.020	0.237	-0.001	92.7
III quartile	0.242	0.204	0.038	0.247	-0.006	84.8
IV quartile	0.314	0.140	0.174	0.300	0.014	91.7

The subsets created by full matching with a combination of many–to–one and one–to–many are used for the next simulation results.

4.3 Results

Simulation analysis involves false alarm rates and power of DIF methods. We mainly focus on the new methodology. We apply conditional logistic regression with the best matching previously checked: full matching with a combination of many–to–one and one–to–many. In addition, we compare its performance to the performance of traditional DIF methods. We consider Mantel–Heanszel statistic, Lord's χ^2 and conventional logistic regression as traditional DIF methods, described in section 2.3.1.

All figures below present the performance for DIF methods simultaneously, controlling for all manipulated factor[29]. In particular, in all graphs the number of test takers are represented on the x–axis, while y–axis represents false positive and true positive (power) rates. In addition, tables C1– C9, in appendix C, give the point estimates for false alarm rates and power for $\beta \sim N(0, 1)$. For false alarm rates we set a nominal alpha level to 0.05 (the red line in the graphs), that is, we tolerate that at most 5% of false positive detected may be due to chance.

4.3.1 Type I error inflation

First of all, we analyze and comment simulation results with βs drawn from a standard normal distribution (Tutz and Berger, 2016; Berger and Tutz, 2016; Magis, Tuerlinckx, et al., 2015; Khalid and Glass, 2013; Magis, Raîche, et al., 2011). When we control for no biased items (figure 6) all methods perform satisfactorily: they present false alarm rates under or very close to nominal alpha level. In addition, all methods performances are perfectly under 5% when test contains large number of items (J=60).

Introducing 10% biased items with moderate DIF size (figure 7), false positives identified by all methods tend to increase compared to no bias scenarios. Nevertheless, they are still under or very close to nominal alpha level. Once again, all methods perform better in situations in which large number of items forms the test.

Now, if we double DIF size (figure 8), we find that general false alarm rates increase. Furthermore, as sample size increases false positives detected increase. The new method-

[29] See section 3.2.5 for a detailed description.

ology and Lord's χ^2 outperform MH statistic and LR, especially for large sample size. Nevertheless, all methods exceed nominal alpha level for large samples. When we control for 20% of biased items and moderate DIF size (figure 9) we can comment similar to scenarios of figure 8.

Finally, figure 10 represents performances for 20% of biased items with large DIF size. Here, general false alarm rates increase dramatically and as sample size increases false positives detected increase. Furthermore, the new methodology outperforms traditional DIF detection methods for large sample size. Here, it seems to be no significant differences among different test lengths.

To sum up, Zumbo's methodology outperforms, in terms of false alarm rates, traditional DIF detection methods for large sample size. In addition, this result is more evident in situations of moderate number of biased items and large DIF size. One possible explanation for association between large sample and inflated type I error rate is the nature of statistical test used. All considered methods are based on a statistic test which is approximately distributed as a χ^2. This statistic is sensitive to large sample sizes (Jodoin and Gierl, 2001), amplifying the inflated type I error rate.

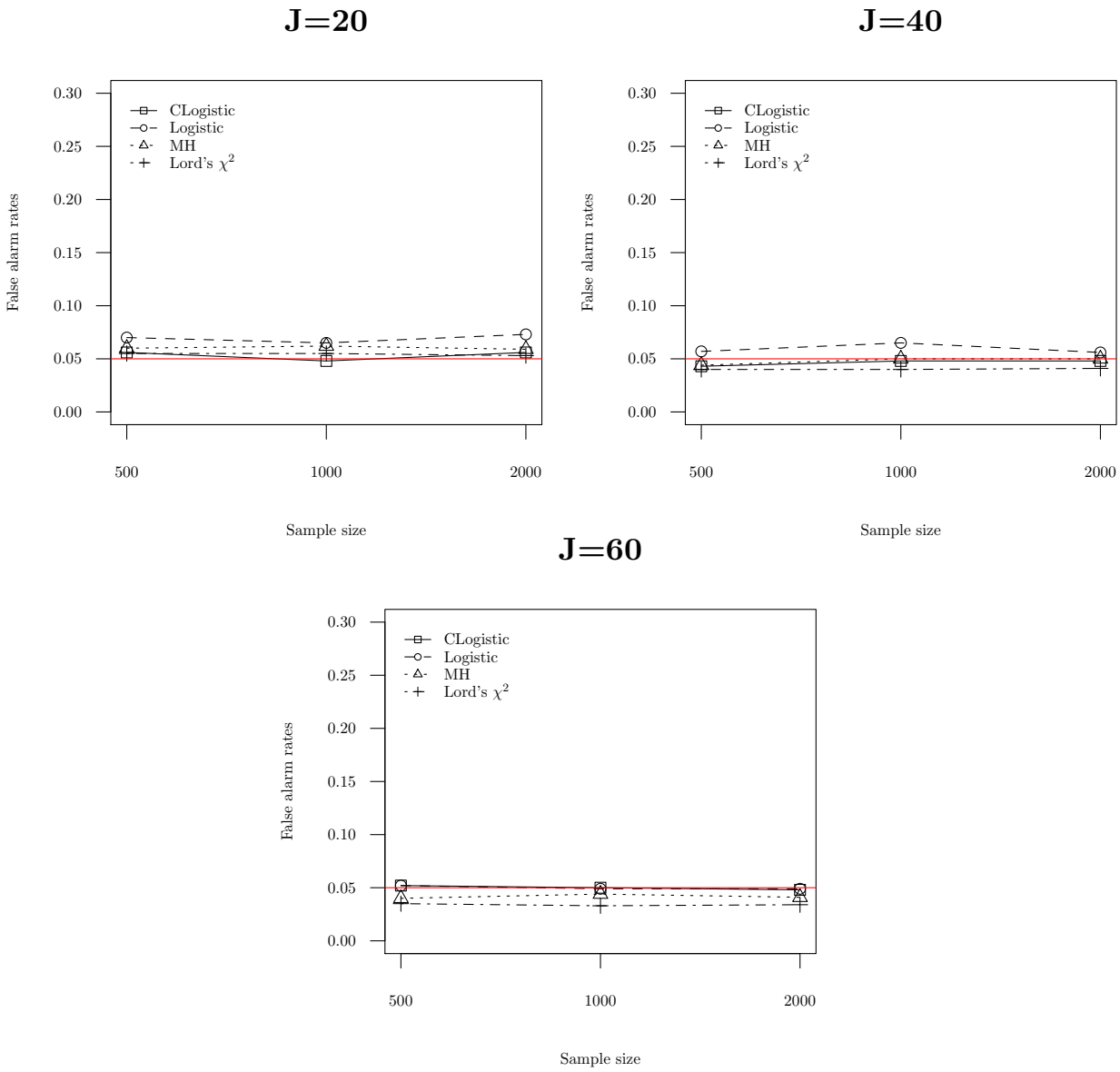


Figure 6: False alarm rates of DIF methods: no biased items.

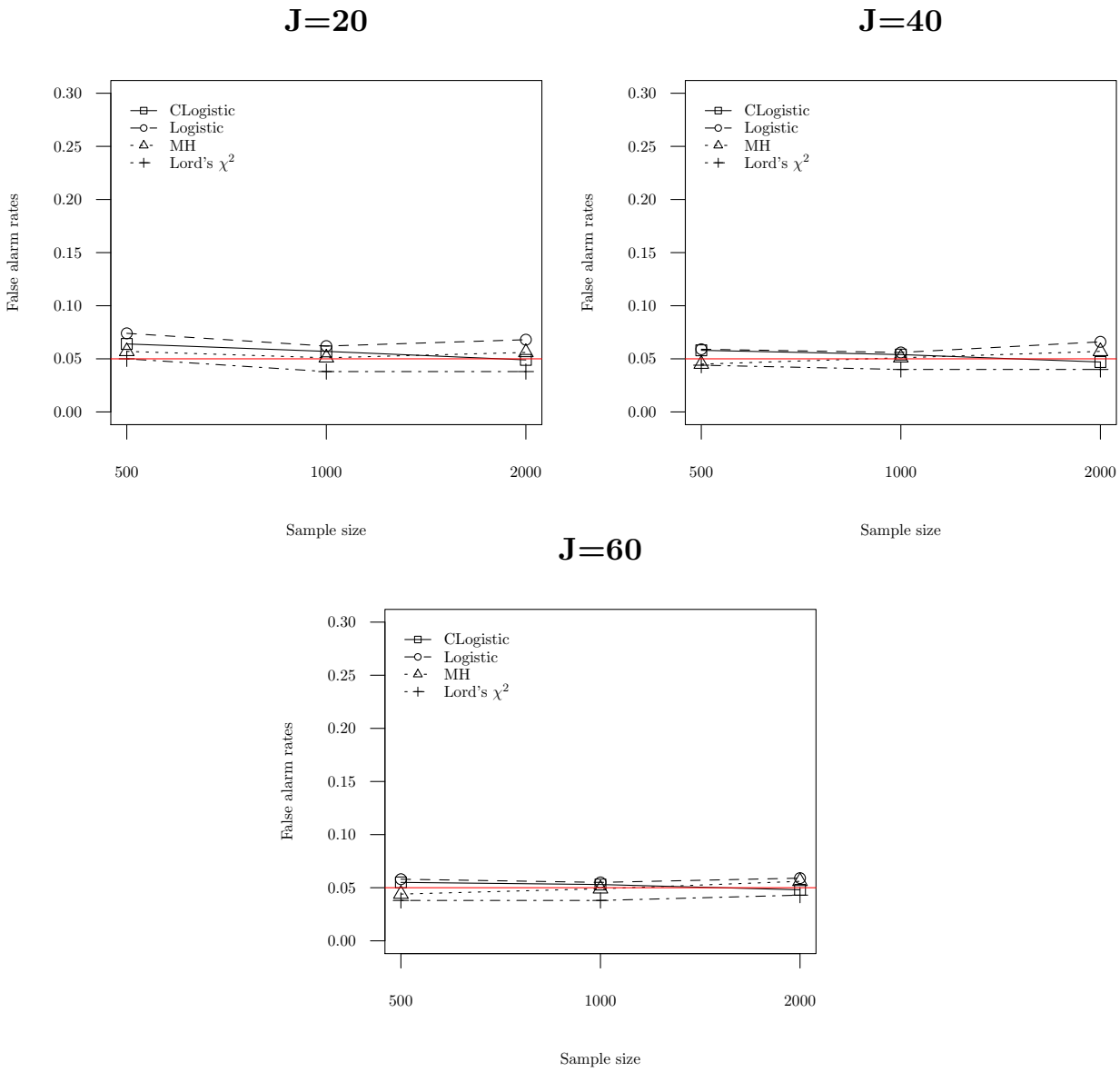


Figure 7: False alarm rates of DIF methods: 10% biased items and δ=0.4.

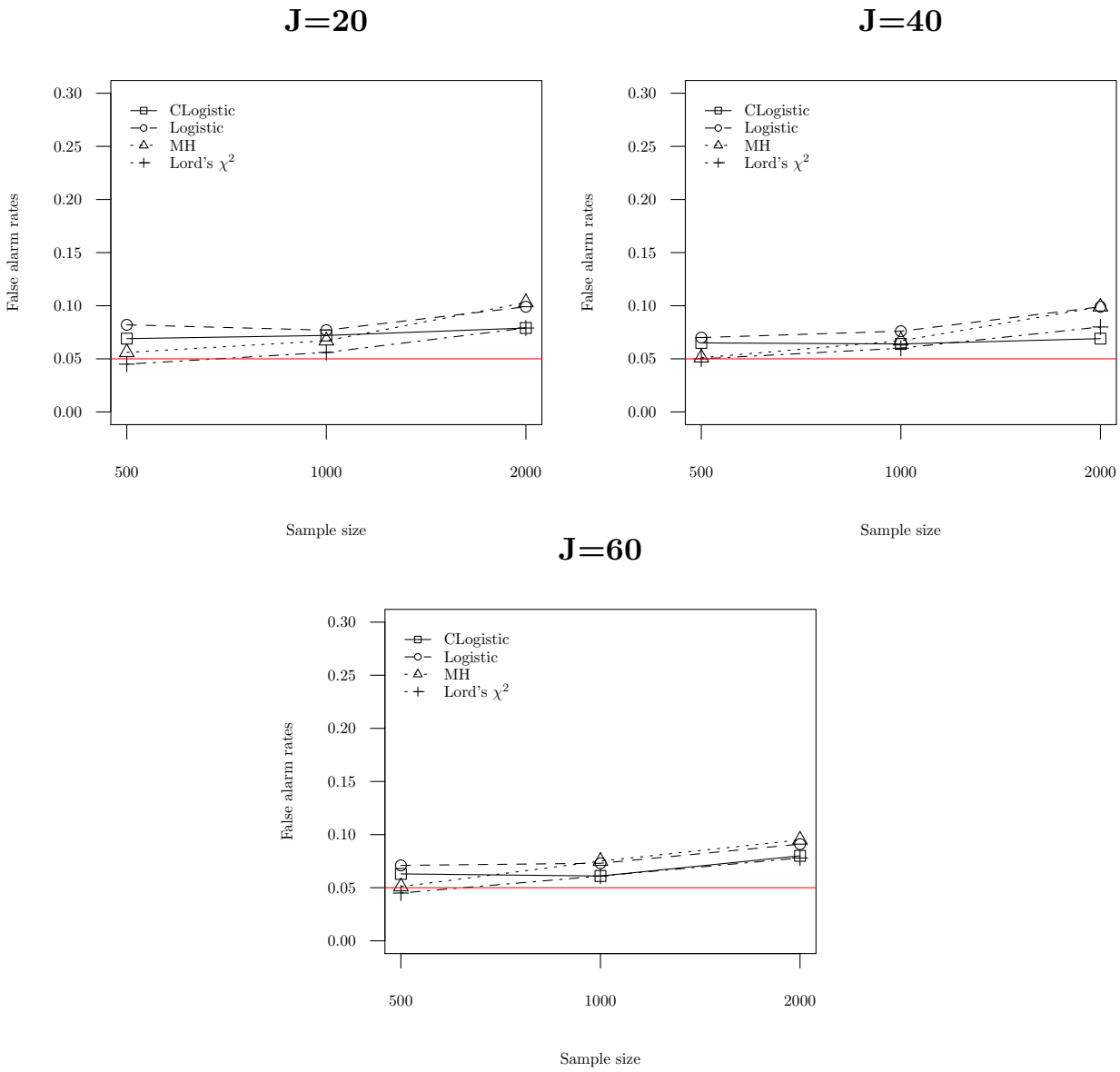


Figure 8: False alarm rates of DIF methods: 10% biased items and δ=0.8.

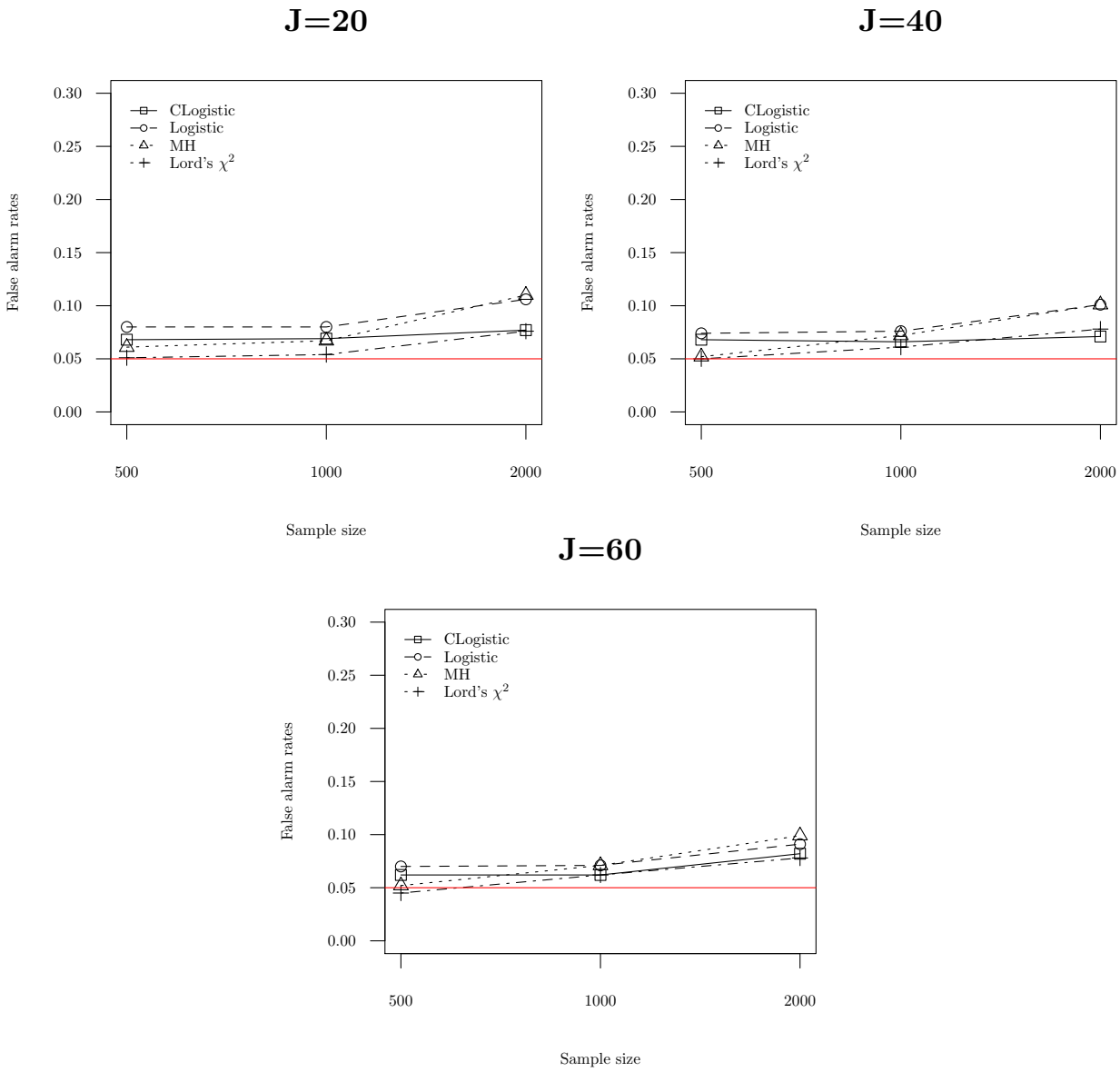


Figure 9: False alarm rates of DIF methods: 20% biased items and δ=0.4.

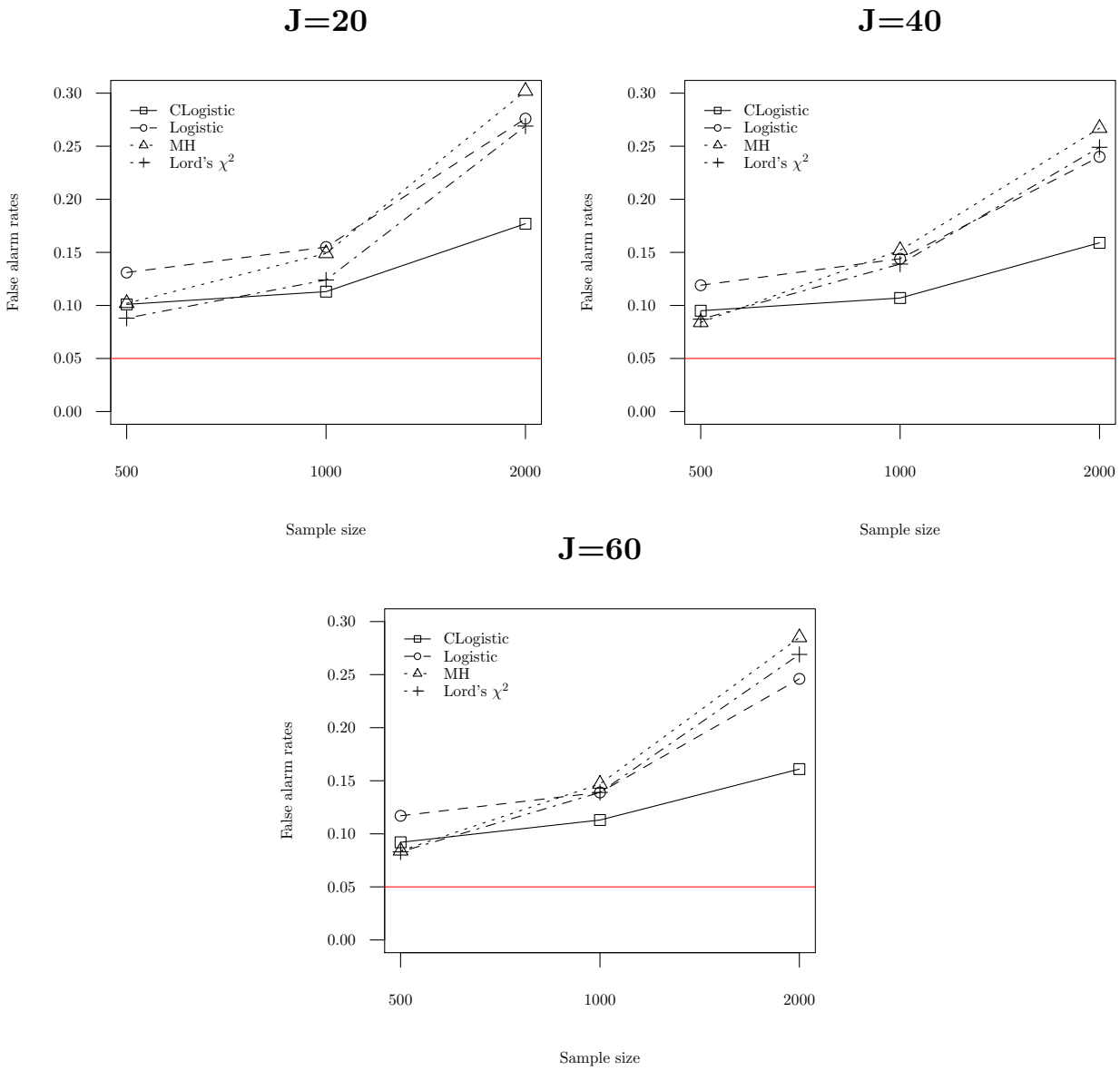


Figure 10: False alarm rates of DIF methods: 20% biased items and δ=0.8.

4.3.2 Power rates

Now, we shift the attention towards test power linked to DIF methods. Of course, power analysis does not include situations with no biased items. For moderate DIF size (figures 11 and 13), as number of test takers increases, true positives detected increase for all methods. The global performance are not very satisfactory for small size, while it increases with larger sample size. In addition, the new methodology underperforms traditional DIF detection methods. Across test length, there seems to be no evident pattern: test length does not impact the power rates.

Differently, for large DIF size (figures 12 and 14) all methods present satisfactory trends, with a constant increase of true positives detected as sample size increases; only MH statistic presents fluctuating trends. These results are not surprising because methods should identify better a true positive when DIF size is large. Once again, number of items does not significantly impact all performances. Furthermore, here, if the new methodology presents lower power rates for small sample size, it tends to detect perfectly true positives for large sample size[30].

[30]This perfect detection is observed also for Lord's χ^2 and LR.

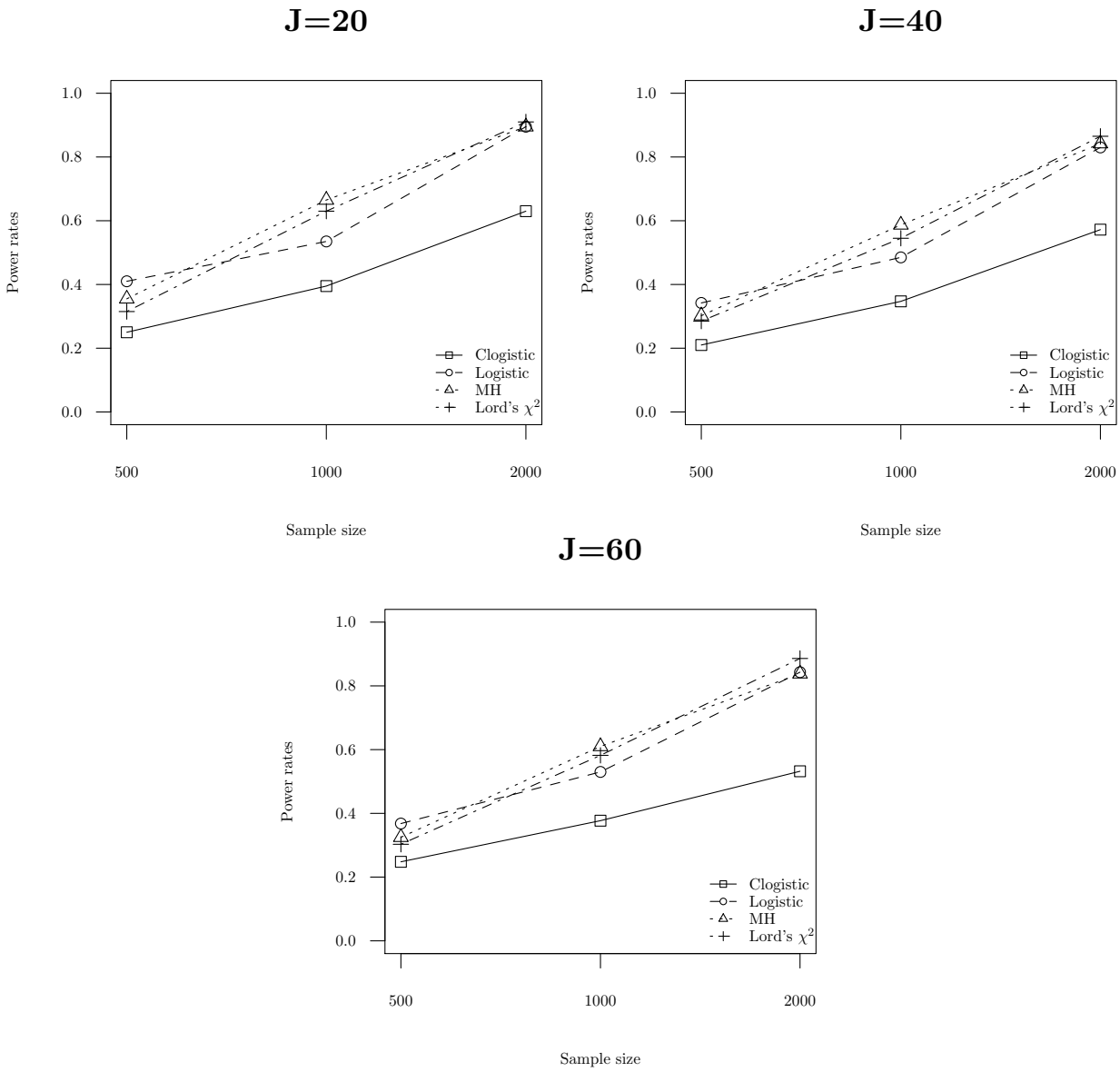


Figure 11: Power rates of DIF methods: 10% biased items and δ=0.4.

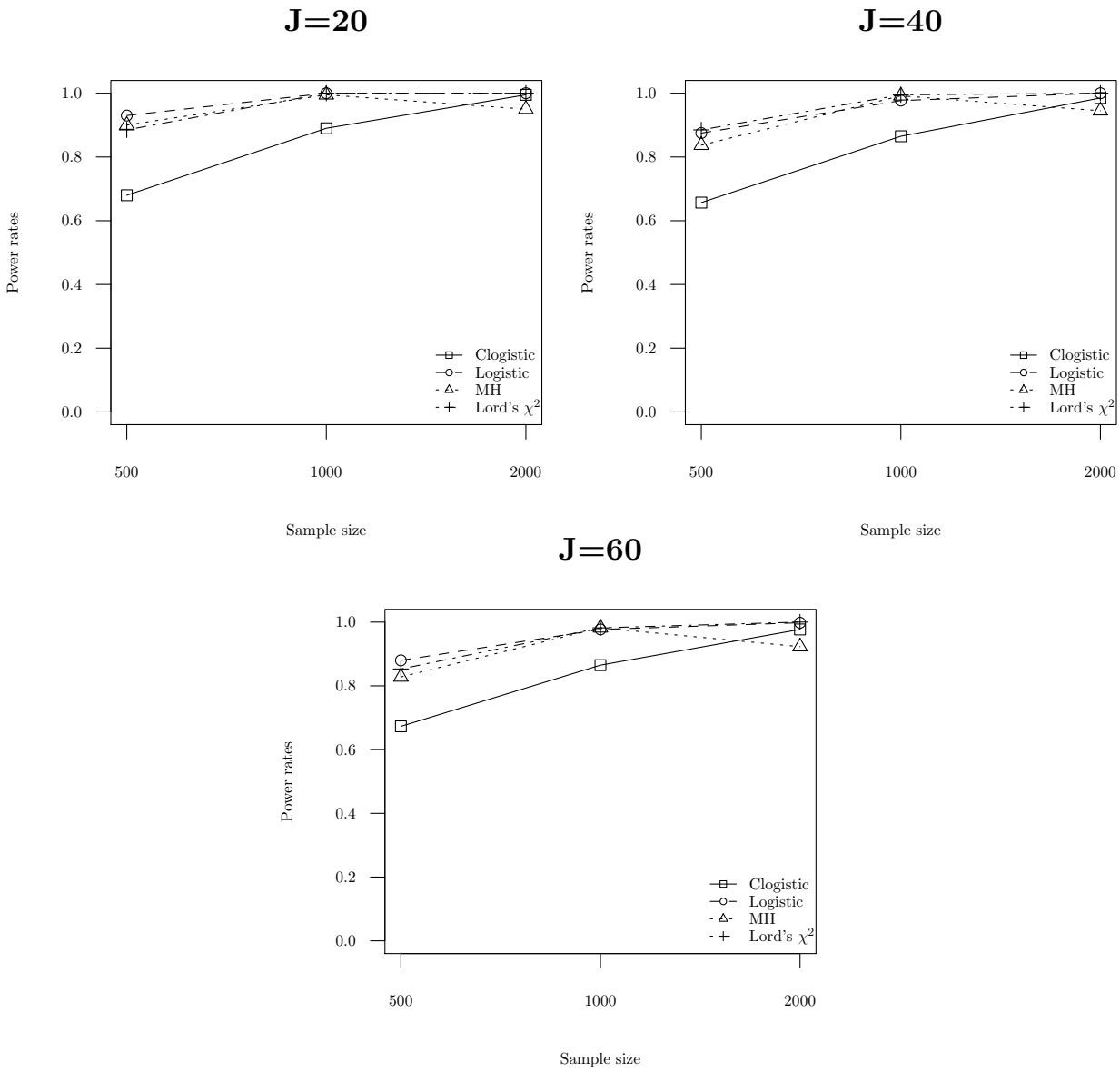


Figure 12: Power rates of DIF methods: 10% biased items and δ=0.8.

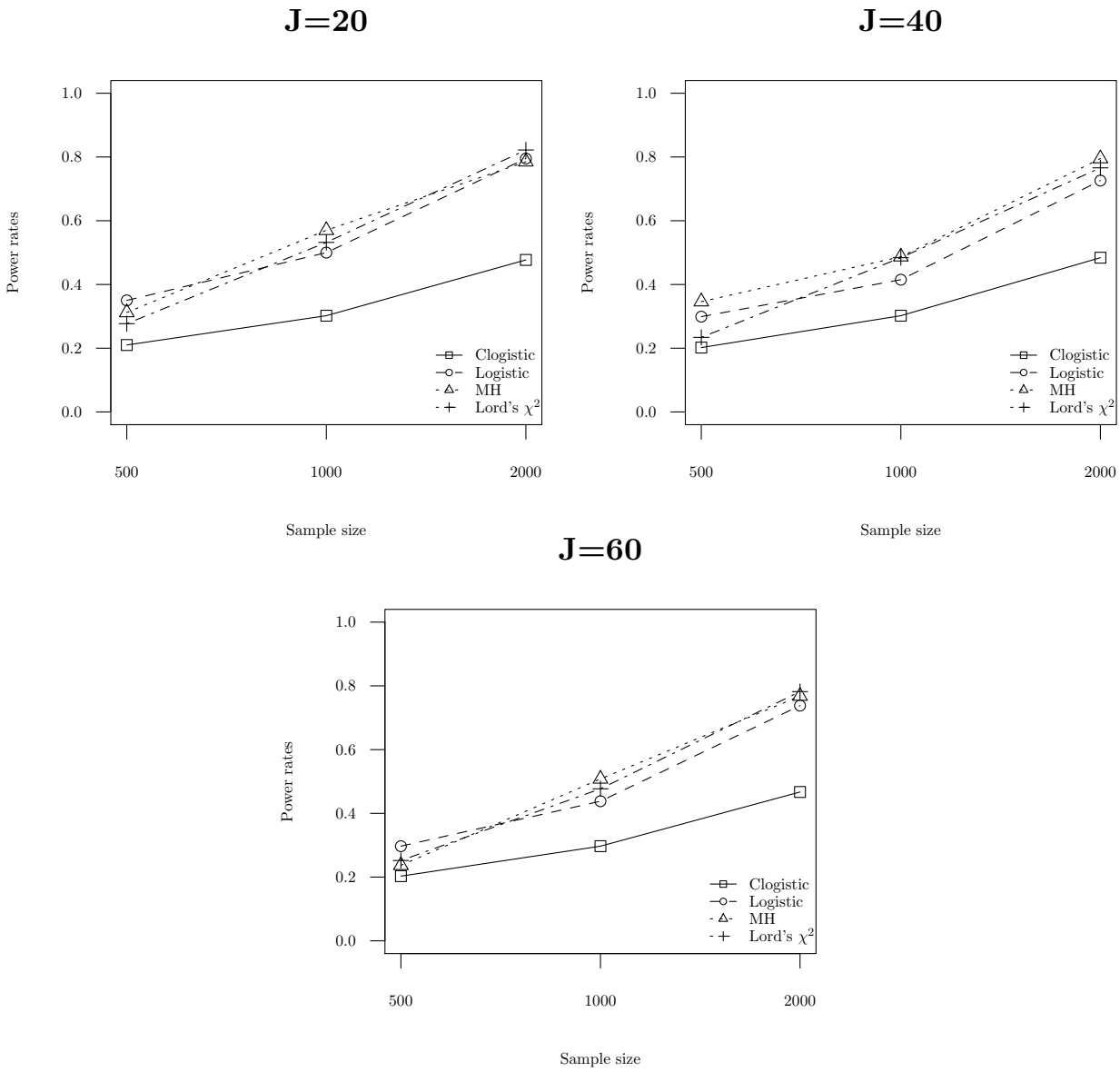


Figure 13: Power rates of DIF methods: 20% biased items and δ=0.4.

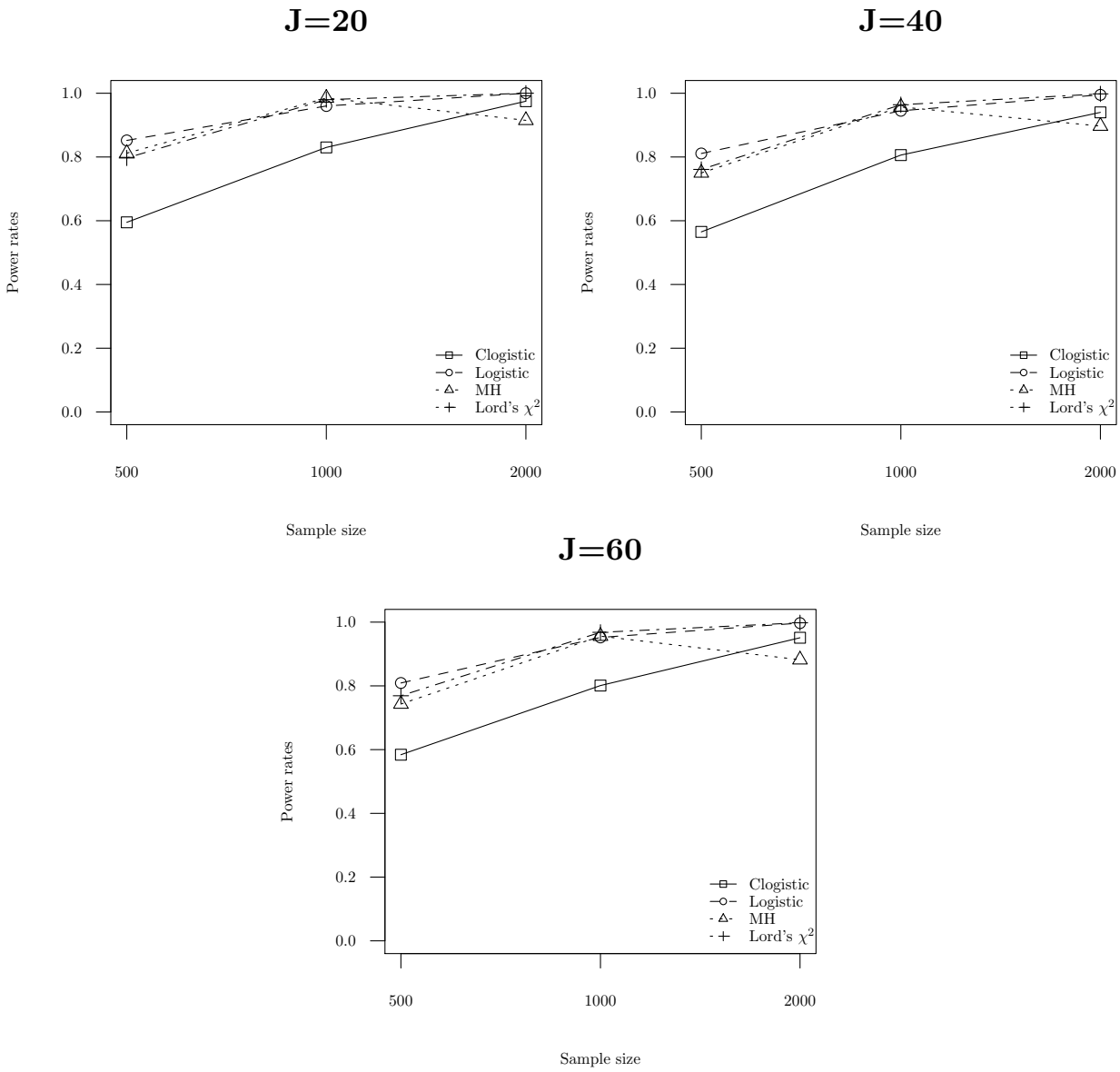


Figure 14: Power rates of DIF methods: 20% biased items and δ=0.8.

To sum up, simulation results show that the new methodology outperforms, in several situations, traditional DIF detection methods in imbalanced groups. In particular, it presents the best performance for false alarm rates in situations of large number of DIF items, large DIF size and large number of test takers. Nevertheless, it has some disadvantages to correctly detect an item as DIF, especially for small samples. However, this disadvantage disappears for large DIF size and large sample size. Therefore, if you choose to apply the new methodology, you must be aware there exists a trade–off of false positive and true positive detected when DIF is present: low type I error inflation involves

low power rates. Nevertheless, we pay more attention to false positive inflation rather than true positive identification. It is "dangerous" that an item is mistakenly flagged as DIF. Traditionally, if an item is flagged as DIF, it is assessed from an expert *equipe* and, subsequently, it is deleted from analysis. In this case, test mistakenly loses useful information linked to the item in the analysis.

Previous simulation studies have shown that false alarm rates and power rates increase for conventional logistic regression as the sample sizes increase (Jodoin and Gierl, 2001; Narayanon and Swaminathan, 1996; Swaminathan and Rogers, 1990). MH statistic also depends on sample size (Narayanon and Swaminathan, 1996). In addition, it is shown that 20 items tests present more inflation of type I error rates than one for the 40 items[31] (Uttaro and Millsap, 1994). Lord's χ^2 performance are very similar to MH procedure, though MH identifies slightly more DIF items (Raju et al., 1993). In addition, conventional logistic regression generally detected more DIF items than MH statistic (Gómez-Benito, Hidalgo, et al., 2009; Hidalgo and LóPez-Pina, 2004). Finally, it was found that conventional LR and MH procedure have similar powerful in detecting uniform DIF (Swaminathan and Rogers, 1990). Our results are consistent with respect to previous results for traditional DIF detection methods. Therefore, this can be considered a robustness check of our results.

As said in section 3.2.5, we want to assume also a different distribution for difficulty parameter β. This occurs because simulating βs from standard normal distribution makes the difficulty parameters focus around zero, excluding extreme values. This strategy precludes the presence of very easy or very difficult items. Therefore, βs are also drawn from an uniform distribution with parameters set to minus 2 and plus 2 (Weiss, 2014; Magis and Facon, 2012). Uniform distribution allows to consider both very easy and very difficult items. Nevertheless, we do not comment this results because we found the same patterns and trends for situations in which βs are drawn from standard normal distribution. For details, you can find figures and tables of performance analysis with uniform assumption in appendix D.

In conclusion, when groups are imbalanced we suggest the use of new methodology for two reasons. First of all, simulation results showed that Zumbo's methodology presents the same performance in some situation (small samples and small biased items) and better

[31] Our results present the same patterns for all DIF considered procedure; anyway, 60 items test presents results similar to 20 items test.

performance in others (large samples, large biased items and large DIF size). Secondly, it is useful for detect casual effects of group allocation mechanism according to DIF analysis. Thus, the new methodology is recommended for observational studies since traditional DIF detection techniques do not present better performance in considered scenarios.

4.4 Effect size measure

Previous simulation results have demonstrated that the new methodology outperforms traditional DIF detection techniques in some situations, especially for large sample size. Nevertheless, also Zumbo's methodology presents no satisfactory type I error inflation for large sample size. That is, the false alarm rates exceed the nominal alpha level of 0.05. Therefore, here, we use an effect size measure in order to reduce I error inflation for the new methodology. In particular, we assess the performance using an effect size measure, always, in terms of false alarm rates and power rates. As in previous simulation studies about the use of effect size measure for conventional LR (Gómez-Benito, Hidalgo, et al., 2009; Jodoin and Gierl, 2001), we expect that it reduces both type I error and power rates when sample is large.

As said in section 3.3.4, using only null hypothesis significance testing for DIF detection has been criticized (Kirk, 1996; J. Cohen, 1994) because statistical test is sensitive to sample size. Therefore, using null hypothesis significance testing with an effect size measure could overcome this issue. *"Moreover, there is a broad consensus about the need to bring together the interpretation of effect size with significance tests in all types of research"* (Gómez-Benito, Hidalgo, et al., 2009, p. 24). Therefore, it is crucial to assess the performance of the effect size measure use for different scenarios.

Tables 11 and 12 present the performance of effect size measure for the new methodology across different scenarios. Table 11 shows the percentage of false positive (FP), where, DIF–U represents the FP ratio detected from the new methodology only with statistical test, while $\Delta R^2 - U$ represents the FP ratio detected using the effect size measure. Finally, the 4th, the 7th and the 10th columns present the reduction of FP using the effect size measure rather than null hypothesis significance testing[32]. Differently, table 12 shows the percentage of correct identification (CI).

[32]If the reduction is not reported, it means that or $\Delta R^2 - U$ detects zero false positive either the reduction exceeds 100%.

As threshold for DIF detection using effect size measure we use the criterion proposed by Gierl and McEwen (1998): items is flagged as no DIF if $\Delta R^2 < 0.035$ or χ^2 is non significant. We can see that effect size measure produces very large reduction of false positives. Especially, for small sample sizes the reduction occurs between about 10% and 20%. For example, in the first scenario (no biased items, J=20 and N=500) the new methodology with only statistical test identifies, in average, about 5% of FPs, while using effect size measure this percentage goes down to minus of 1% (0.30%). For moderate and large sample sizes the reduction is greater than 100% and, for many scenarios, effect size measure identifies no false positives.

In contrast, correct identification of DIF items suffers from a net reduction using effect size measure. The reduction for small simple size is quite weak, between about 2% and 9%. Here, there seems to be an effect of DIF size because the reduction is bigger for moderate DIF size than for large DIF size. For moderate sample size, the reduction oscillates between about 7% and 90% with, once again, greater reduction for moderate DIF size than for large DIF size. Finally, for large sample size the reduction is greater than 100% and, for most of scenarios, effect size measure identifies no true positives.

For uniform assumption of β (tables D10 and D11 in appendix D) the results are very similar to normal assumption. As expected, false alarm rates benefits from using effect size measure. For large samples, this allows to have no type I error inflation. Nevertheless, a reduction of type I error rates involves also a reduction in power. That is, the effect size measure guarantees to not mistakenly flag items as DIF, but it does not identify correctly an item as DIF. Therefore, you must keep in mind the existence of this trade–off when using effect size measure. For further developments, new effect size measures should be explored in order to balance the FP and the CI percentages (Gómez-Benito, Hidalgo, et al., 2009).

Table 11: Percentage of false positive using effect size measure.

	J=20			*J=40*			*J=60*		
	DIF–U	$\Delta R^2 - U$	Reduction (%)	DIF–U	$\Delta R^2 - U$	Reduction	DIF–U	$\Delta R^2 - U$	Reduction (%)
No Bias									
N=500	5.60	0.30	18.67	5.30	0.35	15.14	5.22	0.48	10.87
N=1000	4.85	0.00	–	4.97	0.00	–	5.03	0.02	–
N=2000	5.61	0.00	–	4.80	0.00	–	4.80	0.00	–
10% Bias $\delta = 0.4$									
N=500	6.44	0.27	23.85	5.81	0.36	16.14	5.50	0.48	11.46
N=1000	5.72	0.00	–	5.36	0.00	–	5.26	0.02	–
N=2000	4.90	0.00	–	4.70	0.00	–	4.80	0.00	–
10% Bias $\delta = 0.8$									
N=500	6.94	0.22	31.54	6.50	0.33	19.70	6.35	0.48	13.23
N=1000	7.17	0.00	–	6.42	0.00	–	6.15	0.02	–
N=2000	7.90	0.00	–	6.95	0.00	–	8.00	0.00	–
20% Bias $\delta = 0.4$									
N=500	6.81	0.31	21.97	6.78	0.34	19.94	6.23	0.48	12.98
N=1000	6.87	0.00	–	6.59	0.00	–	6.17	0.00	–
N=2000	7.70	0.00	–	7.10	0.00	–	8.20	0.00	–
20% Bias $\delta = 0.8$									
N=500	10.06	0.87	11.56	9.47	0.59	16.05	9.17	0.94	9.75
N=1000	11.31	0.06	–	10.66	0.03	–	11.27	0.02	–
N=2000	17.70	0.00	–	15.90	0.00	–	16.11	0.00	–

– refers to reduction $\geq 100\%$ or reduction impossible to compute due to denominator equal to zero (no false positive detected).

Table 12: Percentage of correct identification using effect size measure.

	J=20			*J=40*			*J=60*		
	DIF–U	$\Delta R^2 - U$	Reduction (%)	DIF–U	$\Delta R^2 - U$	Reduction	DIF–U	$\Delta R^2 - U$	Reduction (%)
10% Bias $\delta = 0.4$									
N=500	25.00	3.50	7.14	21.00	4.25	4.94	24.83	5.00	4.97
N=1000	39.50	0.00	–	34.75	0.75	46.33	37.66	0.66	57.06
N=2000	63.00	0.00	–	57.25	0.00	–	53.17	0.00	–
10% Bias $\delta = 0.8$									
N=500	68.00	30.50	2.23	65.75	24.75	2.66	67.33	29.00	2.32
N=1000	89.00	9.50	9.37	86.50	11.00	7.86	86.50	11.17	7.74
N=2000	99.50	1.00	99.50	98.50	1.75	56.28	97.70	0.67	–
20% Bias $\delta = 0.4$									
N=500	21.00	2.25	8.89	20.25	3.25	6.23	20.33	3.17	6.41
N=1000	30.25	0.00	–	30.50	0.12	–	29.67	0.33	89.91
N=2000	47.75	0.00	–	48.37	0.00	–	46.67	0.00	–
20% Bias $\delta = 0.8$									
N=500	59.50	22.50	2.64	56.50	20.37	2.77	58.42	20.25	2.88
N=1000	83.00	5.25	15.81	80.62	7.25	11.12	80.08	6.42	12.47
N=2000	97.50	0.25	–	94.00	0.37	–	95.10	0.17	–

– refers to reduction $\geq 100\%$ or reduction impossible to compute due to denominator equal to zero (no correct identification detected).

5. Application to a real dataset

Now, we provide an application of DIF detection analysis with particular focus on the Zumbo's methodology. The application involves different academic tracks (section 5.1) from INVALSI sample 2016/2017, described into section 5.2. We carry out DIF detection analysis two groups at a time (section 5.3), and we have three different subsamples (subsections 5.3.1, 5.3.2 and 5.3.3) in which we carry out DIF detection analysis (subsection 5.3.4). Finally, the last section (section 5.4) presents conclusions and discussion about DIF results.

5.1 Academic tracks

The Italian education system presents an horizontal stratification at upper secondary school level. The stratification involves academic, technical, vocational schools and vocational training courses. All curricular programs are decided at national level and the schools provide some similar subjects, such as Italian language and literature, mathematics, history, one or more foreign language and so on, while they differ for specific subjects. This differentiation is due to different track purpose. As said in section 2.2.1, academic schools provide academic and general curricula, technical schools aim to prepare students for labor market, especially for technical and economic positions and vocational schools transfer to pupils vocational skills oriented to industry and handicraft and services[33].

For the simulation we have considered academic and technical schools, while for the application, now, we consider only academic schools. Therefore, here, we exclusively focus on academic track. Italian academic track presents a further horizontal stratification that differs schools for the main subjects. Students can choose from classic lyceum (*liceo classico*) characterized by Latin and Ancient Greek; scientific lyceum (*liceo scientifico*) focused on scientific studies with mathematics, physics, chemistry, biology, earth science and computer science; linguistic lyceum (*liceo linguistico*) characterized by modern foreign languages such as English, French, Spanish and German, but also Russian, Arabic and Chinese; artistic lyceum (*liceo artistico*) where the emphasis is on theoretical and practical arts; human sciences lyceum (*liceo delle scienze umane*) oriented on sociology, psychology,

[33]In addition to these three branches, Italian pupils can opt for vocational training courses. We do not consider them because there are no pupils from this track into INVALSI sample.

anthropology and pedagogy; music and dance lyceum (*liceo musicale e coreutico*) oriented to teach students music, playing instruments, dance and choreography.

The choice of apply the new methodology to different academic tracks leads an important advantage. These groups should be more similar than groups of first stratification (academic, technical, vocational tracks and vocational training courses), making easier matching and results interpretation. Indeed, propensity score matching should perform well since groups should be low imbalanced. This helps and eases the DIF attribution to group allocation mechanism. Moreover, we find very interesting to consider pupils from academic tracks because they should exhibit similar achievement, although scholastic curricula differ from each other.

At the end of this chapter (section 5.4), discussion about bias results are proposed. We expect to find differences in raw scores among academic schools because these different curricula should transfer different abilities to pupils. However, we expect that the instrument is fair with respect to academic track and the new methodology allows to assess this issue and to attribute differential item functioning to group membership. For possible DIF item detected, we analysis them more in detail, through a qualitative analysis about items format and content. It is possible that some item format favors pupils from one particular track and some school tracking could benefit some item contents. Therefore, if DIF item is detected, this kind of analysis helps detect possible sources of DIF due to group allocation.

5.2 Data

For the application we use sample INVALSI referred to 2016/2017 academic year. Sample contains 38285 units for the Italian language test, while the sample counts 38120 pupils for the maths test. Our analytic sample contains only pupils with information about both tests and pupils from academic tracks. Therefore, the final sample size contains 15699 pupils due to merge and data cleaning.

Table 13: Sample composition by different academic tracks.

	N	%
Artistic	1281	8.16
Classic	1815	11.56
Human sciences	2642	16.83
Linguistic	2826	18.00
Scientific	7135	45.45

Table 13 presents the distribution of pupils into different academic schools. We aggregate music and dance lyceum with artistic lyceum because of low sample size of pupils[34] into first lyceum. We opt for this choice because these schools teach similar main subjects. Almost one half of sample attends to scientific lyceum (45.5%). Linguistic and human sciences (HS) lyceum are frequented by about one fifth of students (respectively, 18.0% and 16.3%), while one pupil on ten is enrolled in classic (11.6%) or artistic lyceum (8.2%).

Table 14: Descriptives of maths and Italian language raw scores by different tracks.

	Maths				*Italian language*			
	Mean	SD	Max	Min	Mean	SD	Max	Min
Artistic	39.87	18.01	95.00	2.50	55.19	15.30	91.84	2.04
Classic	54.48	19.19	97.50	2.50	73.68	11.69	97.96	8.16
Human sciences	39.22	17.29	95.00	2.50	58.72	13.60	93.88	10.20
Linguistic	48.80	18.30	97.50	2.50	66.43	12.54	93.88	4.08
Scientific	69.18	18.89	97.50	2.50	68.05	13.30	97.96	2.04

From the table above (table 14), pupils from artistic and human sciences presents the lowest performance, in terms of raw scores, both in math and Italian language. Scientific outperforms other tracks in maths, while only classic overcomes it in Italian language test. Classic and linguistic present results very similar in both tests, although the former outperforms the latter. We are interested in DIF detection analysis among pupils from different academic tracks, therefore we need to reduce the groups number in order to carry

[34]Only 1% of pupils is enrolled in music and dance lyceum (174 pupils in the former and 29 in the latter).

out DIF analysis in simple way. We decide to aggregate academic tracks as follow: artistic with human sciences (HS), classic with linguistic and scientific alone. This aggregation takes the sample size and similar raw score in both tests into consideration.

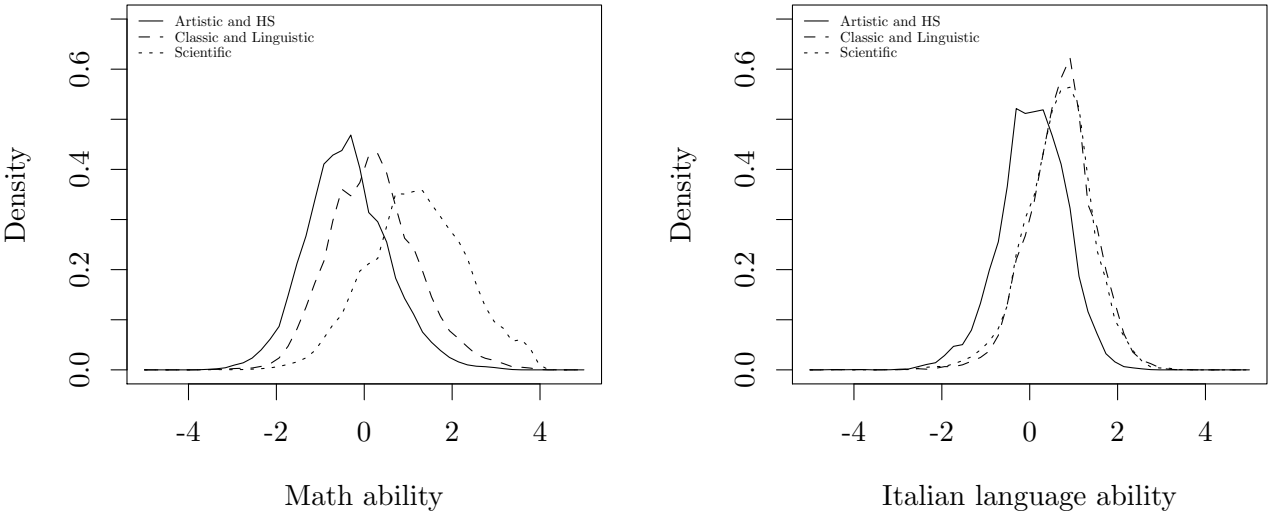


Figure 15: Kernel density in maths and Italian language test for the three groups.

The aggregation leads to three new groups and figure 15 shows math and Italian language abilities distributions across the groups. Abilities are estimated under a Rasch model. We can see that scientific schools tend to have students with higher math ability than others, while classic and linguistic schools tend to have students with higher math ability than artistic and human sciences schools. Differently, students from scientific and classic and linguistic schools present similar Italian language abilities and higher than students from artistic and human sciences schools.

5.3 Results

Now, we provide DIF detection analysis among pupils from these three groups, two groups at a time. This is because traditional DIF detection methods and Zumbo's methodology consider usually only two groups[35], aware that it is a limitation of our

[35]In literature, some works provide methods of DIF detection for multiple groups (Finch, 2016; Woods et al., 2013; Magis, Raîche, et al., 2011).

research. Consequently, we have three different applications: scientific *vs* classic and linguistic (application 1), scientific *vs* artistic and human sciences (application 2) and classic and linguistic *vs* artistic and human sciences (application 3).

We apply conditional logistic regression. Therefore, first of all, we check covariates balancing between groups and apply propensity score matching in order to reduce selection bias. Conditional logistic regression, as traditional DIF methods, is based on statistical tests that are distributed as a χ^2 and they have been criticized (Kirk, 1996; J. Cohen, 1994) because of sensitive to sample size. Here, our sample size is very large (N=15699), consequently, we can meet high risk of false positive identification. Therefore, we limit out applications to Lombardy. We choose this region because it presents the higher sampling number (N=1579).

The new methodology suggests to apply DIF detection analysis only for groups which are comparable (without imbalanced covariates). Propensity score matching allows to reduce covariates imbalance, but this technique requires no missing data. Therefore, sample reduction occurs for guaranteeing the matching. In particular, sample reduction is equal about to 10% and the final sample size contains 1427 pupils. We apply propensity score matching only for that covariates which are significantly imbalanced in reference and focal group. Consequently, the covariates can be different among the three applications. The covariates are the same used for the simulation[36] and described in section 3.1.2, with a difference: ESCS index is treated as continuous variable. In addition, we consider other variables which can affect the pupils' upper secondary school choice: the regularity of previous study (*regular*), if dialect is spoken at home (*dialect*), books number at home (*books*), student attendance to primary school more than one year (*primary*) and material deprivation index[37] (*material deprivation*).

5.3.1 Scientific *vs* classic and linguistic

Before to apply conditional logistic regression, balancing check and propensity score matching analysis is conducted for each application. First of all, we consider application between scientific schools and classic and linguistic schools. In this case, scientific track

[36] No presence of area variable because we circumscribe application to one region.

[37] The index is an additive index composed by the student's tenure of quiet place to study, computer to study, desk for homework, encyclopedia (made up of books or CD-ROMs or DVDs), Internet and single room.

refers to reference group, while classic and linguistic track refers to focal group.

Table 15: Covariates balancing between scientific and classic and linguistic tracks.

	Scientific	Classic and Linguistic	p	SMD
n	642	402		
Gender				
Male (%)	56.4	17.2	<0.001	0.809
Citizen				
Not Italian (%)	8.1	10.2	0.239	0.081
Aspiration				
Low (%)	10.1	20.1	<0.001	0.288
Regular				
Yes (%)	93.8	90.8	0.096	0.112
Books *(%)*			0.038	0.162
< 25	7.3	10.4		
> 26 *and* < 200	48.6	52.5		
> 201	44.1	37.1		
Dialect				
Yes (%)	29.4	25.1	0.149	0.097
Primary				
No or less than one year (%)	5.0	5.5	0.839	0.022
ESCS				
(mean)	0.56	0.42	0.011	0.161
Material deprivation				
(mean)	0.91	0.84	0.742	0.021

Here, reference and focal groups are very similar with respect to considered covariates. Table 15 shows standardized mean differences (SMD) between the two groups. Only gender, aspiration, books and ESCS index are statistically different between groups. Scientific schools are more attended by males and students with higher aspiration. Although books and ESCS index present significant differences, they have negligible or low differences because SMD are close to 0.100. Consequently, these two groups are very similar.

Table 16: Percentage of Bias Reduction (PBR): scientific vs classic and linguistic.

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.682	0.508	0.174	0.682	0.000	99.9
Gender (Male)	0.564	0.172	0.392	0.564	0.000	100.0
Aspiration (Low)	0.099	0.201	-0.101	0.099	0.000	100.0
Books (> 26 and < 200)	0.486	0.525	-0.039	0.491	-0.005	87.9
Books (> 201)	0.441	0.371	0.070	0.436	0.005	93.3
ESCS	0.564	0.421	0.143	0.554	0.010	93.1

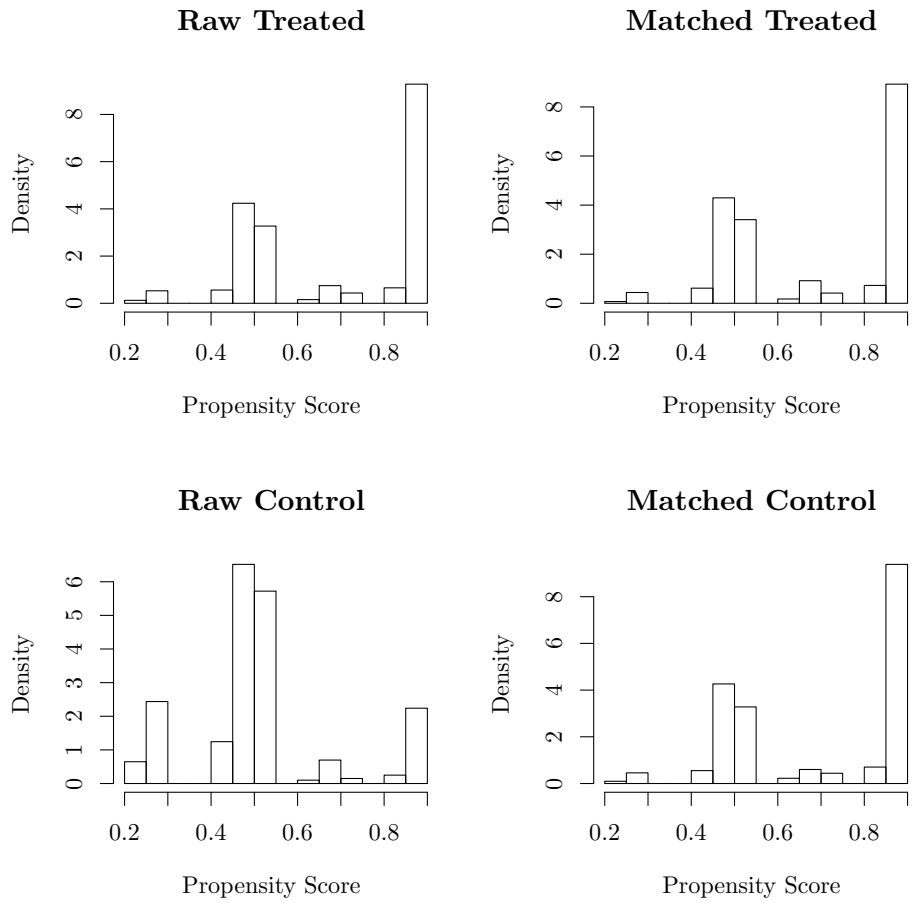


Figure 16: Percentage of Bias Reduction (PBR): scientific vs classic and linguistic.

Propensity score matching is performed for gender, aspiration, books and ESCS index. In particular, we opt for a full matching with a combination of one–to–many and many–to–one, where maximum control and treated are set to 13 subjects. Table 16 shows that propensity score matching reduces the covariates differences between groups with high reduction for all covariates from 100.0% for gender and aspiration to 87.9% for the first category of books variable. Despite the groups are very similar before matching, propensity score matching balances covariates for students with high value of propensity score (figure 16).

5.3.2 Scientific *vs* artistic and human sciences

Now, we carry out matching analysis between scientific track (reference group) and artistic and human sciences track (focal group).

Table 17: Covariates balancing between scientific and artistic and HS tracks.

	Scientific	Artistic and HS	p	SMD
n	642	383		
Gender				
Male (%)	56.4	24.3	<0.001	0.693
Citizen				
Not Italian (%)	8.1	9.4	<0.549	0.046
Aspiration				
Low (%)	10.0	36.0	<0.001	0.651
Regular				
Yes (%)	93.8	85.4	<0.001	0.277
Books *(%)*			0.001	0.236
< 25	7.3	12.0		
> 26 *and* < 200	48.6	54.0		
> 201	44.1	33.9		
Dialect				
Yes (%)	29.4	35.8	0.042	0.135
Primary				
No or less than one year (%)	5.0	8.4	0.043	0.135
ESCS				
(mean)	0.56	0.17	<0.001	0.455
Material deprivation				
(mean)	0.71	0.87	<0.009	0.166

Scientific and artistic and HS schools are very imbalanced with respect to covariates (table 17). Reference group is more represented by males (56%), pupils with high aspiration (90%) and pupils with higher value of economic–social–cultural index (0.56), on average, than focal group. Reference and focal group are imbalanced, although in less marked way, with respect to regular, books, dialect, primary and material deprivation. Citizen does not present statistically significant differences between the groups.

Table 18: Percentage of Bias Reduction (PBR): scientific vs artistic and human sciences.

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.711	0.484	0.227	0.707	0.004	99.3
Gender (Male)	0.564	0.243	0.321	0.559	0.004	98.7
Aspiration (Low)	0.099	0.360	-0.261	0.003	-0.003	98.8
Regular (Yes)	0.938	0.854	0.084	0.929	0.009	89.8
Book (> 26 and < 200)	0.486	0.540	-0.054	0.513	-0.027	50.1
Book (> 201)	0.441	0.339	0.101	0.407	0.034	66.7
Dialect (Yes)	0.294	0.358	-0.063	0.250	0.044	30.1
Primary (No or less 1 year)	0.049	0.084	-0.034	0.073	-0.024	29.9
ESCS	0.564	0.167	0.397	0.484	0.079	79.9
Material deprivation	0.710	0.869	-0.159	0.785	-0.075	52.8

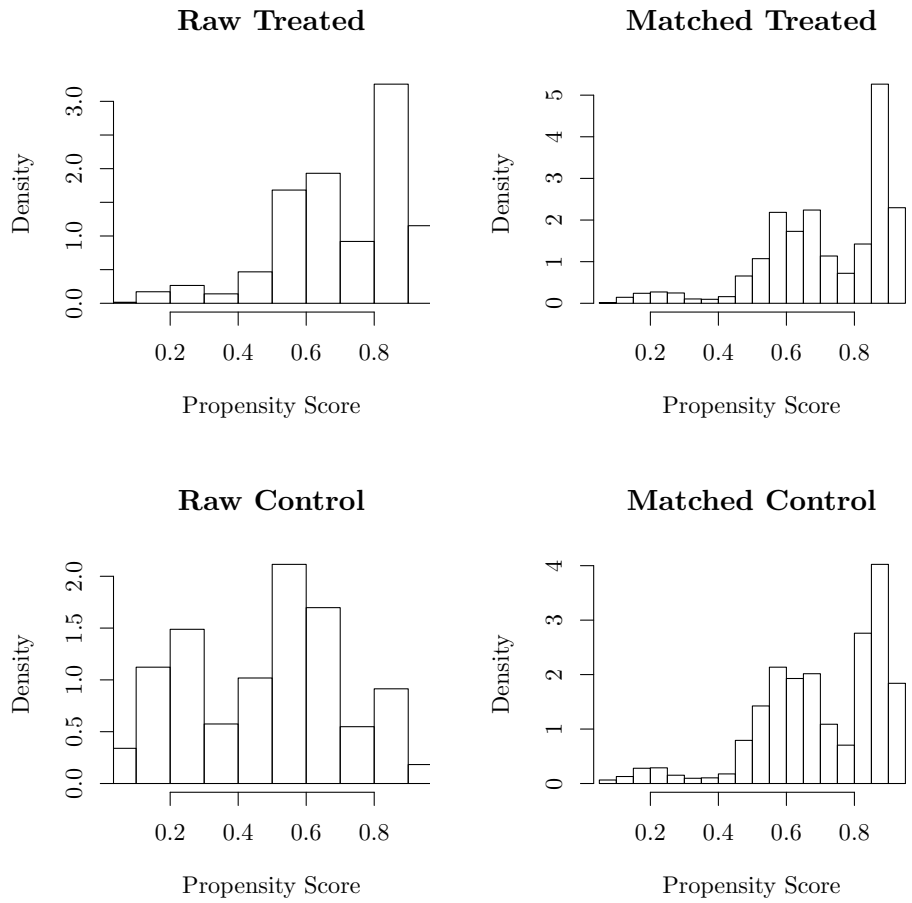


Figure 17: Percentage of Bias Reduction (PBR): scientific vs artistic and human sciences.

Here, propensity score matching is performed for gender, aspiration, regular, books, dialect, primary, ESCS index and material deprivation. We opt for a full matching with a combination of one–to–many and many–to–one and set 8 as maximum number controls and treated. From table 18 and figure 17, it is possible to observe that propensity score matching reduces significantly selection bias in reference and focal groups. In particular, we find high reduction for gender, aspiration, regular and ESCS (from 79.9 % to 98.8%) and moderate for the others (from 30.1% to 66.7%).

5.3.3 Classic and linguistic *vs* artistic and human sciences

The final application considers pupils from classic and linguistic schools (reference group) and from artistic and human sciences schools (focal group).

Table 19: Covariates balancing between classic and linguistic and artistic and HS tracks.

	Classic and Linguistic	Artistic and HS	p	SMD
n	402	383		
Gender				
Male (%)	17.2	24.3	0.018	0.176
Citizen				
Not Italian (%)	10.4	9.4	0.710	0.035
Aspiration				
Low (%)	20.1	36.0	<0.001	0.359
Regular				
Yes (%)	90.8	85.4	0.025	0.168
Books *(%)*			0.590	0.073
< 25	10.4	12.0		
> 26 *and* < 200	52.5	54.0		
> 201	37.1	33.9		
Dialect				
Yes (%)	25.1	35.8	0.002	0.233
Primary				
No or less 1 year (%)	5.5	8.4	0.146	0.114
ESCS				
(mean)	0.42	0.17	<0.001	0.288
Material deprivation				
(mean)	0.73	0.87	<0.033	0.152

Here, reference and focal group are imbalanced with respect to all covariates, except for citizen, books and primary (table 19). In particular, males (17.2%), pupils with low school aspiration (20.1%), dialect spoken at home (25.1%) and ESCS index (mean of 0.73) are under–represented in classic and linguistic rather than in artistic and human sciences schools. Conversely, regular students (90.8%) and material deprivation index (mean of 0.73) are over–represented in reference group.

Table 20: Percentage of Bias Reduction (PBR): classic and linguistic vs artistic and human sciences.

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.542	0.480	0.062	0.542	0.000	99.5
Gender (Male)	0.172	0.243	-0.071	0.180	-0.008	88.1
Aspiration (Low)	0.201	0.360	-0.159	0.201	0.003	99.7
Regular (Yes)	0.908	0.854	0.054	0.906	0.002	95.8
Dialect (Yes)	0.251	0.358	-0.106	0.271	-0.020	81.1
ESCS	0.438	0.421	0.167	0.459	-0.038	84.9
Material deprivation	0.729	0.869	-0.141	0.632	0.097	30.9

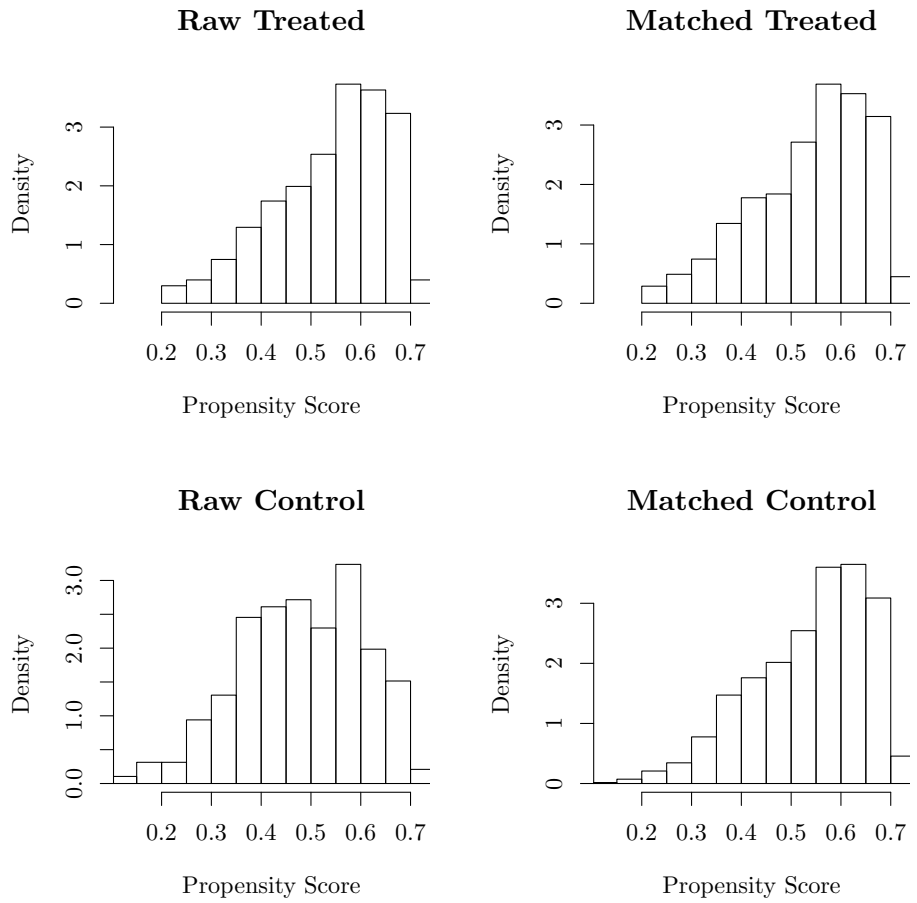


Figure 18: Percentage of Bias Reduction (PBR): classic and linguistic vs artistic and human sciences.

Table 20 shows percentage of bias reduction of full matching with a combination of one–to–many and many–to–one, where maximum control and treated are set to 8 subjects. Matching is performed for gender, aspiration, regular, dialect, ESCS index and material deprivation. It reduces differences in the propensity score distribution between the two groups. The bias reduction is moderate for material deprivation (30.9%) and high for other covariates (from 81.1% to 99.7%).

5.3.4 DIF detection analysis

Now, we can carry out DIF detection analysis following the new methodology. After matching, we are able to apply conditional logistic regression considering the data dependence structure. In the following pages, DIF detection analysis is carried out for both maths and Italian language test. As described in previous section, three different

applications are considered: scientific *vs* classic and linguistic (application 1), scientific *vs* artistic and HS (application 2) and classic and linguistic *vs* artistic and HS (application 3). DIF detection analysis considers both null hypothesis significance test and the effect size measure. Complete DIF analysis are inserted in appendix E for all three applications: table E2, E3 and E4 show group coefficients, hypothesis test values, with relative statistical significance, and ΔR^2 values, where necessary[38], for both tests.

Maths INVALSI test 2016/2017 for secondary schools consists of 40 items. These items aim to investigate students' knowledge of mathematics, hence they have mathematical contents. The maths test involves four different content areas: quantity, space and shape, change and relationship, uncertainty and data. A second dimension, on which maths items are constructed, concerns three moment of processes[39]: formulating, employing, interpreting (INVALSI, 2012b). The items have four different formats: multiple–choice questions with four possible choices (14 items), open–ended questions (19 items), complex multiple–choice questions (6 items) and one cloze question[40]. Finally, each item is transformed in dichotomous variable, where 1 refers to correct answer and 0 refers to wrong answer (INVALSI, 2017c, p. 95).

Table 21 shows DIF results about maths test. Considering only null hypothesis test, the first two applications present large DIF items, respectively, the 40% (16 items) and 45% (18 items), while the last application presents the 12.5% (5 items) of DIF items. Although DIF items are numerous, if we use the effect measure size we find that only three items are significantly flagged as DIF. In particular, item $M30$ has large DIF size effect in the first application and moderate DIF size effect in the second application. The first application presents also item $M31$ as DIF (with moderate effect size), while the last application has no DIF items.

[38] Values are not reported for hypothesis significance test not statistical significance.
[39] Activities for problem resolution.
[40] It is an exercise where some words or signs are removed from text portion. The test participants must replace the missing words or signs.

Table 21: DIF results about maths test among pupils from different academic tracks.

	1st application		2nd application		3rd application	
	Test	Size	Test	Size	Test	Size
M1	DIF	A	NoDIF	A	NoDIF	A
M2	NoDIF	A	NoDIF	A	NoDIF	A
M3	NoDIF	A	NoDIF	A	NoDIF	A
M4a	NoDIF	A	NoDIF	A	NoDIF	A
M4b	DIF	A	NoDIF	A	NoDIF	A
M5	DIF	A	NoDIF	A	DIF	A
M6	DIF	A	NoDIF	A	NoDIF	A
M7	NoDIF	A	NoDIF	A	NoDIF	A
M8	DIF	A	DIF	A	NoDIF	A
M9	NoDIF	A	NoDIF	A	NoDIF	A
M10	NoDIF	A	NoDIF	A	NoDIF	A
M11	NoDIF	A	NoDIF	A	NoDIF	A
M12	DIF	A	DIF	A	NoDIF	A
M13	DIF	A	DIF	A	DIF	A
M14a	DIF	A	DIF	A	NoDIF	A
M14b	NoDIF	A	DIF	A	DIF	A
M14c	DIF	A	DIF	A	NoDIF	A
M15	NoDIF	A	DIF	A	NoDIF	A
M16a	NoDIF	A	DIF	A	NoDIF	A
M16b	NoDIF	A	NoDIF	A	NoDIF	A
M16c	NoDIF	A	DIF	A	NoDIF	A
M17	NoDIF	A	NoDIF	A	NoDIF	A
M18	NoDIF	A	NoDIF	A	NoDIF	A
M19	NoDIF	A	DIF	A	DIF	A
M20a	NoDIF	A	NoDIF	A	NoDIF	A
M20b	NoDIF	A	DIF	A	NoDIF	A
M21	NoDIF	A	NoDIF	A	NoDIF	A
M22	NoDIF	A	DIF	A	NoDIF	A
M23	NoDIF	A	DIF	A	NoDIF	A
M24	DIF	A	DIF	A	NoDIF	A
M25	NoDIF	A	NoDIF	A	NoDIF	A

Continued on next page

Table 21 – continued from previous page.

	1st application		2nd application		3th application	
	Test	Size	Test	Size	Test	Size
M26a	NoDIF	A	NoDIF	A	NoDIF	A
M26b	NoDIF	A	NoDIF	A	NoDIF	A
M27	NoDIF	A	NoDIF	A	NoDIF	A
M28	DIF	A	DIF	A	DIF	A
M29a	DIF	A	DIF	A	NoDIF	A
M29b	DIF	A	DIF	A	NoDIF	A
M30	DIF	C	DIF	B	NoDIF	A
M31	DIF	B	NoDIF	A	NoDIF	A
M32	DIF	A	NoDIF	A	NoDIF	A

A = Negligible, B = Moderate and C = Large effect.

Italian language INVALSI test 2016/2017 for secondary schools consists of 49 items. This test wants to assess reading (comprehension, interpretation, reflection and evaluation of written text) and grammatical competences (INVALSI, 2012a). So, the Italian language test consists of two different part. The first one aims to investigate student's reading comprehension, while the second one aims to investigate student's ability and knowledge of language. The first part presents questions about two argumentative texts (text A with 10 questions and D with 9 questions), 10 questions about an argumentative–expositive text (text B) and 10 questions about a poetical text (text C). The second part (text E) is composed by 10 questions about student's ability and knowledge of language. Items have different formats: multiple–choice questions with four possible choices (32 items), open–ended questions (10 items), complex multiple–choice questions (6 items) and one cloze question. Finally, as for maths test, each item is transformed in dichotomous variable, where 1 refers to correct answer and 0 refers to wrong answer (INVALSI, 2017c, p. 83).

The Italian language test shows less DIF items than maths test, using only statistical test. The first application shows about the 8% (4 items) of DIF items, while the second and the third application show, respectively, about the 20% (10 items) and the 25% (12 items) DIF items. The effect size measure flags only item $A4\_3$ as DIF (moderate effect) for application 1, while in application 2 items $C3$, $C8$ and $E10$ present DIF with moderate

effect. Differently, comparing pupils of third application, two items are flagged as moderate DIF ($B3$ and $E3$) and four items as large DIF ($A4\_2$, $A4\_2$, $A4\_3$ and $E10$).

Table 22: DIF results about Italian language test among pupils from different academic tracks.

	1st application		2nd application		3rd application	
	Test	Size	Test	Size	Test	Size
A1	NoDIF	A	NoDIF	A	NoDIF	A
A2	NoDIF	A	NoDIF	A	NoDIF	A
A3	NoDIF	A	NoDIF	A	NoDIF	A
A4_1	NoDIF	A	NoDIF	A	DIF	C
A4_2	NoDIF	A	NoDIF	A	DIF	C
A4_3	DIF	B	NoDIF	A	DIF	C
A4_4	NoDIF	A	NoDIF	A	NoDIF	A
A4_5	NoDIF	A	DIF	A	NoDIF	A
A4_6	NoDIF	A	NoDIF	A	NoDIF	A
A5	NoDIF	A	NoDIF	A	NoDIF	A
B1	NoDIF	A	NoDIF	A	NoDIF	A
B2	NoDIF	A	DIF	A	DIF	A
B3	DIF	A	NoDIF	A	DIF	B
B4	NoDIF	A	NoDIF	A	NoDIF	A
B5	NoDIF	A	NoDIF	A	NoDIF	A
B6	NoDIF	A	NoDIF	A	NoDIF	A
B7	NoDIF	A	NoDIF	A	NoDIF	A
B8	NoDIF	A	NoDIF	A	NoDIF	A
B9	DIF	A	DIF	A	NoDIF	A
B10	NoDIF	A	NoDIF	A	NoDIF	A
C1	NoDIF	A	NoDIF	A	NoDIF	A
C2	NoDIF	A	NoDIF	A	NoDIF	A
C3	NoDIF	A	DIF	B	NoDIF	A
C4	NoDIF	A	NoDIF	A	DIF	A
C5	NoDIF	A	NoDIF	A	NoDIF	A
C6	NoDIF	A	DIF	A	DIF	A
C7	NoDIF	A	NoDIF	A	NoDIF	A
C8	NoDIF	A	DIF	B	NoDIF	A

Continued on next page

Table 22 – continued from previous page.

	1st application		2nd application		3th application	
	Test	Size	Test	Size	Test	Size
C9	NoDIF	A	NoDIF	A	NoDIF	A
C10	NoDIF	A	NoDIF	A	NoDIF	A
D1	NoDIF	A	NoDIF	A	NoDIF	A
D2	NoDIF	A	NoDIF	A	NoDIF	A
D3	DIF	A	NoDIF	A	NoDIF	A
D4	NoDIF	A	NoDIF	A	NoDIF	A
D5	NoDIF	A	NoDIF	A	NoDIF	A
D6	NoDIF	A	NoDIF	A	NoDIF	A
D7	NoDIF	A	NoDIF	A	NoDIF	A
D8	NoDIF	A	NoDIF	A	DIF	A
D9	NoDIF	A	NoDIF	A	NoDIF	A
E1	NoDIF	A	NoDIF	A	NoDIF	A
E2	NoDIF	A	DIF	A	DIF	A
E3	NoDIF	A	DIF	A	DIF	B
E4	NoDIF	A	NoDIF	A	NoDIF	A
E5	NoDIF	A	NoDIF	A	NoDIF	A
E6	NoDIF	A	DIF	A	DIF	A
E7	NoDIF	A	NoDIF	A	NoDIF	A
E8	NoDIF	A	NoDIF	A	NoDIF	A
E9	NoDIF	A	NoDIF	A	NoDIF	A
E10	NoDIF	A	DIF	B	DIF	C

A = Negligible, B = Moderate and C = Large effect.

5.4 Discussion

Previous analysis has shown that the amount of items flagged as DIF is significant, especially for maths test. Nevertheless, results change if we use misuse effect size in order to identify DIF items. Indeed, it occurs a significant reduction of DIF items with the second method. In particular, three items are flagged as DIF for the mathematics test, while ten for the Italian language test. Applications involve situations in which the number of test takers is

around one thousand units[41] for length tests equal to 40 and 49 items. Section 4.4 has shown that, in these situations, using effect size measure leads to better DIF identification. In particular, we are sure to not mistakenly flag items as DIF, although the correct identification of DIF items is low. Nevertheless, as said in section 4.3.2, it is more import to keep false positive inflation low rather than keep correct identification high. Therefore, we consider only effect size measure to understand DIF results, because it allows to keep false positive inflation low (close to zero).

Now, we analyze these DIF items more in detail. Tables E2, E3 and E4, in appendix E, show DIF results in details. In particular, they show group coefficients of the conditional logistic regressions that allow to assess which group has advantage or not. For the maths test, scientific presents an advantage for $M30$ and $M31$ over classic and linguistic track and for $M30$ over artistic and human sciences track. For Italian language test, $A4\_3$ is unfair in favor of scientific rather than classic and linguistic. Scientific exhibits also some advantage for $E10$ over artistic and HS, but for items $C3$ and $C8$ fairness changes in favor of artistic and HS schools. Finally, classic and linguistic track presents disadvantage for $A4\_1$, $A4\_2$, $A4\_3$ and $B3$ over artistic and human sciences, while it exhibits some advantage for $E3$ and $E10$.

Considering the DIF items content, item $M30$ involves questions about mathematical function, while item $M31$ concerns natural numbers (INVALSI, 2017b). DIF results show a systematic advantage of scientific schools in these items. For the Italian language test (INVALSI, 2017a), scientific exhibits advantage in one item ($A4\_3$) of over classic and linguistic in argumentative text. Poetical text is unfair for two items ($C3$ and $C8$) in favor of artistic and HS rather than scientific, which has an advantage for item $E10$ in language knowledge. Finally, argumentative text presents systematic disadvantage ($A4\_1$, $A4\_2$, $A4\_1$ and $B3$) for classic and linguistic rather than artistic and HS which, in turn, exhibits advantage in two items ($E3$ and $E10$) about language knowledge.

Table E1, in appendix E, exhibits items format for both tests. There is no pattern for the maths test due to few items flagged as DIF. Conversely, the Italian language test seems to show some patterns. Artistic and human sciences schools present advantages about multiple–choice questions over scientific ($C3$ and $C8$) and classic and linguistic ($A4\_1$, $A4\_2$ and $A4\_1$) schools, but it shows disadvantages in complex multiple–choice questions:

[41] Applications counts, respectively, 1044, 1025 and 785 pupils.

$E10$ over scientific and $E3$ and $E10$ over classic and linguistic. Nevertheless, item format analysis does not lead to a clear evidence and explanation.

6. Conclusions

Test users and policy makers often direct their public actions from the educational standardized test results. They usually operate basing on the test total raw scores. This is possible only assuming that the test is comparable among different groups, but this is not alway correct. It is possible that the test, or a part of the test, advantages some subgroups rather than others, therefore it results biased and unfair. Psychometric literature refers to differential item functioning when you want to detect possible unfair items among individuals from different groups. In educational context, DIF occurs when individuals with the ability but allocated into different groups present different probability of success to the item. Therefore, it is necessary to conduct DIF detection analysis in order to assess test fairness.

This thesis work has had as first research goal that of assessing the performance of a new methodology, recently proposed in literature for DIF detection analysis. This new method, based on a redefinition of biased item, allows to reduce pre–existing differences among groups. For this first goal, a simulation study has been implemented in which groups are constructed to be imbalanced with respect to covariates. The simulation study supports the new methodology in some situations. The assumptions on item difficulty parameters and test length have no significance impact, while the other manipulated factors exhibit an impact on DIF methods performances. Although the new methodology performs similar to traditional DIF detection methods in some situations (no DIF items and small simple size), it outperforms traditional DIF detection methods in situations where sample size is large, DIF is present and the DIF size is large. Therefore, we recommend the new methodology for imbalanced groups both because it presents the best performances and it allows to attribute DIF to group allocation.

Despite the new methodology presents the best performances, it suffers from high false alarm rates for large sample size. Therefore, we have integrated simulation study with an effect size measure for the new methodology. This measure is based on ΔR^2, similar to measure for conventional logistic regression for DIF detection. The simulation results have suggested that using the proposed effect size measure reduces sensibility the I error inflation. In addition, the reduction for large samples is close to zero per cent of chance

to commit this kind of error, but the effect size measure reduces sensibility the correct identification of DIF items. Nevertheless, this measure is recommended because is more "dangerous" to mistakenly identify an item as DIF, losing useful information linked to the item in the analysis.

The second aim of this work has been to assess INVALSI tests, comparing different academic schools in optic of test fairness. It seems that INVALSI tests present fair items comparing pupils from different academic tracks. In this context, the INVALSI instrument is robust, especially for mathematics. Some problems seems to be comparing classical and linguistic to artistic and human sciences schools for Italian test. However, these results could be affected by chosen aggregation. In addition, we tried to analyze deeper the few items identified as DIF. In particular, we proposed a content and format item analysis. Results exhibited no significant patterns for the second one. Differently, the content analysis presented significant patterns. Pupils from scientific schools show a systematic advantage in two items about mathematical function and natural numbers. In addition, poetical text tend to advantage artistic and human science rather than scientific schools. Argumentative text presents systematic disadvantage for classic and linguistic rather than artistic and human science which, in turn, exhibits advantage in two items about language knowledge. However, very few items are flagged as DIF by the new methodology and we can conclude that INVALSI tests are fair for pupils from different academic schools.

6.1 Policy implications

The differential item functioning item presence leads consequences for a standardized test. Traditionally, when an item is flagged as DIF, a qualitative assessment from an expert equip is required in order to decide the "fate" of DIF item (Ramsey, 1993; Berk, 1982). The new approach gives a statistical instrument to DIF detection analysis since it allows to attribute DIF item to group allocation mechanism. Consequently, this new methodology attributes DIF presence to allocation to one group rather than other, excluding the DIF sources due to other possible confounding factors. When groups are not random and the allocation depends on other individual characteristics, this statistical technique is very useful and it develops possible policy implications. Test users and policy makers should pay attention to DIF detection analysis when they use and assess standardized test results.

DIF presence leads two critical issues that test users and policy makers should keep in mind. On the one hand, traditionally, their evaluations and public decisions (actions) on standardized tests take place on the basis of raw scores, e.g., the total test scores. As has already been explained, this could be self-defeating whether some items advantage one group rather the others. Consequently, policy actions could act in a wrong way, maintaining social inequalities or even persisting inequalities and aggravating the situation. On the other hand, understanding DIF sources should help policy makers act: one group could be systematically disadvantaged in items with particular topic, so policy action should understand where and why this happens in order to modify the situation. For example, Le (2009) shows gender DIF in science PISA test 2006 that depends on item formats and content domains. In particular, males tend to be advantaged on multiple choice and closed response items and on items about science knowledge. Therefore, test users and policy makers should act according to these results. Firstly, when they use and interpret test results they should keep in mind that some items do not measure the same ability. Secondly, they should promote policies in order to improve future tests and delete unfairness.

Now, we provide two final considerations, linked to results from this work, that allow to formulate two policy implications. Firstly, our analysis exhibits the fairness for the majority of items. In other words, INVALSI tests are robust measurement instrument, comparing pupils from different academic tracks. Therefore, for Italian academic schools, the administration of the same standardized test on school competences is fair and a robust instrument to measure school pupils ability. Secondly, the few DIF items have been analysed by format and content. If the first one does not present interesting results, the second one does. For example, in mathematics, scientific schools exhibit an advantage in mathematical function and natural numbers. On the other side, human science schools have advantages in poetical and argumentative texts with respect to, respectively, scientific and classic and linguistic, while language texts advantages this last with respect to human sciences. The Italian educational system differentiates academic schools for different specific teaching content. Probably, these various specifications transfer to pupils different meanings of content according to specific school subject. Indeed, we can note that DIF items concern the specific contents of schools: mathematical function and natural numbers for scientific schools, poetical and argumentative texts for human science schools and language texts

for classic and linguistic schools. In conclusion, test takers should be quite satisfied to INVALSI results among pupils from academic schools. Possible test improvements should regard the fortification of the teach subjects not specific of the schools in order to make the test completely fair.

6.2 Limitations and further developments

Now, we consider the limits of our research, useful for possible future developments. First of all, our simulation strategy are restricted to the assumption that the conditional dependence between item performance and grouping variable remains constant (uniform DIF). Nevertheless, the conditional logistic regression allows to detect nonuniform DIF. Indeed, as said in section 3.3.3, it is possible to detect nonuniform DIF adding the interaction term between pupils' total score and grouping variable to models 24, 25 and 26. We restricted our analysis to uniform DIF because our simulation design allows to generate only uniform DIF items. In the future a different simulation strategy should be considered for assessing how the new methodology performs in the presence of nonuniform DIF items.

Our second goal concerned the evaluation of INVALSI tests. The data presents information on each item with dichotomous variable where 1 refers to correct answer and 0 refers to wrong answer. Despite the nature of INVALSI data, there exists standardized tests, both in educational field and others, that contain and trait outcome variables with more than two possible responses (polytomous variables). An algorithm of conditional logistic regression for polytomous response variable has not been developed yet, so it is possible apply this methodology only for dichotomous response variables (Liu et al., 2016).

The new methodology is computed for comparing two groups at a time. Therefore, our simulation study considers only groups with two possible allocations. In addition, our application involves chosen group aggregation in order to maintain dichotomous groups. Nevertheless, Svetina and Rutkowski (2014) and Magis at al. (2011) proposed Generalized Logistic Regression, while Woods et al. (2013) improved version of Lords χ^2 Wald test for DIF detection analysis in multiple groups. Finch (2016) assessed Generalized Mantel-Haenszel test, Generalized Logistic Regression, Lords χ^2 test for multiple group, showing that the first method outperforms the others as an optimal combination of type I error control and power. Hence, future researches should integrate matching techniques

to generalized DIF detection methods for multiple groups in order to improve the bias detection.

Finally, from an applicative point of view, future developments should concern DIF detection analysis among pupils from, not only academic, but also technical and vocational schools. In addition, the new methodology helps attribute possible DIF items to particular secondary school track. Probably, matching will be less precise than matching for academic tracks because pupils from academic schools are more similar, in term of covariates distribution, than pupils from technical and vocational schools. Nevertheless, this comparison is very interesting because INVALSI results present systematic performance gap between academic, technical vocational schools (INVALSI 2017c, 2016). Therefore, an evaluation of INVALSI instrument validity is fundamental in order to robust the results and make decision based on the standardized test.

A. Variables check

Table A1: INVALSI sample: school tracking composition.

	N	%
Academic	14185	0.44
Technical	10873	0.34
Vocational	7193	0.22

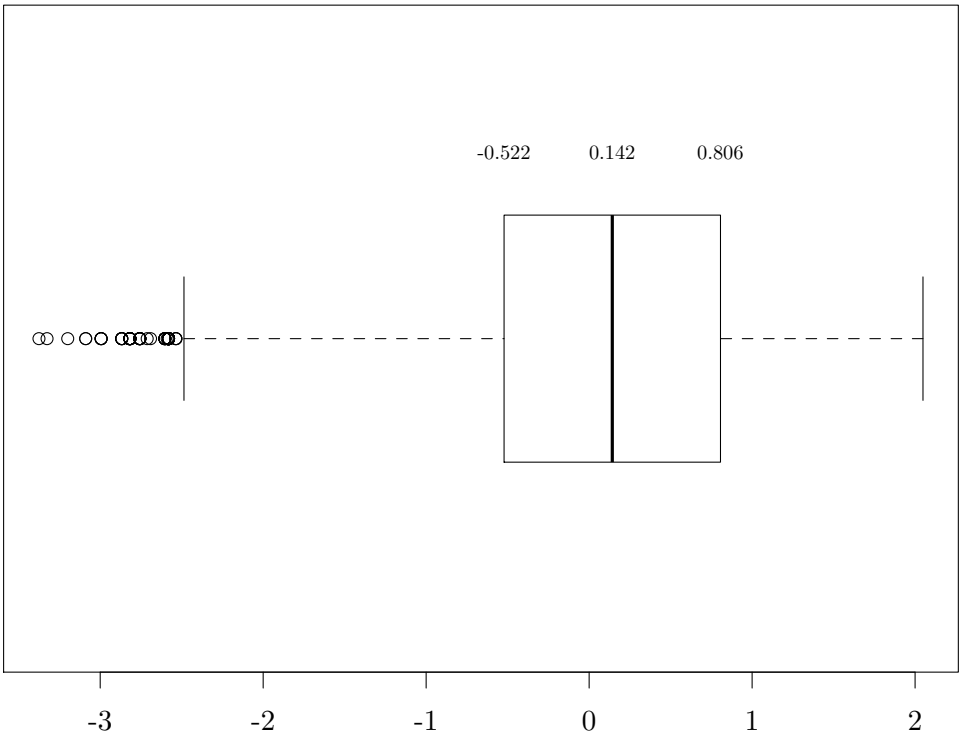


Figure A1: Box plot of pupils' ESCS index of INVALSI sample.

Table A2: INVALSI sample: maths score means and standard deviations by covariates.

	Means	Standard Deviations
Gender		
Male	50.74	23.07
Female	43.13	21.11
Citizen		
Italian	47.45	22.56
Not Italian	42.04	20.59
Aspiration		
University degree	51.87	22.66
Not University degree	39.64	19.99
Area		
North	52.46	21.69
Middle	45.89	22.08
South	40.17	21.69
ESCS		
I quartile	42.14	21.28
II quartile	45.79	22.08
III quartile	48.56	22.45
IV quartile	52.77	22.77

Table A3: Pupils' ESCS index composition among other variables.

	I quartile	II quartile	III quartile	IV quartile
Male Italian HighAspiration North	0.156	0.212	0.255	0.376
Male Italian HighAspiration Middle	0.153	0.210	0.247	0.389
Male Italian HighAspiration South	0.224	0.230	0.247	0.299
Male notItalian HighAspiration North	0.404	0.174	0.230	0.191
Male notItalian HighAspiration Middle	0.368	0.241	0.203	0.188
Male notItalian HighAspiration South	0.373	0.229	0.169	0.229
Male Italian LowAspiration North	0.319	0.282	0.236	0.162
Male Italian LowAspiration Middle	0.348	0.246	0.249	0.156
Male Italian LowAspiration South	0.449	0.256	0.191	0.104
Male notItalian LowAspiration North	0.503	0.245	0.156	0.095
Male notItalian LowAspiration Middle	0.537	0.245	0.116	0.102
Male notItalian LowAspiration South	0.534	0.329	0.091	0.045
Female Italian HighAspiration North	0.198	0.250	0.263	0.288
Female Italian HighAspiration Middle	0.187	0.211	0.277	0.325
Female Italian HighAspiration South	0.269	0.247	0.211	0.272
Female notItalian HighAspiration North	0.499	0.218	0.144	0.140
Female notItalian HighAspiration Middle	0.494	0.196	0.190	0.119
Female notItalian HighAspiration South	0.429	0.248	0.143	0.181
Female Italian LowAspiration North	0.342	0.286	0.230	0.142
Female Italian LowAspiration Middle	0.372	0.265	0.212	0.151
Female Italian LowAspiration South	0.524	0.265	0.212	0.151
Female notItalian LowAspiration North	0.593	0.217	0.147	0.043
Female notItalian LowAspiration Middle	0.572	0.221	0.137	0.069
Female notItalian LowAspiration South	0.061	0.167	0.167	0.061

Table A4: Composition of pupils' ESCS index among other variables (INVALSI sample).

		I quantile	II quantile	III quantile	IV quantile
Gender					
	Male (%)	0.29	0.24	0.23	0.24
	Female (%)	0.30	0.25	0.22	0.23
Citizen					
	Italian (%)	0.28	0.25	0.23	0.24
	Not Italian (%)	0.50	0.22	0.16	0.12
Aspiration					
	University degree (%)	0.22	0.23	0.24	0.30
	Not University degree (%)	0.41	0.26	0.20	0.13
Area					
	North (%)	0.26	0.25	0.24	0.25
	Middle (%)	0.28	0.23	0.24	0.25
	South (%)	0.34	0.25	0.20	0.21

Table A5: Composition of pupils' ESCS index among other variables (simulation N=500).

		I quantile	II quantile	III quantile	IV quantile
Gender					
	Male (%)	0.27	0.24	0.23	0.26
	Female (%)	0.28	0.25	0.22	0.25
Citizen					
	Italian (%)	0.26	0.25	0.23	0.26
	Not Italian (%)	0.37	0.27	0.16	0.20
Aspiration					
	University degree (%)	0.21	0.23	0.25	0.31
	Not University degree (%)	0.37	0.26	0.19	0.16
Area					
	North (%)	0.24	0.26	0.23	0.27
	Middle (%)	0.25	0.23	0.25	0.27
	South (%)	0.33	0.25	0.20	0.22

Table A6: Composition of pupils' ESCS index among other variables (simulation N=1000).

		I quantile	II quantile	III quantile	IV quantile
Gender					
	Male (%)	0.28	0.24	0.23	0.25
	Female (%)	0.29	0.25	0.22	0.24
Citizen					
	Italian (%)	0.27	0.25	0.23	0.25
	Not Italian (%)	0.44	0.23	0.16	0.17
Aspiration					
	University degree (%)	0.22	0.23	0.25	0.30
	Not University degree (%)	0.39	0.27	0.19	0.15
Area					
	North (%)	0.25	0.25	0.24	0.26
	Middle (%)	0.28	0.22	0.24	0.26
	South (%)	0.33	0.25	0.20	0.22

Table A7: Composition of pupils' ESCS index among other variables (simulation N=2000).

		I quantile	II quantile	III quantile	IV quantile
Gender					
	Male (%)	0.28	0.25	0.22	0.25
	Female (%)	0.30	0.25	0.22	0.23
Citizen					
	Italian (%)	0.27	0.25	0.23	0.25
	Not Italian (%)	0.47	0.23	0.15	0.15
Aspiration					
	University degree (%)	0.22	0.23	0.24	0.31
	Not University degree (%)	0.39	0.27	0.20	0.14
Area					
	North (%)	0.26	0.25	0.24	0.25
	Middle (%)	0.29	0.23	0.23	0.25
	South (%)	0.33	0.25	0.21	0.21

B. Propensity score matching for simulations

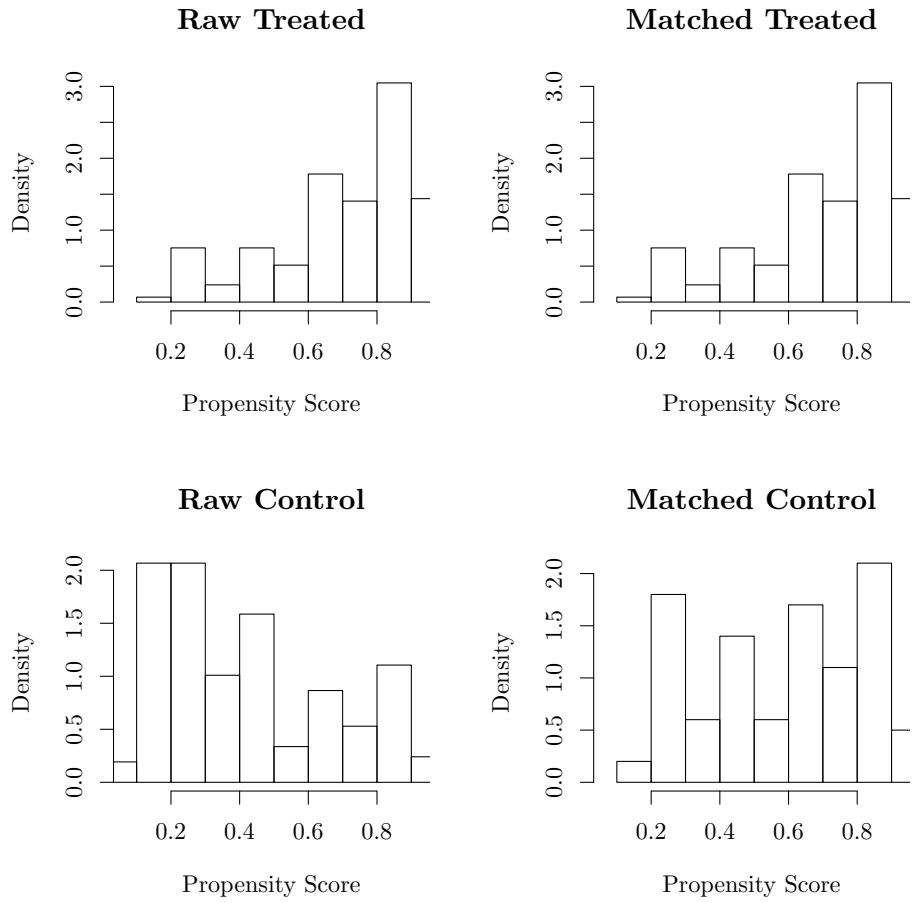


Figure B1: Propensity score distributions before and after matching using greedy matching (N=500).

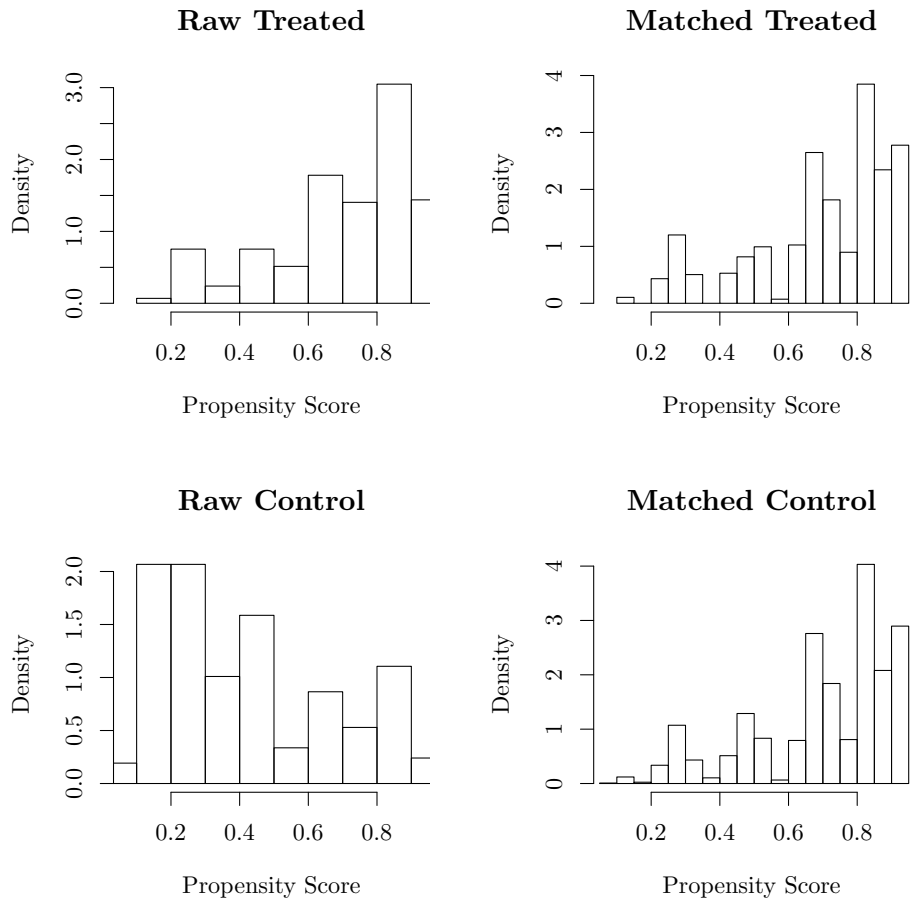


Figure B2: Propensity score distributions before and after matching using full matching (N=500).

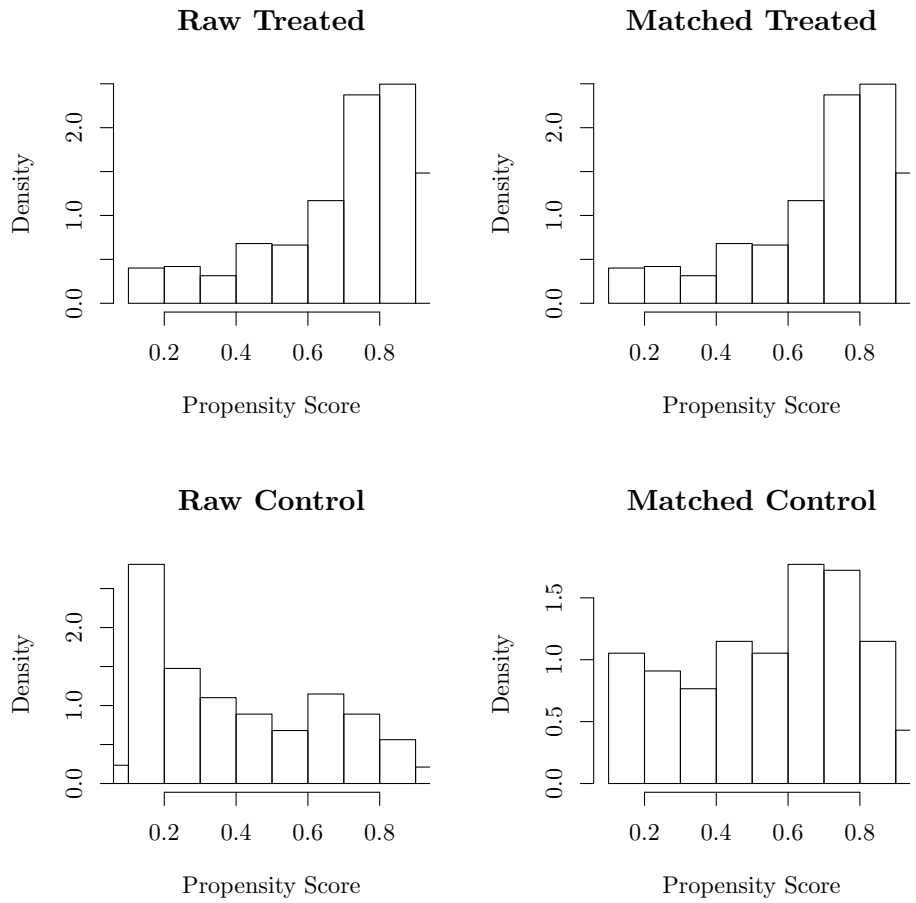


Figure B3: Propensity score distributions before and after matching using greedy matching (N=1000).

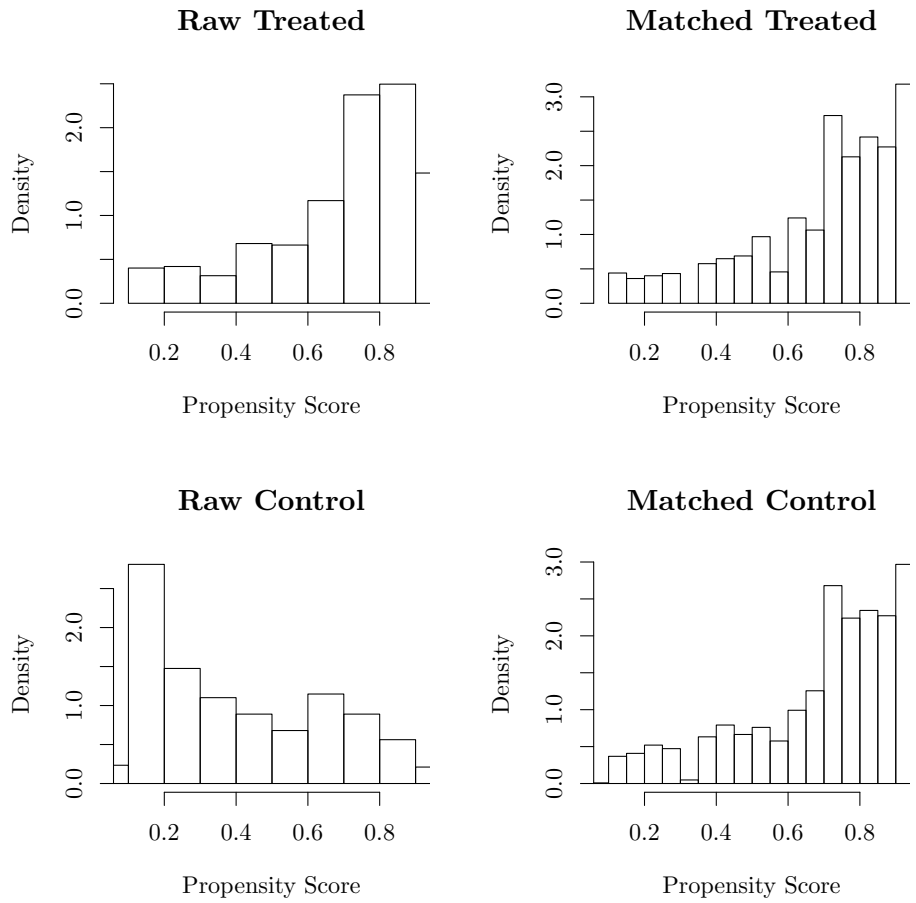


Figure B4: Propensity score distributions before and after matching using full matching (N=1000).

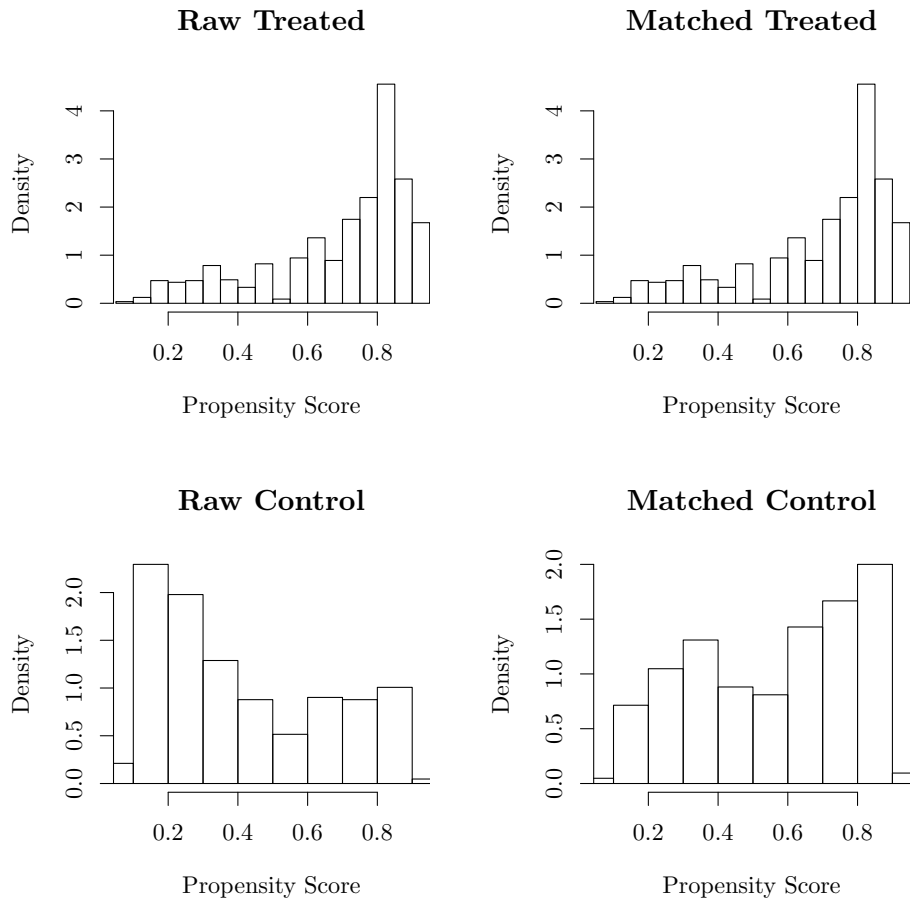


Figure B5: Propensity score distributions before and after matching using greedy matching (N=2000).

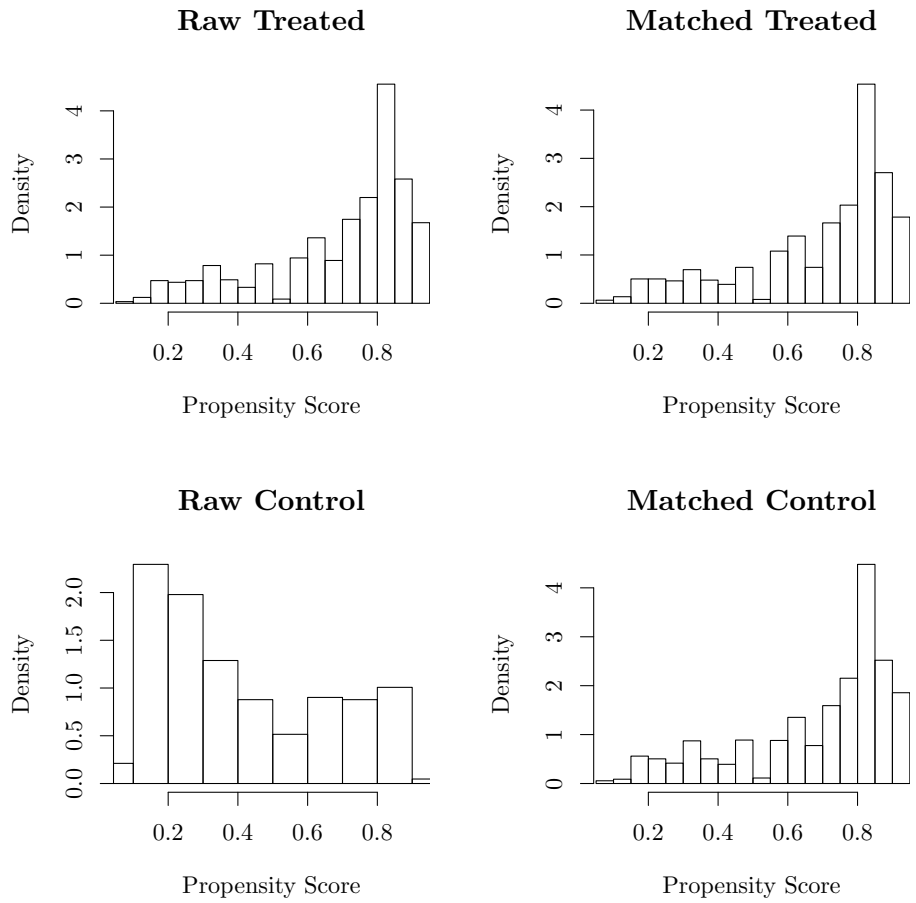


Figure B6: Propensity score distributions before and after matching using full matching (N=2000).

Table B1: Percentage of Bias Reduction (PBR) using greedy matching (N=500).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.701	0.420	0.281	0.701	0.002	99.9
Gender (Male)	0.394	0.625	-0.231	0.428	-0.034	85.2
Citizen (Not Italian)	0.065	0.125	-0.060	0.038	0.027	54.3
Aspiration (High)	0.794	0.351	0.444	0.805	-0.010	97.7
Middle	0.202	0.221	-0.019	0.130	0.072	-276.5
South	0.318	0.346	0.028	0.366	-0.048	-73.3
II quartile	0.239	0.259	-0.020	0.274	-0.034	-72.2
III quartile	0.253	0.187	0.066	0.257	-0.003	94.8
IV quartile	0.332	0.144	0.188	0.308	0.024	87.2

Table B2: Percentage of Bias Reduction (PBR) using full matching (N=500).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.701	0.420	0.281	0.692	0.008	99.9
Gender (Male)	0.394	0.625	-0.231	0.440	-0.046	85.9
Citizen (Not Italian)	0.065	0.125	0.065	0.038	0.010	83.4
Aspiration (High)	0.794	0.351	0.444	0.797	-0.003	99.4
Middle	0.202	0.221	-0.019	0.201	0.001	-191.5
South	0.318	0.346	0.028	0.336	-0.018	-47.1
II quartile	0.239	0.259	-0.020	0.243	-0.003	5.4
III quartile	0.253	0.187	0.066	0.294	-0.041	83.3
IV quartile	0.332	0.144	0.188	0.381	0.051	87.2

Table B3: Percentage of Bias Reduction (PBR) using greedy matching (N=1000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.698	0.406	0.292	0.697	0.001	99.9
Gender (Male)	0.363	0.660	-0.297	0.363	0.000	100.0
Citizen (Not Italian)	0.073	0.112	-0.039	0.065	0.009	77.7
Aspiration (High)	0.809	0.342	0.468	0.809	0.000	100.0
Middle	0.202	0.220	-0.018	0.187	0.016	11.2
South	0.328	0.333	-0.004	0.337	-0.009	-95.9
II quartile	0.232	0.260	-0.028	0.232	0.000	100.0
III quartile	0.246	0.201	0.045	0.237	0.009	80.5
IV quartile	0.305	0.162	0.144	0.305	0.000	100.0

Table B4: Percentage of Bias Reduction (PBR) using full matching (N=1000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.698	0.406	0.292	0.697	0.002	99.9
Gender (Male)	0.363	0.660	-0.297	0.363	0.000	100.0
Citizen (Not Italian)	0.073	0.112	-0.039	0.069	0.004	89.3
Aspiration (High)	0.809	0.342	0.468	0.809	0.000	100.0
Middle	0.202	0.220	-0.018	0.267	-0.084	-376.8
South	0.328	0.333	-0.004	0.336	-0.008	-83.2
II quartile	0.232	0.260	-0.028	0.235	-0.003	88.9
III quartile	0.246	0.201	0.045	0.237	0.009	80.6
IV quartile	0.305	0.162	0.144	0.306	-0.001	99.6

Table B5: Percentage of Bias Reduction (PBR) using greedy matching (N=2000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.690	0.416	0.274	0.690	0.000	99.9
Gender (Male)	0.387	0.628	-0.240	0.384	-0.003	98.5
Citizen (Not Italian)	0.067	0.121	-0.053	0.054	0.013	75.5
Aspiration (High)	0.805	0.348	0.458	0.805	0.000	100.0
Middle	0.211	0.208	0.002	0.168	0.043	-1461.4
South	0.339	0.318	0.020	0.378	-0.039	-95.7
II quartile	0.236	0.255	-0.020	0.236	0.000	100.0
III quartile	0.242	0.204	0.038	0.234	0.008	79.3
IV quartile	0.314	0.140	0.174	0.310	0.004	97.5

Table B6: Percentage of Bias Reduction (PBR) using full matching (N=2000).

	Before Matching			After Matching		
	Mean Treated	Mean Control	Mean Difference	Mean Control	Mean Difference	PBR (%)
Distance	0.690	0.416	0.274	0.690	0.000	99.9
Gender (Male)	0.387	0.628	-0.240	0.378	-0.009	96.2
Citizen (Not Italian)	0.067	0.121	-0.053	0.055	0.012	77.8
Aspiration (High)	0.805	0.348	0.458	0.801	0.005	98.9
Middle	0.211	0.208	0.002	0.169	0.042	-1432.4
South	0.339	0.318	0.020	0.382	-0.043	-116.9
II quartile	0.236	0.255	-0.020	0.237	-0.001	94.5
III quartile	0.242	0.204	0.038	0.238	0.004	90.5
IV quartile	0.314	0.140	0.174	0.310	0.005	97.1

C. Simulation results $\beta \sim N(0,1)$

Table C1: False alarm rates of DIF methods: no biased items.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.056	0.070	0.060	0.055
J=40	0.043	0.057	0.044	0.040
J=60	0.052	0.052	0.040	0.035
N=1000				
J=20	0.048	0.065	0.062	0.055
J=40	0.050	0.053	0.050	0.040
J=60	0.050	0.049	0.044	0.033
N=2000				
J=20	0.056	0.073	0.059	0.053
J=40	0.048	0.056	0.050	0.041
J=60	0.048	0.049	0.041	0.034

Table C2: False alarm rates of DIF methods: 10% biased items and δ=0.4.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.064	0.074	0.057	0.050
J=40	0.058	0.059	0.045	0.044
J=60	0.055	0.058	0.044	0.038
N=1000				
J=20	0.057	0.062	0.051	0.038
J=40	0.054	0.056	0.051	0.040
J=60	0.053	0.055	0.049	0.038
N=2000				
J=20	0.049	0.068	0.056	0.038
J=40	0.047	0.066	0.057	0.040
J=60	0.048	0.059	0.056	0.043

Table C3: False alarm rates of DIF methods: 10% biased items and δ=0.8.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.069	0.082	0.056	0.045
J=40	0.065	0.070	0.051	0.050
J=60	0.063	0.071	0.051	0.045
N=1000				
J=20	0.072	0.077	0.067	0.056
J=40	0.064	0.076	0.067	0.060
J=60	0.061	0.073	0.075	0.061
N=2000				
J=20	0.079	0.099	0.103	0.079
J=40	0.069	0.099	0.099	0.080
J=60	0.080	0.091	0.095	0.078

Table C4: False alarm rates of DIF methods: 20% biased items and δ=0.4.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.068	0.080	0.061	0.051
J=40	0.068	0.074	0.052	0.050
J=60	0.062	0.070	0.052	0.045
N=1000				
J=20	0.069	0.080	0.067	0.054
J=40	0.066	0.076	0.072	0.061
J=60	0.062	0.071	0.071	0.062
N=2000				
J=20	0.077	0.106	0.110	0.076
J=40	0.071	0.101	0.101	0.078
J=60	0.082	0.091	0.099	0.078

Table C5: False alarm rates of DIF methods: 20% biased items and δ=0.8.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.101	0.131	0.102	0.088
J=40	0.095	0.119	0.084	0.087
J=60	0.092	0.117	0.084	0.083
N=1000				
J=20	0.113	0.155	0.149	0.124
J=40	0.107	0.144	0.152	0.139
J=60	0.113	0.139	0.147	0.139
N=2000				
J=20	0.177	0.276	0.302	0.269
J=40	0.159	0.240	0.267	0.249
J=60	0.161	0.246	0.285	0.269

Table C6: Power rates of DIF methods: 10% biased items and δ=0.4.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.250	0.410	0.355	0.315
J=40	0.210	0.342	0.302	0.285
J=60	0.248	0.368	0.325	0.303
N=1000				
J=20	0.395	0.535	0.665	0.630
J=40	0.347	0.485	0.587	0.545
J=60	0.377	0.530	0.610	0.582
N=2000				
J=20	0.630	0.895	0.895	0.910
J=40	0.572	0.830	0.842	0.865
J=60	0.532	0.843	0.838	0.886

Table C7: Power rates of DIF methods: 10% biased items and δ=0.8.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.680	0.930	0.900	0.885
J=40	0.657	0.875	0.837	0.885
J=60	0.673	0.880	0.828	0.853
N=1000				
J=20	0.890	1.000	0.995	1.000
J=40	0.865	0.977	0.992	0.995
J=60	0.865	0.977	0.982	0.982
N=2000				
J=20	0.995	1.000	0.950	1.000
J=40	0.985	1.000	0.945	1.000
J=60	0.977	0.998	0.923	1.000

Table C8: Power rates of DIF methods: 20% biased items and δ=0.4.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.210	0.350	0.312	0.277
J=40	0.202	0.299	0.346	0.234
J=60	0.203	0.297	0.237	0.252
N=1000				
J=20	0.302	0.500	0.570	0.532
J=40	0.305	0.415	0.487	0.484
J=60	0.297	0.438	0.508	0.477
N=2000				
J=20	0.477	0.795	0.785	0.822
J=40	0.484	0.726	0.795	0.766
J=60	0.467	0.738	0.768	0.782

Table C9: Power rates of DIF methods: 20% biased items and δ=0.8.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.595	0.852	0.812	0.797
J=40	0.565	0.811	0.750	0.761
J=60	0.584	0.809	0.743	0.769
N=1000				
J=20	0.830	0.960	0.985	0.980
J=40	0.806	0.945	0.957	0.964
J=60	0.801	0.952	0.956	0.968
N=2000				
J=20	0.975	1.000	0.915	1.000
J=40	0.940	0.995	0.897	0.998
J=60	0.951	0.997	0.882	0.998

D. Simulation results $\beta \sim U(-2, +2)$

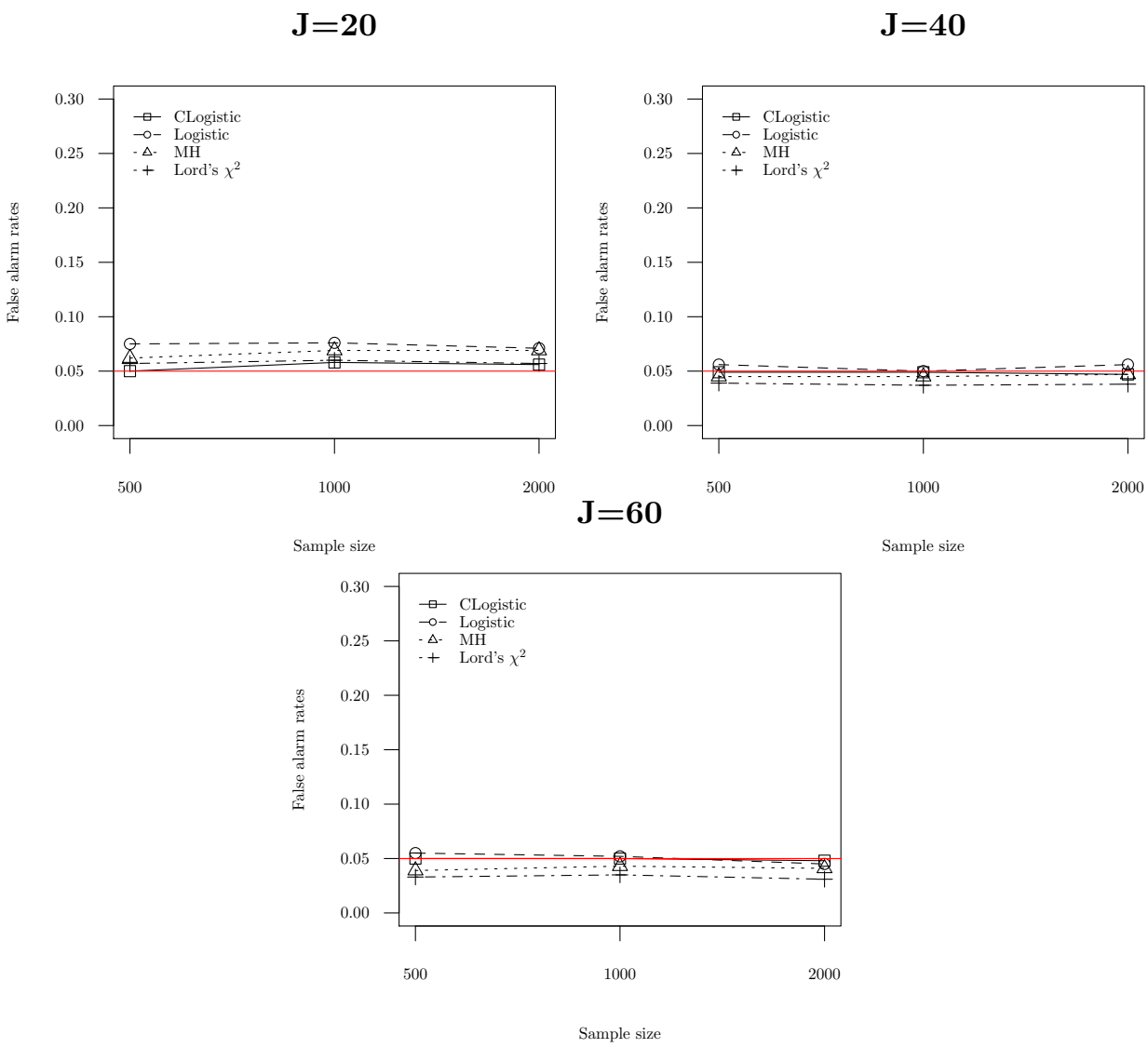


Figure D1: False alarm rates of DIF methods: no biased items $\beta \sim U(-2, +2)$.

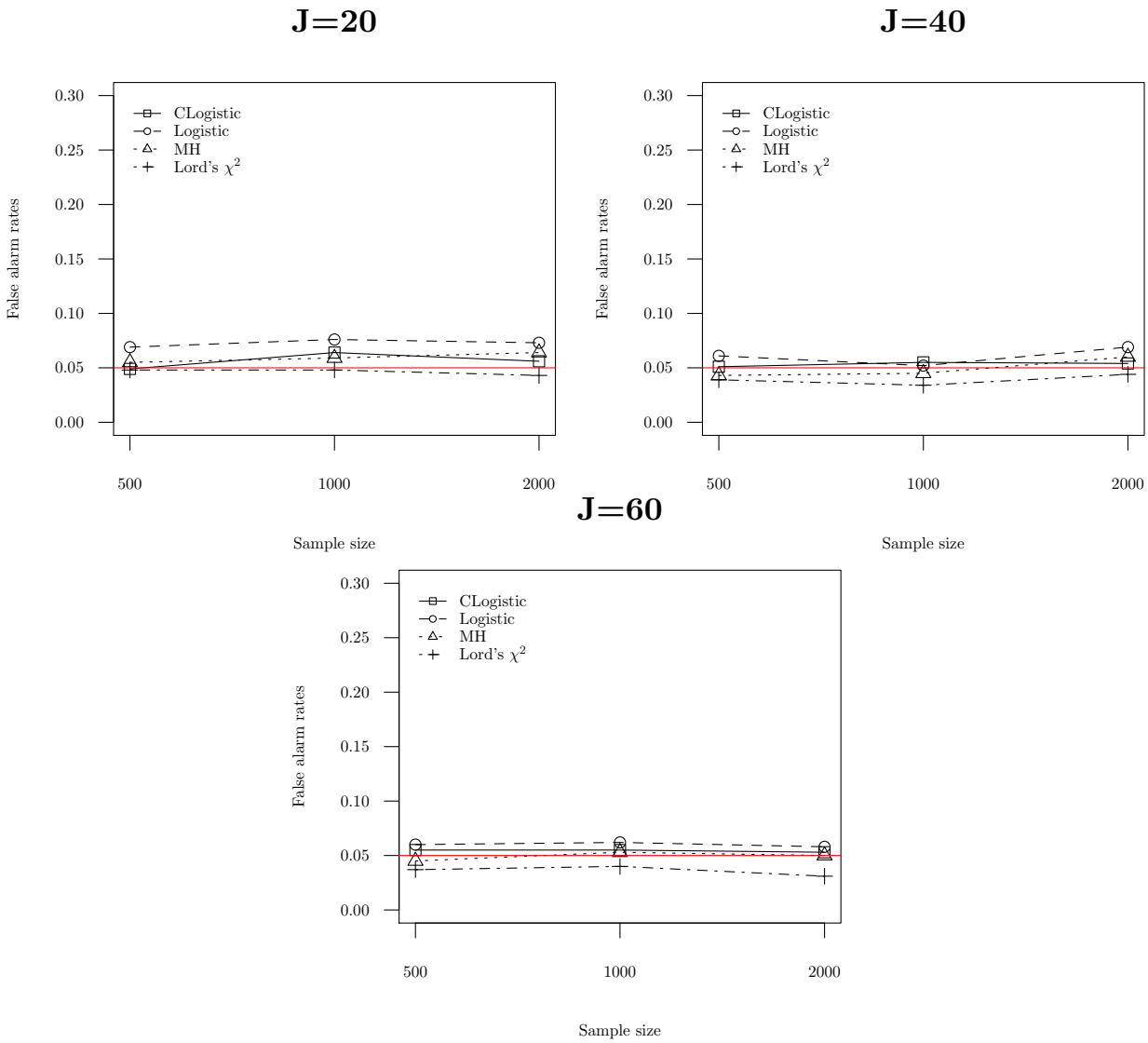


Figure D2: False alarm rates of DIF methods: 10% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

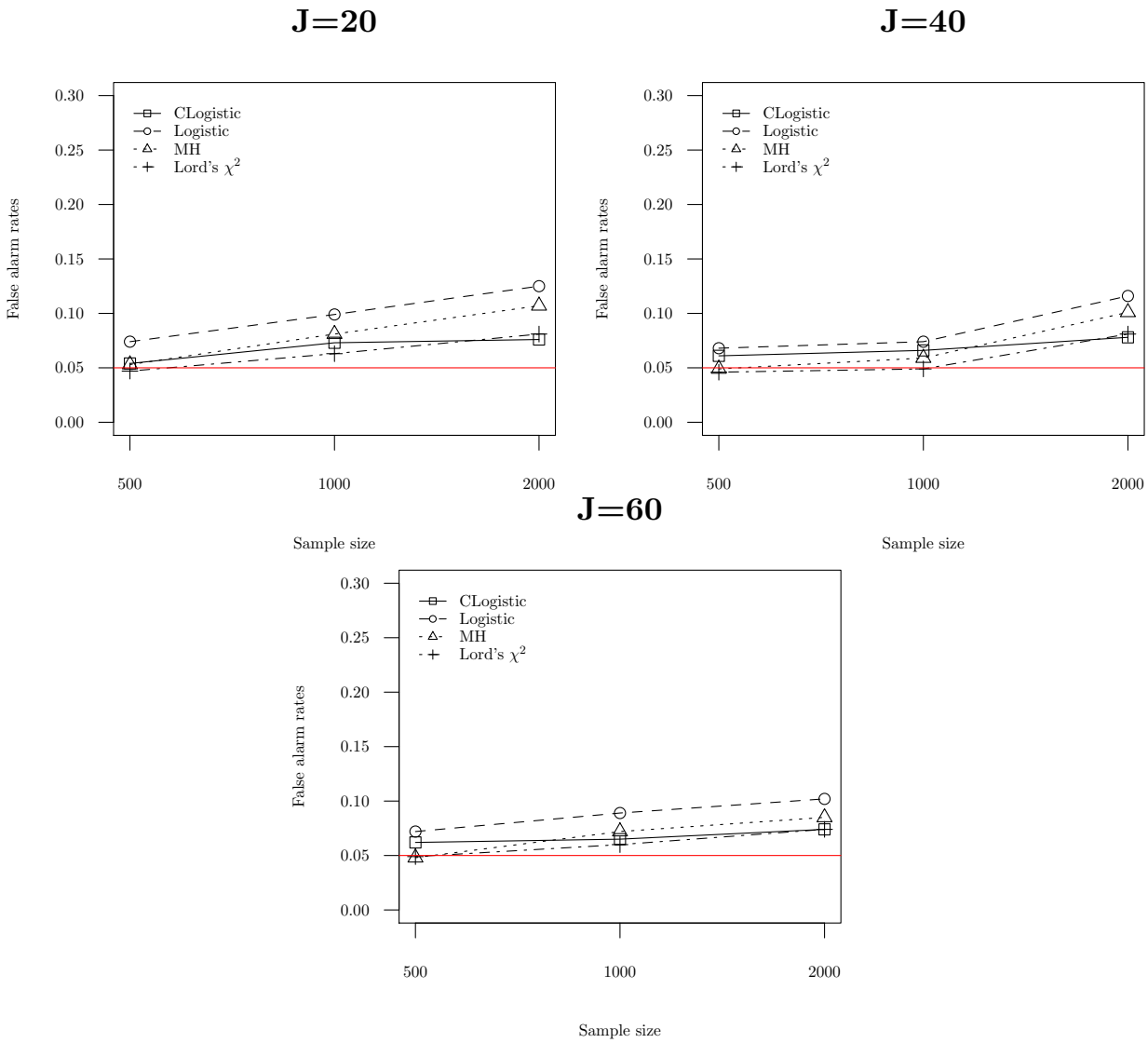


Figure D3: False alarm rates of DIF methods: 10% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

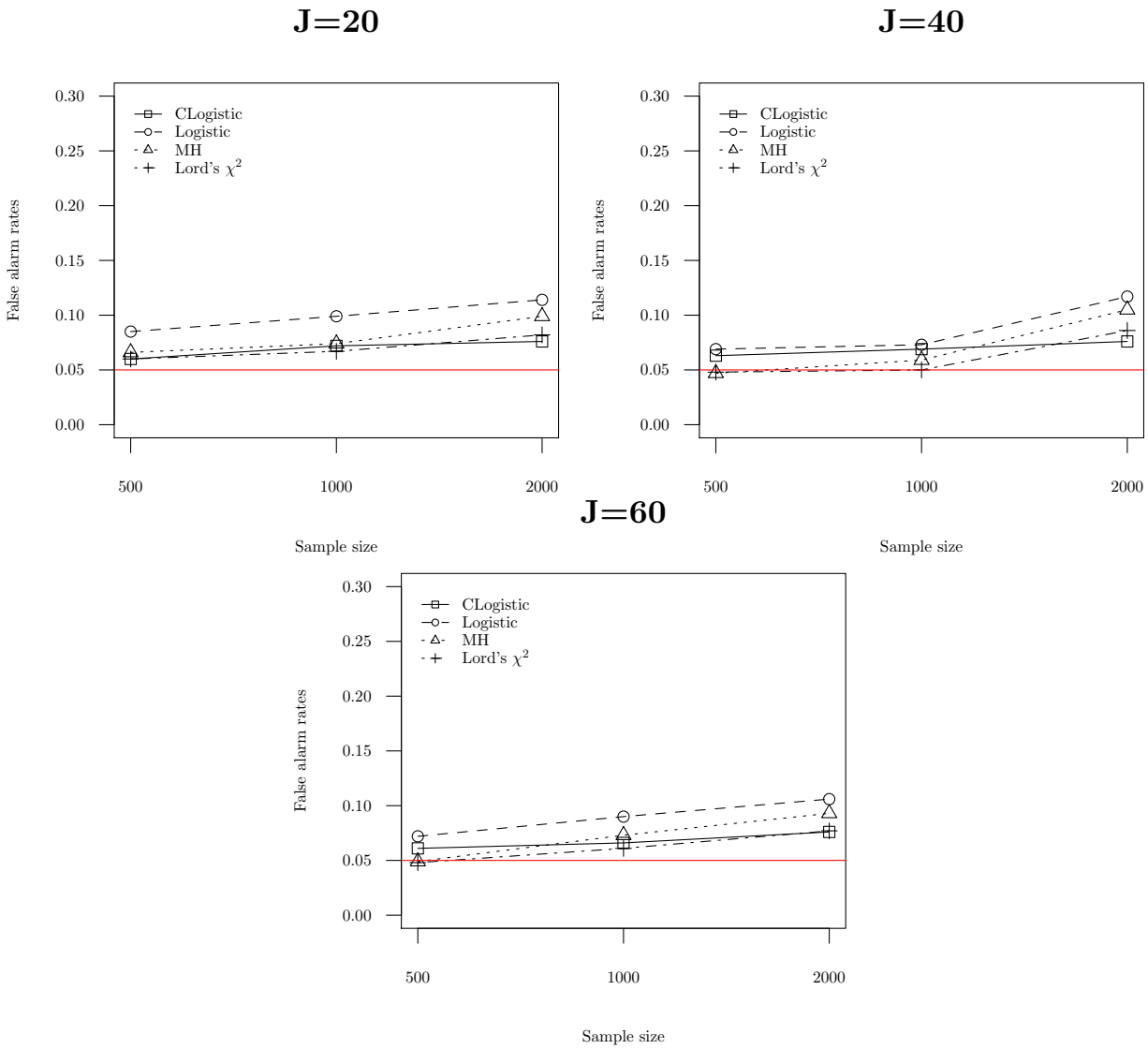


Figure D4: False alarm rates of DIF methods: 20% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

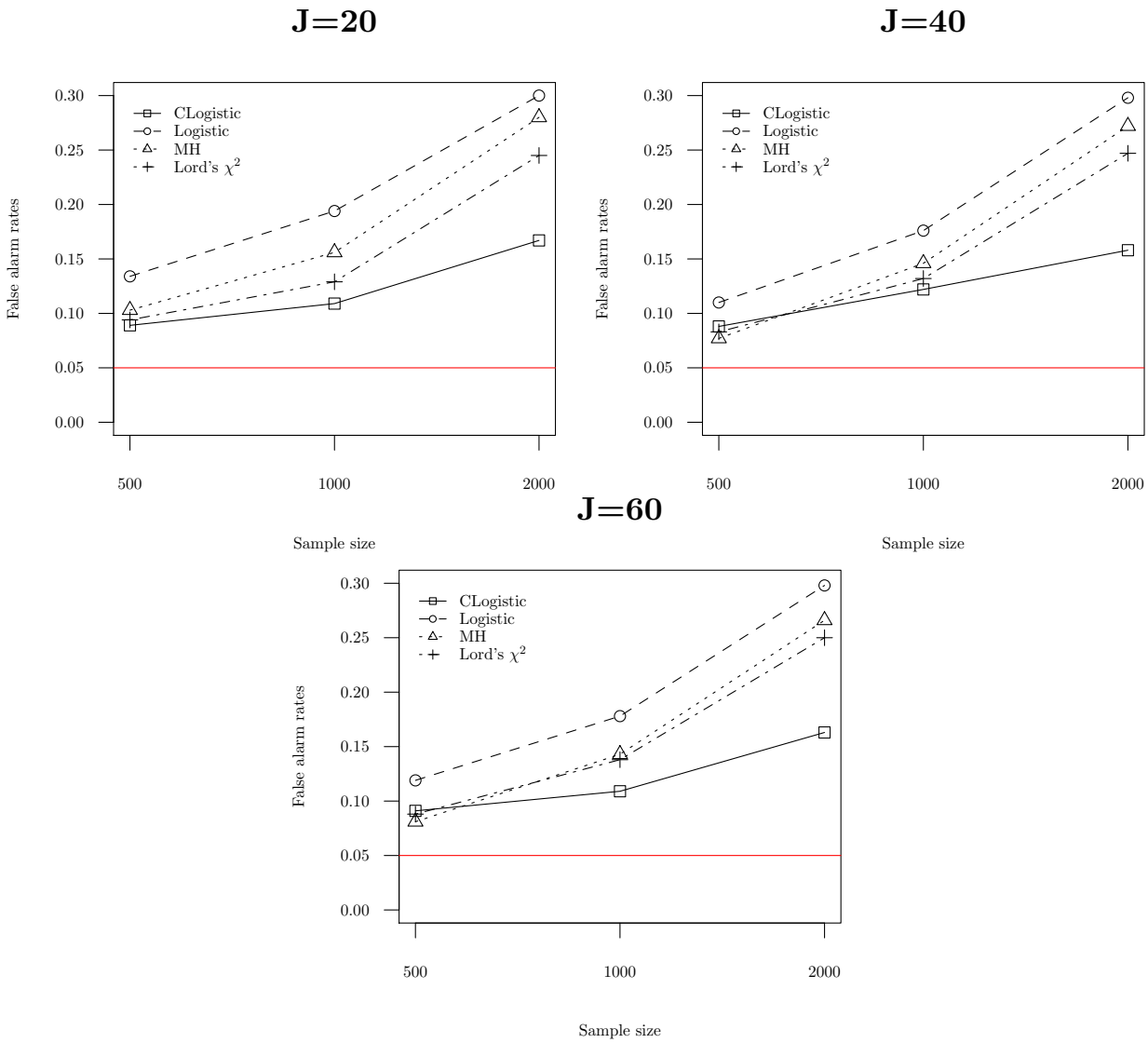


Figure D5: False alarm rates of DIF methods: 20% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

Table D1: False alarm rates of DIF methods: no biased items $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.050	0.075	0.062	0.057
J=40	0.049	0.056	0.045	0.039
J=60	0.050	0.055	0.039	0.033
N=1000				
J=20	0.058	0.076	0.069	0.060
J=40	0.049	0.050	0.045	0.037
J=60	0.050	0.052	0.043	0.035
N=2000				
J=20	0.056	0.071	0.069	0.057
J=40	0.047	0.056	0.047	0.038
J=60	0.048	0.045	0.041	0.031

Table D2: False alarm rates of DIF methods: 10% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.049	0.069	0.055	0.048
J=40	0.051	0.061	0.043	0.039
J=60	0.055	0.060	0.045	0.037
N=1000				
J=20	0.064	0.076	0.059	0.048
J=40	0.055	0.052	0.045	0.034
J=60	0.055	0.062	0.053	0.040
N=2000				
J=20	0.056	0.073	0.064	0.043
J=40	0.054	0.069	0.060	0.044
J=60	0.053	0.058	0.050	0.031

Table D3: False alarm rates of DIF methods: 10% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.054	0.074	0.053	0.047
J=40	0.061	0.068	0.049	0.046
J=60	0.062	0.072	0.048	0.049
N=1000				
J=20	0.073	0.099	0.081	0.063
J=40	0.066	0.074	0.059	0.049
J=60	0.065	0.089	0.072	0.060
N=2000				
J=20	0.076	0.125	0.107	0.081
J=40	0.078	0.116	0.101	0.081
J=60	0.074	0.102	0.085	0.074

Table D4: False alarm rates of DIF methods: 20% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.060	0.085	0.066	0.060
J=40	0.063	0.069	0.047	0.048
J=60	0.061	0.072	0.049	0.048
N=1000				
J=20	0.072	0.099	0.074	0.067
J=40	0.069	0.073	0.059	0.050
J=60	0.066	0.090	0.073	0.061
N=2000				
J=20	0.076	0.114	0.099	0.082
J=40	0.078	0.117	0.105	0.086
J=60	0.076	0.106	0.093	0.077

Table D5: False alarm rates of DIF methods: 20% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.089	0.134	0.103	0.094
J=40	0.088	0.110	0.077	0.083
J=60	0.091	0.119	0.081	0.088
N=1000				
J=20	0.109	0.194	0.156	0.129
J=40	0.122	0.176	0.146	0.132
J=60	0.109	0.178	0.143	0.138
N=2000				
J=20	0.167	0.310	0.280	0.245
J=40	0.158	0.298	0.272	0.247
J=60	0.163	0.298	0.266	0.250

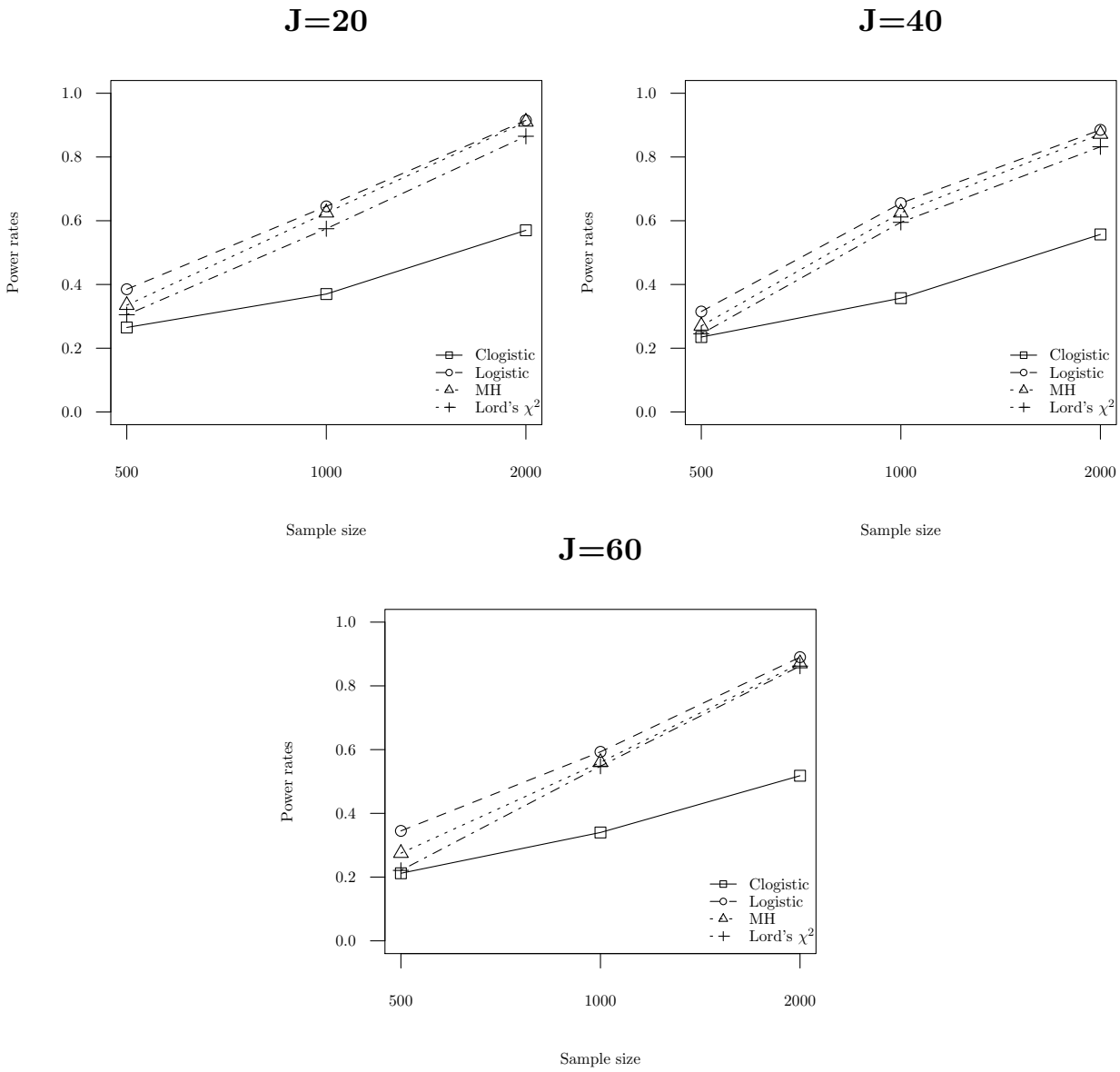


Figure D6: Power of DIF methods: 10% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

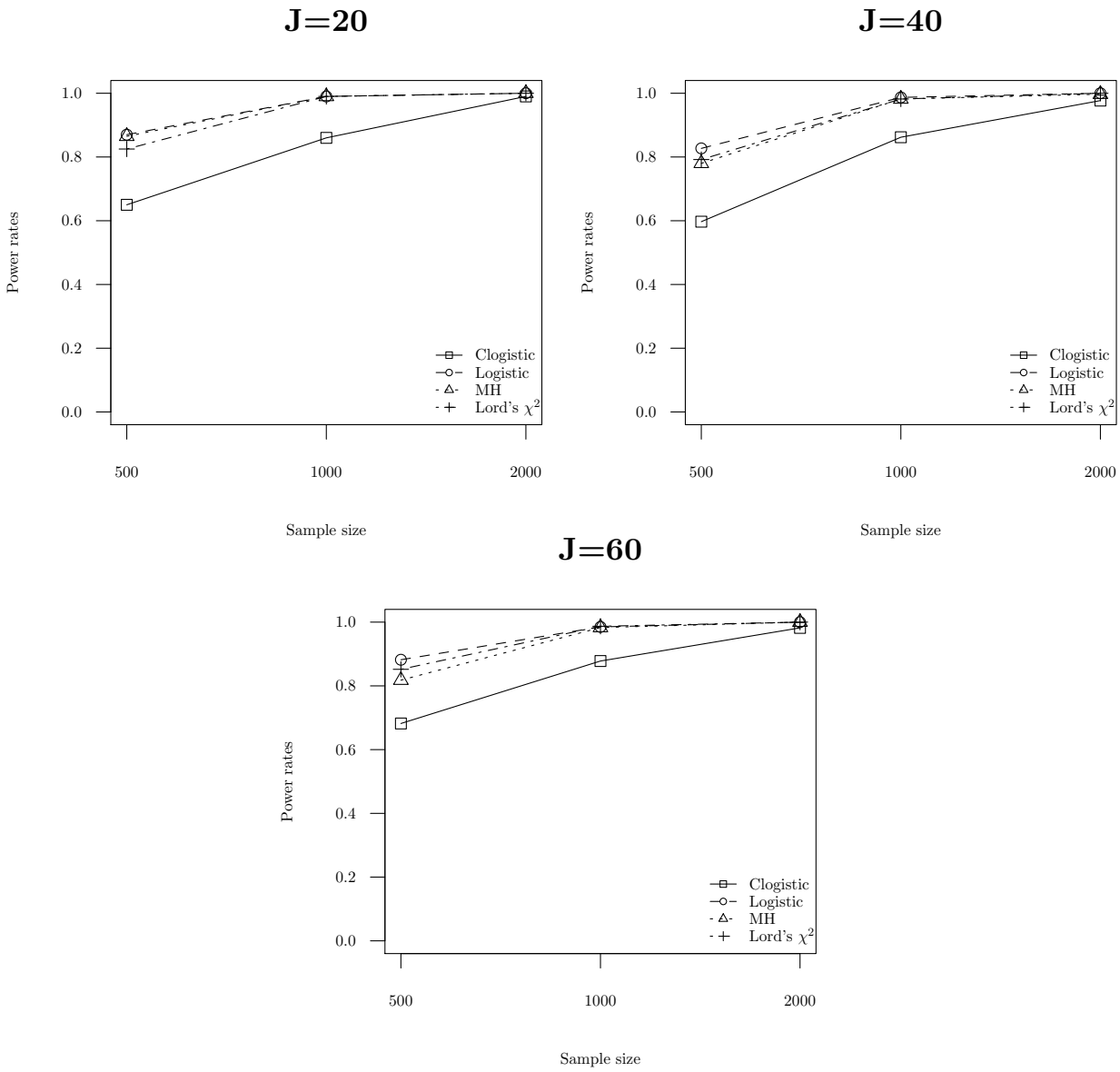


Figure D7: Power of DIF methods: 10% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

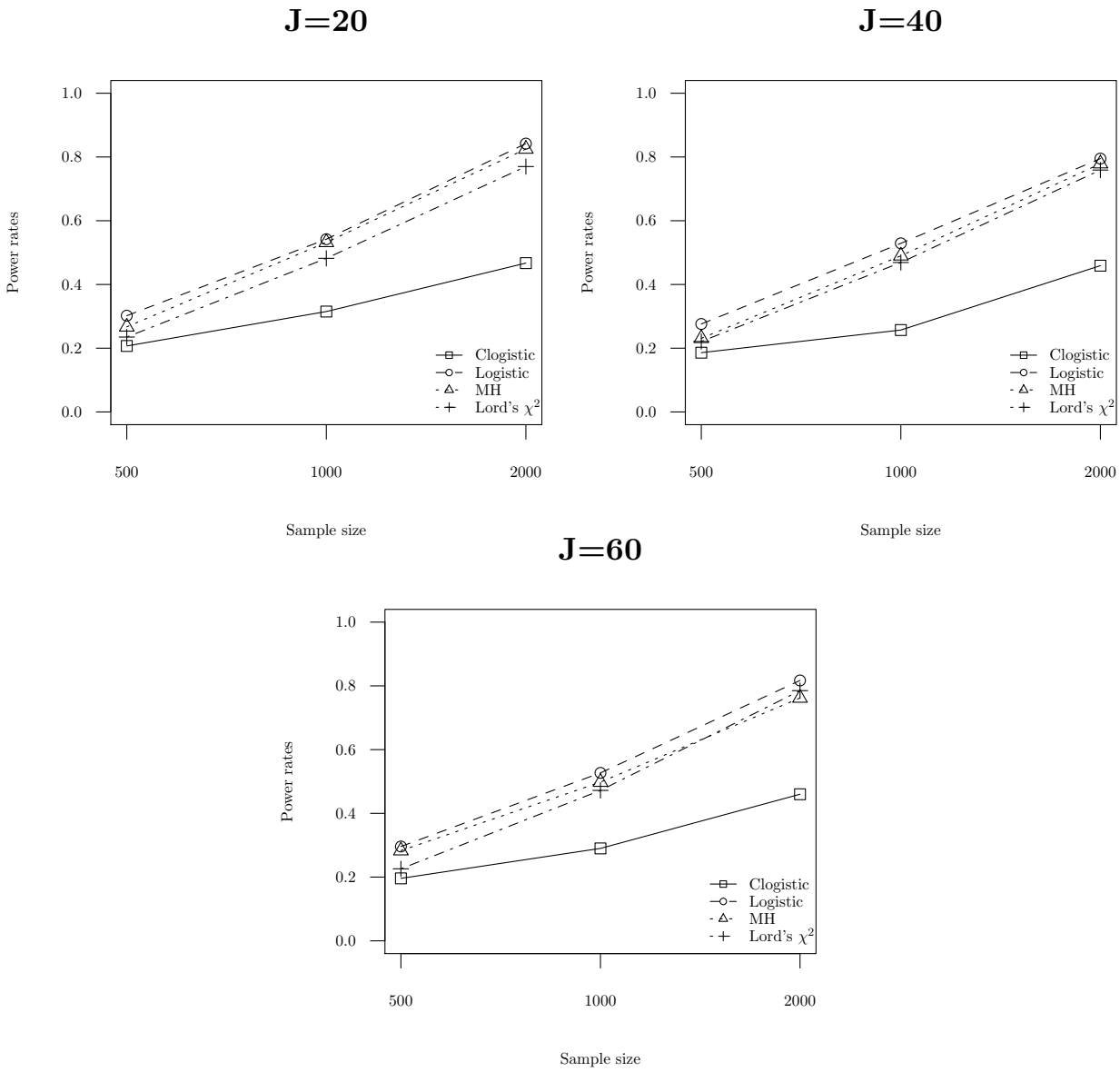


Figure D8: Power of DIF methods: 20% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

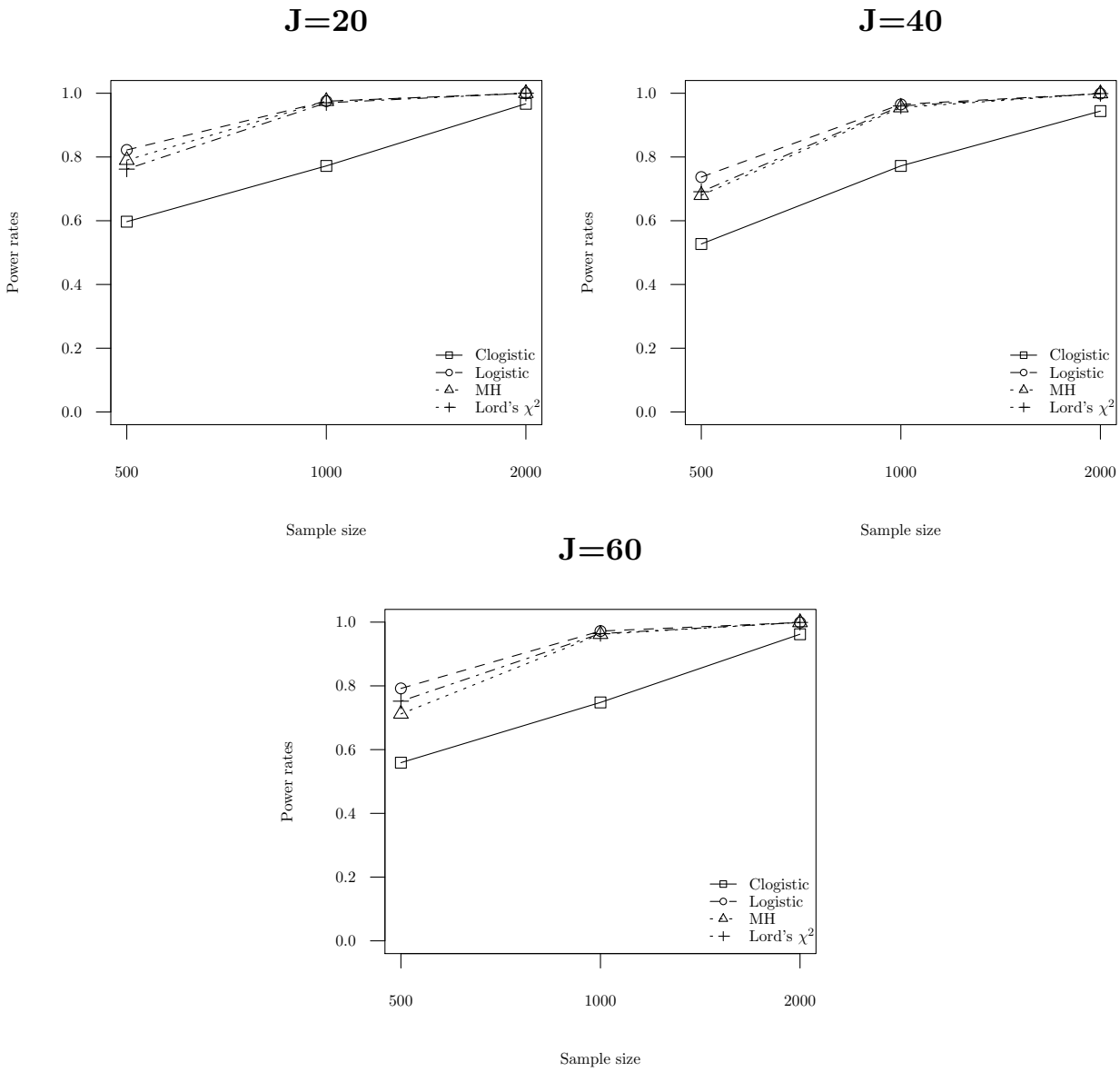


Figure D9: Power of DIF methods: 20% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

Table D6: Power of DIF methods: 10% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.265	0.385	0.335	0.305
J=40	0.235	0.315	0.270	0.245
J=60	0.212	0.345	0.275	0.221
N=1000				
J=20	0.370	0.645	0.625	0.575
J=40	0.357	0.655	0.625	0.595
J=60	0.340	0.593	0.560	0.548
N=2000				
J=20	0.570	0.915	0.910	0.865
J=40	0.557	0.885	0.872	0.832
J=60	0.518	0.890	0.870	0.862

Table D7: Power of DIF methods: 10% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.650	0.870	0.865	0.825
J=40	0.597	0.827	0.780	0.792
J=60	0.682	0.882	0.818	0.852
N=1000				
J=20	0.860	0.990	0.990	0.990
J=40	0.862	0.987	0.982	0.982
J=60	0.878	0.985	0.983	0.987
N=2000				
J=20	0.990	1.000	1.000	1.000
J=40	0.977	1.000	0.997	1.000
J=60	0.982	1.000	1.000	1.000

Table D8: Power of DIF methods: 20% biased items and δ=0.4 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.207	0.302	0.267	0.235
J=40	0.186	0.276	0.231	0.221
J=60	0.196	0.296	0.283	0.226
N=1000				
J=20	0.315	0.542	0.532	0.482
J=40	0.257	0.529	0.490	0.469
J=60	0.290	0.527	0.498	0.472
N=2000				
J=20	0.467	0.842	0.825	0.770
J=40	0.459	0.795	0.779	0.759
J=60	0.460	0.817	0.762	0.785

Table D9: Power of DIF methods: 20% biased items and δ=0.8 $\beta \sim U(-2, +2)$.

	CLogistic	Logistic	MH	Lord's χ^2
N=500				
J=20	0.597	0.822	0.790	0.762
J=40	0.527	0.737	0.680	0.691
J=60	0.559	0.792	0.712	0.752
N=1000				
J=20	0.772	0.975	0.975	0.970
J=40	0.772	0.965	0.955	0.960
J=60	0.748	0.972	0.962	0.964
N=2000				
J=20	0.967	1.000	1.000	1.000
J=40	0.944	0.999	0.999	0.999
J=60	0.962	0.999	0.999	0.999

Table D10: Percentage of false positive using effect size measure $\beta \sim U(-2, +2)$.

	J=20			*J=40*			*J=60*		
	DIF–U	$\Delta R^2 - U$	Reduction (%)	DIF–U	$\Delta R^2 - U$	Reduction	DIF–U	$\Delta R^2 - U$	Reduction (%)
No Bias									
N=500	5.00	0.05	100.00	4.90	0.42	11.67	5.02	0.37	13.57
N=1000	5.80	0.00	–	4.87	0.00	–	5.00	0.02	–
N=2000	5.65	0.00	–	4.75	0.00	–	4.80	0.00	–
10% Bias $\delta = 0.4$									
N=500	4.89	0.11	44.45	5.14	0.42	12.4	5.50	0.39	14.10
N=1000	6.44	0.00	–	5.50	0.00	–	5.52	0.02	–
N=2000	5.61	0.00	–	5.42	0.00	–	5.29	0.00	–
10% Bias $\delta = 0.8$									
N=500	5.44	0.33	16.48	6.14	0.50	12.28	6.17	0.42	14.69
N=1000	7.33	0.00	–	6.64	0.00	–	6.50	0.00	–
N=2000	7.56	0.00	–	7.78	0.00	–	7.44	0.00	–
20% Bias $\delta = 0.4$									
N=500	6.00	0.31	19.35	6.34	0.56	11.32	6.15	0.44	13.98
N=1000	7.19	0.00	–	6.87	0.00	–	6.25	0.00	–
N=2000	7.56	0.00	–	7.78	0.00	–	7.60	0.00	–
20% Bias $\delta = 0.8$									
N=500	8.87	0.69	12.85	8.88	0.87	10.21	9.08	0.75	12.11
N=1000	10.87	0.08	–	12.19	0.02	–	10.92	0.08	–
N=2000	16.75	0.00	–	15.84	0.00	–	16.29	0.00	–

– refers to reduction $\geq 100\%$ or reduction impossible to compute due to denominator equal to zero (no false positive detected).

Table D11: Percentage of correct identification using effect size measure $\beta \sim U(-2, +2)$.

	J=20			*J=40*			*J=60*		
	DIF–U	$\Delta R^2 - U$	Reduction (%)	DIF–U	$\Delta R^2 - U$	Reduction	DIF–U	$\Delta R^2 - U$	Reduction (%)
10% Bias $\delta = 0.4$									
N=500	26.50	4.50	5.89	23.50	4.00	5.87	21.17	3.00	7.06
N=1000	37.00	0.00	–	35.75	0.25	–	34.00	0.17	–
N=2000	57.00	0.00	–	55.75	0.00	–	51.83	0.00	–
10% Bias $\delta = 0.8$									
N=500	65.00	31.50	2.06	59.75	27.75	2.15	68.17	29.83	2.28
N=1000	86.00	10.00	8.60	86.25	11.00	7.84	87.83	11.17	7.86
N=2000	99.00	1.50	66.00	97.75	0.75	–	98.17	0.12	–
20% Bias $\delta = 0.4$									
N=500	20.75	2.25	9.22	18.62	2.25	8.28	19.58	2.92	6.70
N=1000	31.50	0.00	–	25.75	0.12	–	29.00	0.25	–
N=2000	46.75	0.00	–	45.87	0.00	–	46.00	0.00	–
20% Bias $\delta = 0.8$									
N=500	59.75	23.00	2.60	52.75	20.12	2.62	55.92	21.58	2.59
N=1000	77.25	5.50	14.04	77.25	5.00	15.45	78.42	6.92	11.33
N=2000	96.75	0.50	–	94.37	0.50	–	96.17	0.42	–

– refers to reduction $\geq 100\%$ or reduction impossible to compute due to denominator equal to zero (no correct identification detected).

E. DIF results

Table E1: Items format of maths and Italian language INVALSI tests 2016/2017.

Item	Format	**Item**	Format	**Item**	Format	**Item**	Format
M1	MC	**M2**	CMC	**M3**	MC	**M4a**	CMC
M4b	OE	**M5**	MC	**M6**	OE	**M7**	MC
M8	OE	**M9**	OE	**M10**	MC	**M11**	CMC
M12	OE	**M13**	OE	**M14a**	OE	**M14b**	OE
M14c	OE	**M15**	Cloze	**M16a**	OE	**M16b**	OE
M16c	OE	**M17**	OE	**M18**	MC	**M19**	MC
M20a	OE	**M20b**	OE	**M21**	MC	**M22**	MC
M23	MC	**M24**	MC	**M25**	MC	**M26a**	OE
M26b	MC	**M27**	CMC	**M28**	MC	**M29a**	OE
M29b	OE	**M30**	CMC	**M31**	CMC	**M32**	OE
A1	MC	**A2**	Cloze	**A3**	CMC	**A4_1**	MC
A4_2	MC	**A4_3**	MC	**A4_4**	MC	**A4_5**	MC
A4_6	MC	**A5**	MC	**B1**	CMC	**B2**	OE
B3	OE	**B4**	MC	**B5**	OE	**B6**	MC
B7	MC	**B8**	OE	**B9**	MC	**B10**	MC
C1	MC	**C2**	MC	**C3**	MC	**C4**	MC
C5	MC	**C6**	MC	**C7**	MC	**C8**	MC
C9	OE	**C10**	MC	**D1**	OE	**D2**	MC
D3	OE	**D4**	OE	**D5**	OE	**D6**	MC
D7	MC	**D8**	MC	**D9**	MC	**E1**	OE
E2	MC	**E3**	CMC	**E4**	MC	**E5**	CMC
E6	MC	**E7**	CMC	**E8**	MC	**E9**	MC
E10	CMC						

MC=multiple–choice questions with four possible choices, CMC=complex multiple–choice and OE=open–ended.

Table E2: DIF results for maths and Italian language test: application 1.

Item	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2
M1	-0.449	4.228*	0.005	**M25**	0.044	0.043	–	**B10**	0.280	2.368	–
M2	-0.240	0.393	–	**M26a**	-0.106	0.131	–	**C1**	0.038	0.057	–
M3	0.253	0.393	–	**M26b**	-0.123	0.471	–	**C2**	-0.126	0.671	–
M4a	0.158	0.393	–	**M27**	0.032	0.009	–	**C3**	-0.233	0.907	–
M4b	-0.427	4.275*	0.007	**M28**	-0.449	5.123*	0.006	**C4**	0.306	3.037	–
M5	0.513	4.982*	0.010	**M29a**	-0.591	9.575**	0.019	**C5**	0.167	1.094	–
M6	0.728	4.865*	0.021	**M29b**	-0.568	5.745*	0.014	**C6**	0.097	0.114	–
M7	0.054	0.039	–	**M30**	1.288	31.476***	0.071	**C7**	0.357	3.021	–
M8	-0.592	5.733*	0.010	**M31**	0.921	9.156**	0.041	**C8**	-0.189	1.195	–
M9	0.303	1.205	–	**M32**	0.557	5.817*	0.014	**C9**	-0.052	0.106	–
M10	-0.330	2.548	–	–	–	–	–	**C10**	-0.225	1.862	–
M11	-0.243	1.149	–	**A1**	-0.021	0.017	–	**D1**	0.151	0.721	–
M12	0.689	7.419**	0.010	**A2**	-0.112	0.249	–	**D2**	-0.299	1.798	–
M13	0.518	6.049*	0.010	**A3**	0.187	0.538	–	**D3**	-0.803	8.015**	0.031
M14a	-0.761	8.788**	0.016	**A4_1**	0.230	0.232	–	**D4**	-0.264	0.443	–
M14b	-0.377	2.368	–	**A4_2**	0.338	0.429	–	**D5**	0.070	0.197	–
M14c	-1.181	20.021***	0.033	**A4_3**	0.917	12.046***	0.053	**D6**	-0.012	0.002	–
M15	0.389	2.378	–	**A4_4**	-0.007	0.002	–	**D7**	0.199	1.614	–
M16a	-0.127	0.079	–	**A4_5**	0.058	0.048	–	**D8**	-0.150	0.881	–
M16b	0.025	0.011	–	**A4_6**	0.033	0.042	–	**D9**	-0.194	0.819	–
M16c	-0.447	3.174	–	**A5**	-0.221	0.649	–	**E1**	-0.112	0.554	–
M17	-0.425	3.711	–	**B1**	-0.049	0.065	–	**E2**	-0.188	1.109	–
M18	0.114	0.284	–	**B2**	0.089	0.326	–	**E3**	0.139	0.721	–
M19	-0.079	0.101	–	**B3**	0.773	4.365*	0.031	**E4**	-0.232	0.709	–
M20a	-0.390	0.695	–	**B4**	0.117	0.589	–	**E5**	-0.164	0.656	–
M20b	-0.172	0.357	–	**B5**	0.149	0.277	–	**E6**	0.285	0.976	–
M21	0.093	0.238	–	**B6**	-0.273	2.547	–	**E7**	-0.009	0.003	–
M22	-0.218	1.080	–	**B7**	0.304	3.019	–	**E8**	0.313	3.213	–
M23	0.085	0.188	–	**B8**	0.121	0.521	–	**E9**	-0.086	0.019	–
M24	0.745	12.411***	0.021	**B9**	-0.482	8.241**	0.015	**E10**	-0.209	1.514	–

*<0.05, **<0.01, *<0.001. If χ^2 is not significant ΔR^2 is not present.

Table E3: DIF results for maths and Italian language test: application 2.

Item	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2
M1	-0.105	0.203	–	**M25**	-0.431	3.881	–	**B10**	0.098	0.247	–
M2	-0.404	1.247	–	**M26a**	0.145	0.283	–	**C1**	0.347	3.646	–
M3	0.215	0.764	–	**M26b**	-0.301	2.103	–	**C2**	-0.282	2.509	–
M4a	0.134	0.355	–	**M27**	-0.091	0.070	–	**C3**	-0.894	10.039**	0.036
M4b	-0.026	0.012	–	**M28**	-0.644	7.956**	0.010	**C4**	-0.087	0.197	–
M5	-0.120	0.232	–	**M29a**	-0.930	17.473***	0.033	**C5**	-0.254	2.242	–
M6	0.451	1.352	–	**M29b**	-0.922	12.826*	0.027	**C6**	0.607	6.068**	0.019
M7	0.321	1.369	–	**M30**	1.235	22.725***	0.042	**C7**	0.188	0.772	–
M8	-0.959	12.569***	0.024	**M31**	1.065	10.598**	0.028	**C8**	-0.507	7.121**	0.015
M9	-0.104	0.139	–	**M32**	0.364	2.151	–	**C9**	-0.062	0.124	–
M10	-0.156	0.464	–	–	–	–	–	**C10**	0.207	1.354	–
M11	-0.181	0.475	–	**A1**	-0.028	0.026	–	**D1**	-0.307	2.332	–
M12	0.644	6.459*	0.010	**A2**	-0.177	0.585	–	**D2**	0.063	0.081	–
M13	0.937	14.759***	0.023	**A3**	0.290	1.397	–	**D3**	-0.267	0.933	–
M14a	-1.110	11.140***	0.020	**A4_1**	-0.538	1.174	–	**D4**	-0.622	3.057	–
M14b	0.590	4.560*	0.006	**A4_2**	-0.355	0.322	–	**D5**	0.352	3.872	–
M14c	-0.812	8.153***	0.011	**A4_3**	-0.249	0.757	–	**D6**	-0.521	3.219	–
M15	1.070	16.820***	0.027	**A4_4**	0.213	1.321	–	**D7**	0.140	0.664	–
M16a	-1.078	4.804*	0.024	**A4_5**	-0.773	6.677**	0.025	**D8**	0.154	0.783	–
M16b	0.110	0.169	–	**A4_6**	-0.202	1.256	–	**D9**	-0.212	1.024	–
M16c	-0.858	9.916**	0.021	**A5**	-0.280	1.050	–	**E1**	-0.325	2.971	–
M17	-0.242	0.849	–	**B1**	0.243	1.584	–	**E2**	0.410	4.877*	0.007
M18	-0.085	0.116	–	**B2**	-0.392	4.968*	0.009	**E3**	0.401	4.565*	0.006
M19	-0.698	7.646*	0.012	**B3**	-0.217	0.222	–	**E4**	0.136	0.302	–
M20a	-0.441	1.133	–	**B4**	-0.259	2.315	–	**E5**	0.234	1.272	–
M20b	-0.841	7.363**	0.019	**B5**	0.480	3.114	–	**E6**	0.900	11.109***	0.034
M21	-0.020	0.008	–	**B6**	-0.353	3.782	–	**E7**	0.299	2.746	–
M22	-0.540	5.180*	0.009	**B7**	-0.217	1.339	–	**E8**	0.024	0.018	–
M23	-0.561	5.947*	0.010	**B8**	0.026	1.338	–	**E9**	0.420	0.578	–
M24	0.912	15.701***	0.026	**B9**	-0.501	8.122**	0.015	**E10**	0.848	23.723***	0.040

*<0.05, **<0.01, *<0.001. If χ^2 is not significant ΔR^2 is not present.

Table E4: DIF results for maths and Italian language test: application 3.

Item	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2	**Item**	Coef.	χ^2	ΔR^2
M1	-0.070	0.131	–	**M25**	-0.326	3.373	–	**B10**	-0.179	0.855	–
M2	-0.078	0.095	–	**M26a**	0.208	0.917	–	**C1**	0.175	0.942	–
M3	0.037	0.039	–	**M26b**	-0.029	0.028	–	**C2**	-0.001	0.001	–
M4a	0.109	0.393	–	**M27**	0.265	1.286	–	**C3**	-0.468	2.502	–
M4b	0.298	2.761	–	**M28**	-0.606	9.292**	0.024	**C4**	-0.518	6.878**	0.022
M5	-0.600	9.775**	0.024	**M29a**	-0.184	1.107	–	**C5**	-0.345	3.749	–
M6	-0.193	0.648	–	**M29b**	-0.135	0.467	–	**C6**	0.876	9.475**	0.022
M7	-0.049	0.052	–	**M30**	0.004	0.001	–	**C7**	-0.277	1.527	–
M8	-0.368	3.395	–	**M31**	0.302	2.079	–	**C8**	-0.171	0.796	–
M9	-0.202	0.823	–	**M32**	0.226	1.570	–	**C9**	0.184	0.931	–
M10	-0.220	1.5254	–	–	–	–	–	**C10**	0.193	1.156	–
M11	-0.316	1.871	–	**A1**	-0.110	0.341	–	**D1**	-0.135	0.507	–
M12	0.195	0.850	–	**A2**	-0.205	0.829	–	**D2**	0.115	0.269	–
M13	0.537	7.054**	0.015	**A3**	-0.013	0.003	–	**D3**	0.179	0.366	–
M14a	0.134	0.353	–	**A4_1**	-1.043	4.507*	0.059	**D4**	0.033	0.006	–
M14b	0.630	6.982**	0.016	**A4_2**	-1.343	6.422*	0.109	**D5**	0.194	1.086	–
M14c	-0.089	0.134	–	**A4_3**	-0.779	8.382***	0.052	**D6**	-0.146	0.201	–
M15	0.352	2.936	–	**A4_4**	0.074	0.154	–	**D7**	-0.210	1.389	–
M16a	0.614	3.831	–	**A4_5**	-0.518	2.986	–	**D8**	0.379	4.351*	0.013
M16b	-0.081	0.168	–	**A4_6**	-0.049	0.071	–	**D9**	0.046	0.055	–
M16c	0.168	0.796	–	**A5**	0.063	0.043	–	**E1**	-0.109	0.367	–
M17	0.228	0.698	–	**B1**	0.255	1.589	–	**E2**	0.584	8.765**	0.013
M18	-0.284	2.043	–	**B2**	0.534	9.659**	0.029	**E3**	0.692	12.603***	0.037
M19	-0.763	13.556***	0.032	**B3**	-0.927	6.527**	0.051	**E4**	0.384	1.857	–
M20a	0.332	1.066	–	**B4**	-0.243	2.153	–	**E5**	0.144	0.438	–
M20b	-0.244	1.013	–	**B5**	0.393	2.485	–	**E6**	0.617	6.036*	0.024
M21	0.274	2.202	–	**B6**	0.293	2.499	–	**E7**	0.231	1.501	–
M22	-0.099	0.275	–	**B7**	-0.272	2.293	–	**E8**	-0.043	0.054	–
M23	-0.268	2.177	–	**B8**	0.028	0.022	–	**E9**	-0.027	0.002	–
M24	0.234	1.497	–	**B9**	-0.149	0.673	–	**E10**	0.998	29.538***	0.080

*<0.05, **<0.01, *<0.001. If χ^2 is not significant ΔR^2 is not present.

References

Angrist, Joshua and Guido Imbens (1995). *Identification and estimation of local average treatment effects*.

Association, American Educational Research, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (US) (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Austin, Peter C (2009a). "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples". In: *Statistics in medicine* 28(25), pp. 3083–3107.

Austin, Peter C (2009b). "Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations". In: *Biometrical journal* 51(1), pp. 171–184.

Austin, Peter C (2011). "An introduction to propensity score methods for reducing the effects of confounding in observational studies". In: *Multivariate behavioral research* 46(3), pp. 399–424.

Azzolini, Davide, Philipp Schnell, and John RB Palmer (2012). "Educational achievement gaps between immigrant and native students in two new immigration countries: Italy and Spain in comparison". In: *The Annals of the American Academy of Political and Social Science* 643(1), pp. 46–77.

Azzolini, Davide and Loris Vergolini (2014). "Tracking, inequality and education policy. Looking for a recipe for the Italian case". In: *Scuola democratica* 2, pp. 0–0.

Barban, Nicola and Michael J White (2011). "Immigrants' children's transition to secondary school in Italy". In: *International Migration Review* 45(3), pp. 702–726.

Barone, Carlo (2006). "Cultural capital, ambition and the explanation of inequalities in learning outcomes: A comparative analysis". In: *Sociology* 40(6), pp. 1039–1058.

Becker, Michael, Oliver Lüdtke, Ulrich Trautwein, Olaf Köller, and Jürgen Baumert (2012). "The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter?" In: *Journal of educational psychology* 104(3), p. 682.

Becker, Rolf (2003). "Educational expansion and persistent inequalities of education: Utilizing subjective expected utility theory to explain increasing participation rates in upper secondary school in the Federal Republic of Germany". In: *European sociological review* 19(1), pp. 1–24.

Becker, Sascha O, Andrea Ichino, et al. (2002). "Estimation of average treatment effects based on propensity scores". In: *The stata journal* 2(4), pp. 358–377.

Beretvas, S Natasha, Stephanie W Cawthon, L Leland Lockhart, and Alyssa D Kaye (2012). "Assessing impact, DIF, and DFF in accommodated item scores: a comparison of multilevel measurement model parameterizations". In: *Educational and Psychological Measurement* 72(5), pp. 754–773.

Berger, Moritz and Gerhard Tutz (2016). "Detection of Uniform and Nonuniform Differential Item Functioning by Item-Focused Trees". In: *Journal of Educational and Behavioral Statistics* 41(6), pp. 559–592.

Berk, R A (1982). *Handbook of methods for detecting test bias*.

Berti, Fabio, Antonella Agostino, Achille Lemmi, and Laura Neri (2014). "Poverty and deprivation of immigrants vs. natives in Italy". In: *International Journal of Social Economics* 41(8), pp. 630–649.

Birnbaum, Allan (1958). "On the estimation of mental ability". In: *Series Rep* 15, pp. 7755–7723.

Birnbaum, Allan (1968). "Some latent trait models and their use in inferring an examinee's ability". In: *Statistical theories of mental test scores*.

Boeck, Paul De, Marjan Bakker, Robert Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx, and Ivailo Partchev (2011). "The estimation of item response models with the lmer function from the lme4 package in R". In: *Journal of Statistical Software* 39(12), pp. 1–28.

Boudon, Rymond (1974). *Education, Opportunity and Social Inequality*. Wiley: New York.

Breslow, Norman E, Nicholas E Day, and Walter Davis (1980). *Statistical methods in cancer research*. Vol. 1. International agency for research on cancer Lyon.

Caliendo, Marco and Sabine Kopeinig (2008). "Some practical guidance for the implementation of propensity score matching". In: *Journal of economic surveys* 22(1), pp. 31–72.

Camilli, Gregory (2006). "Test fairness". In: *Educational measurement* 4, pp. 221–256.

Carstensen, Bendix and Martyn Plummer (2011). "Using Lexis Objects for Multi-State Models in R". In: *Journal of Statistical Software* 38(6), pp. 1–18.

Checchi, Daniele and Luca Flabbi (2013). "Intergenerational mobility and schooling decisions in Germany and Italy: The impact of secondary school tracks". In: *Rivista di politica economica* 3(7-9), pp. 7–60.

Cho, Sun-Joo and Allan Cohen (2010). "A multilevel mixture IRT model with an application to DIF". In: *Journal of Educational and Behavioral Statistics* 35(3), pp. 336–370.

Clauser, Brian, Kathy Mazor, and Ronald K Hambleton (1993). "The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure". In: *Applied Measurement in Education* 6(4), pp. 269–279.

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences. 2nd.*

Cohen, Jacob (1994). "The earth is round (p¡. 05)." In: *American psychologist* 49(12), p. 997.

D'Alessio, Giovanni (2017). "Benessere, Contesto Socio-Economico E Differenze Di Prezzo: Il Divario Tra Nord e Sud (Well-Being, the Socio-Economic Context and Price Differences: The North-South Gap)". In:

Davier, Matthias von and Claus H Carstensen (2007). *Multivariate and mixture distribution Rasch models: Extensions and applications*. Springer Science & Business Media.

Dehejia, Rajeev H and Sadek Wahba (1999). "Causal effects in non-experimental studies: Reevaluating the evaluation of training programs". In: *Journal of the American statistical Association* 94(448), pp. 1053–1062.

Dehejia, Rajeev H and Sadek Wahba (2002). "Propensity score-matching methods for nonexperimental causal studies". In: *Review of Economics and statistics* 84(1), pp. 151–161.

Donoghue, John R, Paul W Holland, and Dorothy T Thayer (1993). "A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning". In: *Differential item functioning*, pp. 137–166.

Doran, Harold, Douglas Bates, Paul Bliese, and Maritza Dowling (2007). "Estimating the multilevel Rasch model: With the lme4 package". In: *Journal of Statistical Software* 20(2), pp. 1–18.

Dorans, Neil J and Paul W Holland (1992). "Dif detection and description: mantel-haenszel and standardization". In: *ETS Research Report Series* 1992(1).

Dustmann, Christian (2004). "Parental background, secondary school track choice, and wages". In: *Oxford Economic Papers* 56(2), pp. 209–230.

Finch, Holmes (2016). "Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods". In: *Applied Measurement in Education* 29(1), pp. 30–45.

Finch, Holmes and Brian French (2007). "Detection of crossing differential item functioning: A comparison of four methods". In: *Educational and psychological Measurement* 67(4), pp. 565–582.

French, Brian and Holmes Finch (2010). "Hierarchical logistic regression: Accounting for multilevel data in DIF detection". In: *Journal of Educational Measurement* 47(3), pp. 299–317.

French, Brian and Holmes Finch (2013). "Extensions of Mantel–Haenszel for multilevel DIF detection". In: *Educational and Psychological Measurement* 73(4), pp. 648–671.

French, Brian and Holmes Finch (2015). "Transforming SIBTEST to account for multilevel data structures". In: *Journal of Educational Measurement* 52(2), pp. 159–180.

Gambacorta, Romina (2017). "Immigration and Poverty: The Case of Italy". In: *Research on Economic Inequality: Poverty, Inequality and Welfare*. Emerald Publishing Limited, pp. 229–257.

Gamoran, Adam (1992). "The variable effects of high school tracking". In: *American Sociological Review*, pp. 812–828.

Glas, Cees AW and Juan Carlos Suárez Falcón (2003). "A comparison of item-fit statistics for the three-parameter logistic model". In: *Applied Psychological Measurement* 27(2), pp. 87–106.

Gómez-Benito, Juana, M Dolores Hidalgo, and José-Luis Padilla (2009). "Efficacy of effect size measures in logistic regression: An application for detecting DIF". In: *Methodology* 5(1), pp. 18–25.

Gómez-Benito, Juana and María José Navas-Ara (2000). "A Comparison of $\chi 2$, RFA and IRT Based Procedures in the Detection of DIF". In: *Quality and Quantity* 34(1), pp. 17–31.

Gu, Xing Sam and Paul R Rosenbaum (1993a). "Comparison of multivariate matching methods: Structures, distances, and algorithms". In: *Journal of Computational and Graphical Statistics* 2(4), pp. 405–420.

Gu, Xing Sam and Paul R Rosenbaum (1993b). "Comparison of multivariate matching methods: structures, distances, and algorithms". In: *Journal of Computational and Graphical Statistics* 2(4), pp. 405–420.

Guill, Karin, Oliver Lüdtke, and Olaf Köller (2017). "Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching study". In: *Learning and instruction* 47, pp. 43–52.

Hanushek, Eric A (2006). "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries". In: *The Economic Journal* 116(510).

Heckman, James J and Richard Robb (1986). "Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes". In: *Drawing inferences from self-selected samples*. Springer, pp. 63–107.

Hidalgo, M Dolores and José Antonio LóPez-Pina (2004). "Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures". In: *Educational and Psychological Measurement* 64(6), pp. 903–915.

Hlavac, Marek (2018). *Stargazer: Well-Formatted Regression and Summary Statistics Tables. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI)*. URL: `https://CRAN.R-project.org/package=stargazer`.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference". In: *Journal of Statistical Software* 42(8), pp. 1–28.

Holland, Paul W and Dorothy Thayer (1985). "An alternate definition of the ETS delta scale of item difficulty". In: *ETS Research Report Series* 1985(2), pp. 1–10.

Holland, Paul W, Dorothy T Thayer, H Wainer, and HI Braun (1988). "Differential item performance and the Mantel-Haenszel procedure". In: *Test validity*, pp. 129–145.

Hosmer, David W, Stanley Lemeshow, and Rodney Sturdivant (2013). *Applied logistic regression*. Vol. 398. John Wiley & Sons.

INVALSI (2012a). *Quadro di riferimento della prova di italiano.* available online. URL: http://www.invalsi.it/snv2012/documenti/QDR/QdR_Italiano.pdf.

INVALSI (2012b). *Quadro di riferimento secondo ciclo di istruzione prova di matematica.* available online. URL: http://www.invalsi.it/snv2012/documenti/QDR/QdR_Mat_II_ciclo.pdf.

INVALSI (2016). *Rilevazioni nazionali degli apprendimenti 2015/2016.* Techical report. INVALSI.

INVALSI (2017a). *PROVA DI ITALIANO - Scuola Secondaria di II grado - Classe Seconda - Fascicolo 1.* available online. URL: https://www.engheben.it/prof/materiali/invalsi/invalsi_seconda_superiore/2016-2017/invalsi_italiano_2016-2017_secondaria_seconda.pdf.

INVALSI (2017b). *PROVA DI MATEMATICA - Scuola Secondaria di II grado - Classe Seconda - Fascicolo 1.* available online. URL: https://www.engheben.it/prof/materiali/invalsi/invalsi_seconda_superiore/2016-2017/invalsi_matematica_2016-2017_secondaria_seconda.pdf.

INVALSI (2017c). *Rilevazioni nazionali degli apprendimenti 2016/2017.* Techical report. INVALSI.

Jodoin, Michael G and Mark J Gierl (2001). "Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection". In: *Applied Measurement in Education* 14(4), pp. 329–349.

Kamata, Akihito (2001). "Item analysis by the hierarchical generalized linear model". In: *Journal of Educational Measurement* 38(1), pp. 79–93.

Khalid, Muhammad N (2011). "A comparison of top-down and bottom-up approaches in the identification of differential item functioning using confirmatory factor analysis". In: *The International Journal of Educational and Psychological Assessment* 7(2), pp. 1–18.

Khalid, Muhammad N and Cees A W Glass (2013). "A Step-wise Method for Evaluation of Differential Item Functioning." In: *Journal of Applied Quantitative Methods* 8(2), pp. 25–47.

Khattab, Nabil (2015). "Students' aspirations, expectations and school achievement: What really matters?" In: *British Educational Research Journal* 41(5), pp. 731–748.

King, Gary and Richard Nielsen (2018). "Why Propensity Scores Should Not Be Used for Matching". In: *Political Analysis*.

Kirk, Roger (1996). "Practical significance: A concept whose time has come". In: *Educational and psychological measurement* 56(5), pp. 746–759.

Lamprianou, Iasonas (2013). "Application of single-level and multi-level Rasch models using the lme4 package". In: *J Appl Meas* 14, pp. 79–90.

Lavrijsen, Jeroen and Ides Nicaise (2016). "Educational tracking, inequality and performance: New evidence from a differences-in-differences technique". In: *Research in Comparative and International Education* 11(3), pp. 334–349.

Lechner, Michael (2001). "A note on the common support problem in applied evaluation studies". In: *Univ. of St. Gallen Economics, Disc. Paper* 1.

Lee, HyeSun and Kurt F Geisinger (2014). "The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF". In: *International Journal of Testing* 14(4), pp. 313–338.

Liu, Yan, Bruno Zumbo, Paul Gustafson, Yi Huang, Edward Kroc, and Amery D Wu (2016). "Investigating Causal DIF via Propensity Score Methods". In: *Practical Assessment, Research & Evaluation* 21(13), p. 2.

Lord, Frederic M (1980). *Applications of item response theory to practical testing problems*. Routledge.

Luc, Le T (2009). "Investigating gender differential item functioning across countries and test languages for PISA science items". In: *International Journal of Testing* 9(2), pp. 122–133.

Magis, David, Sebastien Beland, Francis Tuerlinckx, and Paul De Boeck (2010). "A general framework and an R package for the detection of dichotomous differential item functioning". In: *Behavior Research Methods* 42, pp. 847–862.

Magis, David and Bruno Facon (2012). "Angoff's delta method revisited: Improving DIF detection under small samples". In: *British Journal of Mathematical and Statistical Psychology* 65(2), pp. 302–321.

Magis, David, Gilles Raîche, Sébastien Béland, and Paul Gérard (2011). "A generalized logistic regression procedure to detect differential item functioning among multiple groups". In: *International Journal of Testing* 11(4), pp. 365–386.

Magis, David, Francis Tuerlinckx, and Paul De Boeck (2015). "Detection of differential item functioning using the lasso approach". In: *Journal of Educational and Behavioral Statistics* 40(2), pp. 111–135.

Martinková, Patrícia, Adéla Drabinová, Yuan-Ling Liaw, Elizabeth A Sanders, Jenny L McFarland, and Rebecca M Price (2017). "Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments". In: *CBE–Life Sciences Education* 16(2), rm2.

Millsap, Roger E (2007). "Invariance in measurement and prediction revisited". In: *Psychometrika* 72(4), pp. 461–473.

Millsap, Roger E and Howard T Everson (1993). "Methodology review: Statistical approaches for assessing measurement bias". In: *Applied psychological measurement* 17(4), pp. 297–334.

Mocetti, Sauro (2012). "Educational choices and the selection process: before and after compulsory schooling". In: *Education Economics* 20(2), pp. 189–209.

Morgan, Stephen L and Jennifer J Todd (2008). "A diagnostic routine for the detection of consequential heterogeneity of causal effects". In: *Sociological Methodology* 38(1), pp. 231–281.

Nagelkerke, Nico JD (1991). "A note on a general definition of the coefficient of determination". In: *Biometrika* 78(3), pp. 691–692.

Narayanon, Pankaja and Hariharan Swaminathan (1996). "Identification of items that show nonuniform DIF". In: *Applied Psychological Measurement* 20(3), pp. 257–274.

OECD (2010). *School Factors Related to Quality and Equity. Results from PISA 2000.*

OECD (2015). "PISA 2015 Results (Volume I)". In: DOI: `http://dx.doi.org/10.1787/9789264266490-en`. URL: `/content/book/9789264266490-en`.

Oliveri, María Elena, Kadriye Ercikan, and Bruno Zumbo (2013). "Analysis of sources of latent class differential item functioning in international assessments". In: *International Journal of Testing* 13(3), pp. 272–293.

Oliveri, María Elena, Kadriye Ercikan, and Bruno Zumbo (2014). "Effects of population heterogeneity on accuracy of DIF detection". In: *Applied Measurement in Education* 27(4), pp. 286–300.

Olmos, Antonio and Priyalatha Govindasamy (2015). "Propensity scores: a practical introduction using R". In: *Journal of MultiDisciplinary Evaluation* 11(25), pp. 68–88.

Opdenakker, Marie-Christine and Jan Van Damme (2006). "Differences between secondary schools: A study about school context, group composition, school practice, and school effects with special attention to public and Catholic schools and types of schools". In: *School effectiveness and School improvement* 17(1), pp. 87–117.

Osterlind, Steven J and Howard T Everson (2009). *Differential item functioning*. Vol. 161. Sage Publications.

Özdemir, Burhanettin (2015). "A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest". In: *Procedia-Social and Behavioral Sciences* 174, pp. 2075–2083.

Panichella, Nazareno and Moris Triventi (2014). "Social inequalities in the choice of secondary school: Long-term trends during educational expansion and reforms in Italy". In: *European Societies* 16(5), pp. 666–693.

Pellizzari, Michele (2018). *Propensity Score: Proofs of the Balancing Property and of Unconfoundedness*. URL: `file:///C:/Users/bazoli/Downloads/PROPENSITYSCORE20101004152830%5C%20(1).PDF` (visited on 12/21/2018).

Penfield, Randall D and Gregory Camilli (2006). "5 Differential Item Functioning and Item Bias". In: *Handbook of statistics* 26, pp. 125–167.

Raju, Nambury S (1988). "The area between two item characteristic curves". In: *Psychometrika* 53(4), pp. 495–502.

Raju, Nambury S, Fritz Drasgow, and Jeffrey Slinde (1993). "An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning". In: *Educational and psychological measurement* 53(2), pp. 301–314.

Ramsey, Paul A (1993). "Sensitivity review: The ETS experience as a case study". In: *Differential item functioning*, pp. 367–388.

Rasch, Georg (1960). "Probabilistic models for some intelligence and achievement tests". In: *Copenhagen: Danish Institute for Educational Research*.

Raudenbush, Stephen W and Anthony S Bryk (2002). *Hierarchical linear models: Applications and data analysis methods*. Vol. 1. Sage.

Ricci, Roberto (2010). "The Economic, Social, and Cultural Background: a continuous index for the Italian Students of the fifth grade". In: *Atti Convegno SIS-Società Italiana di Statistica Padova-2010, Sessioni specializzate*.

Richard, P Phelps (2008). *The Role and Importance of Standardized Testing in the World of Teaching and Training*.

Rosenbaum, Paul R (1987). "Model-based direct adjustment". In: *Journal of the American Statistical Association* 82(398), pp. 387–394.

Rosenbaum, Paul R (1991). "A characterization of optimal designs for observational studies". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 597–610.

Rosenbaum, Paul R (2002). "Observational studies". In: *Observational studies*. Springer, pp. 1–17.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70(1), pp. 41–55.

Rosenbaum, Paul R and Donald B Rubin (1984). "Reducing bias in observational studies using subclassification on the propensity score". In: *Journal of the American statistical Association* 79(387), pp. 516–524.

Rosenbaum, Paul R and Donald B Rubin (1985). "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score". In: *The American Statistician* 39(1), pp. 33–38.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and non-randomized studies." In: *Journal of educational Psychology* 66(5), p. 688.

Salgotra, Ajay K and Kumari Roma (2018). "Educational Aspiration and Socio-Economic Status among Secondary School Students". In: *Journal Of Humanities And Social Science* 23(10), pp. 25–29.

Scheuneman, Janice D (1981). "A new look at bias in aptitude tests". In: *New Directions for Testing & Measurement*.

Sewell, William H and Vimal P Shah (1968). "Parents' education and children's educational aspirations and achievements". In: *American sociological review*, pp. 191–209.

Shahidul, SM, AHM Zehadul Karim, and S Mustari (2015). "Social Capital and Educational Aspiration of Students: Does Family Social Capital Affect More Compared to School Social Capital?" In: *International Education Studies* 8(12), p. 255.

Shealy, Robin and William Stout (1993). “A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF”. In: *Psychometrika* 58(2), pp. 159–194.

Simpson, Edward H (1951). “The interpretation of interaction in contingency tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 238–241.

Strobl, Carolin, Julia Kopf, and Achim Zeileis (2015). “Rasch trees: A new method for detecting differential item functioning in the Rasch model”. In: *Psychometrika* 80(2), pp. 289–316.

Svetina, Dubravka and Leslie Rutkowski (2014). “Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments”. In: *Large-scale Assessments in Education* 2(1), p. 4.

Swaminathan, Hariharan and H Jane Rogers (1990). “Detecting differential item functioning using logistic regression procedures”. In: *Journal of Educational measurement* 27(4), pp. 361–370.

Tay, Louis, Qiming Huang, and Jeroen K Vermunt (2016). “Item response theory with covariates (IRT-C): assessing item recovery and differ- ential item functioning for the three-parameter logistic model”. In: *Educational and Psychological Measurement* 76(1).

Tay, Louis, Adam W Meade, and Mengyang Cao (2015). “An overview and practical guide to IRT measurement equivalence analysis”. In: *Organizational Research Methods* 18(1), pp. 3–46.

Tay, Louis, Daniel A Newman, and Jeroen K Vermunt (2011). “Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence”. In: *Organizational Research Methods* 14(1), pp. 147–176.

Thissen, David, Lynne Steinberg, and Howard Wainer (1988). “Use of item response theory in the study of group differences in trace lines.” In: *Test validity*, pp. 147–172.

Tutz, Gerhard and Moritz Berger (2016). “Item-focussed Trees for the Identification of Items in Differential Item Functioning”. In: *Psychometrika* 81(3), pp. 727–750.

Uttaro, Thomas and Roger E Millsap (1994). “Factors influencing the Mantel-Haenszel procedure the detection of differential item functioning”. In: *Applied Psychological Measurement* 18(1), pp. 15–25.

Walker, Cindy M and Sakine G Sahin (2016). "Using a Multidimensional IRT Framework to Better Understand Differential Item Functioning (DIF) A Tale of Three DIF Detection Procedures". In: *Educational and Psychological Measurement*, p. 0013164416657137.

Weiss, David J (2014). *New Horizon Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. Elsevier.

Wiberg, Marie (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test*.

Woods, Carol M, Li Cai, and Mian Wang (2013). "The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT". In: *Educational and Psychological Measurement* 73(3), pp. 532–547.

Wu, Amery D, Yan Liu, Jake Stone, Danjie Zou, and Bruno Zumbo (2017). "Is Difference in Measurement Outcome between Groups Differential Responding, Bias or Disparity? A Methodology for Detecting Bias and Impact from an Attributional Stance". In: *Frontiers in Education*. Vol. 2. Frontiers, p. 39.

Zumbo, Bruno (2007). "Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going". In: *Language assessment quarterly* 4(2), pp. 223–233.

Zumbo, Bruno and Michaela N Gelin (2005). "A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological/Community Moderated (or Mediated) Test and Item Bias." In: *Journal of Educational Research & Policy Studies* 5(1), pp. 1–23.