Systematic Review

Analyzing Measurement Invariance for Studying the Gender Gap in Educational Testing: A Mixed Studies Systematic Review.

M. Carmen Navarro-González[1,2]

José-Luis Padilla[1,2]

Isabel Benítez[1,2]

[1]Affiliation: Department of Methodology for Behavioral Sciences, University of Granada, Granada, Spain

[2]Affiliation: Mind, Brain and Behavior Research Center [CIMCYC], Granada, Spain

Abstract:

Gender DIF is essential to address if fair and valid test scores interpretations are desired. Despite inequivalence having been addressed previously, there are not solid conclusions about the causes provoking this bias nor indications to reduce its presence. The objective of this mixed studies systematic review is to describe how measurement invariance has been addressed when studying the gender gap in educational assessments. We searched for quantitative, qualitative or mixed-methods studies that tested measurement invariance/DIF and/or applied qualitative methods to explore causes of the gender gap in educational assessments with adolescents. We used the QATSDD (Sirriyeh et al., 2012) to assess the risk of bias, and proposed a results-based convergent synthesis design. We included 87 studies, with 3,458,853 adolescent participants. Multigroup CFA and Mantel-Haenszel were the most used strategies to test measurement invariance/detect DIF. Certain methods, such as LCA, MMixIRTM, or SIBTEST were most used by studies that examined sources of DIF. The most used qualitative strategy to examine sources of DIF was content analysis. Limitations due to methodological concerns and missing data are discussed. We provide an important description of invariance testing/DIF detection methods that can serve as a guide to future researchers interested in sources of gender DIF.

Keywords: measurement invariance, differential item functioning, gender gap, systematic review, mixed methods.

José-Luis Padilla
Department of Methodology for Behavioral Sciences
University of Granada
Campus de Cartuja s/n
18071, Granada
Spain

jpadilla@ugr.es

Authorship

Following the CRediT roles taxonomy (available in https://credit.niso.org/), contributions are: M. Carmen Navarro-González: conceptualization, formal analysis, investigation, visualization, writing – original draft; José-Luis Padilla: conceptualization, investigation, supervision, writing – review & editing; Isabel Benítez: supervision, writing – review & editing. M. Carmen Navarro-González is the guarantor of the review.

Open Data

Our systematic review protocol was registered with Open Science Framework (OSF) on April 9th, 2022 (Registration DOI: https://doi.org/10.17605/OSF.IO/XZGDQ). Template data collection forms and Supplementary Material are available here: https://osf.io/rvqk7/?view_only=794676855e194430bc9636d51fa81124.

Funding

This research was funded by the FPU Program of the Spanish Ministry of Education, Culture and Sport under Grant [FPU19/02908], and by FEDER/Junta de Andalucía-Consejería de Economía y Conocimiento / Proyecto P20_00898 (Andalusian Government).

Introduction

Rationale

Testing for measurement equivalence is an agreed-upon requirement in psychometrics in order to make valid comparisons in cross-cultural research (i.e., comparative studies with participants from different sociodemographic, linguistic, and/or cultural groups) (van de Vijver & Leung, 2021). Current validity theory and psychometric validation methods can be useful for obtaining validity evidence of the comparative interpretations derived from survey statistics and test scores in cross-cultural research (e.g., Zumbo & Padilla, 2019).

Gender differences have been found across diverse cross-cultural studies, such as international educational testing projects with middle and high school students, in both cognitive (e.g., mathematics achievement, reading literacy...), and non-cognitive (e.g., school climate, self-efficacy, motivation...) domains. For example, results from the 2018 Programme for International Student Assessment (PISA) show that, despite having a level of academic achievement similar to boys', differences appear between both groups in various variables. For instance, girls have greater fear of failure and less self-efficacy than boys, whereas boys admit to having suffered more bullying episodes than girls, in most countries (OECD, 2019). Furthermore, Ayuso et al. (2020) stated that only 54.9% of girls considered themselves good at mathematics —as opposed to 71.5% of boys— even though more than 50% of teachers thought that boys should "never" consider themselves good at mathematics more frequently than girls. Moreover, Sakellariou and Fang (2021) found that having strong self-efficacy in mathematics skills predicts females' enrollment in higher education STEM (science, technology, engineering and mathematics) programs, but not males'. Such results could have important implications for girls' future career decisions, such as perpetuating the gender gap found in STEM careers, and/or undermining girls' self-efficacy and well-being. Therefore, distinguishing biased differences from true differences becomes vital in order to eventually focus educational policies on guaranteeing equal access to certain educational contexts. Providing validity evidence for comparative inference not only helps discard measurement artifacts (e.g., biases) as explanations of the gender gap, but also uncovers factors responsible for such biases.

Differential item functioning (DIF) analysis, as one of the most used statistical approaches to analyze the lack of measurement invariance at the item level and identify the presence of bias, can provide validity evidence to support comparative interpretations in cross-cultural survey research. Through DIF analysis, psychometricians explore if respondents from different sociodemographic, linguistic and/or cultural groups that are matched on the target construct, show different probabilities of item endorsement (e.g., Chen & Zumbo, 2017). Therefore, DIF analyses could provide information to determine if group differences could stem from some characteristic of the items and/or the testing situation that is irrelevant to the intended construct (Zumbo, 2007).

Over the last years, in addition to the traditional techniques of what Zumbo (2007) has called the first and second generations of DIF analysis, the search for factors responsible for DIF —together with the topics of multidimensionality, fairness and equity in testing, and understanding DIF as a validity issue— have become salient. Zumbo (2007) proposed that the emergence of those approaches is a sign of the third generation of DIF analysis. However, Li et al. (2021) found that, while the analysis of DIF in language research has expanded over time and with more sophisticated procedures, researchers do not address sources of DIF more frequently now than in earlier years. On the other hand, the necessity of searching for possible contextual explanations of DIF has led to the development of more comprehensive conceptual framework to guide DIF analysis, such as the Ecological Model of Item Response

(Zumbo et al., 2015), that allow researchers to investigate DIF causes by broadening their search to look at characteristics of the test items and/or testing situations from a multilevel perspective. Given that obtaining information about DIF causes could imply a contextual analysis, mixed-methods research could be a promising alternative that allows the results of quantitative DIF analysis techniques to be integrated with qualitative data from the ecology of the testing situation (Padilla et al., 2018), overcoming the limitations of the techniques of the first and second generations of DIF analysis. The integration of quantitative and qualitative methods could provide a more comprehensive understanding of DIF sources, measurement equivalence and explanations of the gender gap in cross-cultural research.

In order to establish a starting point from which to begin developing a methodological mixed-methods strategy to provide validity evidence for comparative groups inferences in cross-cultural studies, it could be useful to review available scientific literature about gender DIF/measurement invariance in educational testing projects. The general aim of the review is to have an overview of how measurement invariance and/or DIF has been addressed, quantitatively and qualitatively (qualitative methods for obtaining possible explanations of DIF or the lack of measurement invariance), in our setting of interest: gender DIF/measurement invariance in both cognitive and non-cognitive domains of educational testing projects involving middle and high school students.

Objectives

The main objective of this mixed studies systematic review is to describe how measurement invariance/DIF detection has been addressed, quantitatively and qualitatively, when studying the gender gap in both cognitive and non-cognitive domains of national and international educational testing projects involving middle and high school students. To this end, this systematic review answers the following questions:

1. Which quantitative techniques (psychometric analyses) are used for measurement invariance/DIF detection analysis regarding the gender gap in educational testing studies of middle and high school students?
2. Which qualitative methods are used to obtain insight into the causes of the gender gap revealed in educational testing studies of middle and high school students?
3. How have these quantitative techniques and qualitative methods been applied (e.g., research designs)?
4. What are the pros and cons of these quantitative techniques and qualitative methods for studying measurement invariance/DIF detection in the gender gap?

Methods

This paper was written by following the recommendations from the Preferred Reporting Items for Systematic Review and Meta-analysis [PRISMA] 2020 statement (Page et al., 2021). We report how we determined the number of included studies, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data extraction and synthesis of results, all procedures in the study, and all synthesis strategies.

Eligibility Criteria

In order to respond to the previous research questions, our search strategy and the eligibility criteria were based on the SPIDER (sample, phenomenon of interest, design, evaluation and research type) tool developed by Cooke et al. (2012) to provide a framework for qualitative/mixed studies reviews. Our eligibility criteria are as follows:

Sample (S)

We included educational testing studies in which the participants were female and male adolescents (12-18 years old) in middle or high school (or the international equivalents).

Studies that also included their teachers, and/or their parents were eligible too, in order to include as many perspectives as possible of the students' academic performance and related factors.

Phenomenon of Interest (PI)

We included educational testing studies that tried to analyze, detect and/or explain the presence of a gender gap between female and male students in their cognitive and/or non-cognitive evaluations.

Design (D)

We included studies that tried to study measurement invariance/DIF detection and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Evaluation (E)

Any educational testing study that evaluated students' academic achievement and/or non-cognitive factors that influence such academic achievement was included.

Research Type (R)

Both quantitative and qualitative studies, as well as mixed-methods studies, were included. Therefore, terms related to the present section ("R" terms) were not included in the search strategy. In addition, articles had to be published in peer-reviewed scientific journals, in English or Spanish, and in the fields of Psychology, Education and Social Sciences.

Information Sources

The search was executed in the following databases: Web of Science, Scopus, PsycArticles, PsycExtra, PsycInfo, ProQuest Psychology and Social Science Databases, and ProQuest Education Collection. The final and most recent search was executed on April 11th 2022.

Search Strategy

The search keywords were selected according to the American Psychological Association (APA) Thesaurus of Psychological Index Terms (APA, 2022) and the UNESCO Thesaurus (UNESCO, 2022). The search strategy was developed based on the SPIDER framework (Cooke et al, 2012), with the structure [S AND PI AND D AND E]. The search strategy was adapted to the exigencies and particularities of each database. The final search strategy with English and Spanish terms adapted for Web of Science is presented in Supplementary Material I (Navarro-González et al., 2023a). Here, we introduce the general search strategy with English terms:

1. (student* OR adolescent* OR teenager* OR young*)
2. AND ("gender gap" OR "gender inequalities" OR “gender DIF”)
3. AND (("measurement invariance" OR "measurement equivalence" OR "differential item functioning" OR DIF) OR (("measurement invariance" OR "measurement equivalence" OR "differential item functioning" OR DIF) AND (qualitative OR "qualitative method*")))
4. AND ((educational AND evaluation*) OR (education* AND assessment*) OR "educational testing" OR "testing project*" OR (education* AND measurement*) OR "educational program*" OR "students' evaluation*" OR (student* AND assessment) OR (student* AND testing) OR (student* AND program*) OR (student* AND project*) OR "academic performance" OR "academic achievement" OR "educational achievement" OR (student* AND performance) OR (student* AND achievement))

Selection Process

After the search strategy was implemented, the citations were imported to free online software to conduct the selection process: Rayyan QCRI (Ouzzani et al., 2016). A removal of duplicates was conducted.

To assess for eligibility, two reviewers independently screened the title and abstract of the retrieved studies. Inter-rater agreement was κ = .55, indicating moderate concordance (Landis & Koch, 1977). Then, the full text of the articles included in the first phase was examined by the two reviewers independently to decide conclusively which articles should have been included in the review. Inter-rater agreement in this stage was κ = .66, indicating substantial concordance (Landis & Koch, 1977). Any conflicts were solved by a third reviewer.

Data Collection Process

The data charting process was carried out using the software NVivo (QRS International, 2022). A predefined list of data items or categories for extraction was developed by the research team and implemented in the software. This can be found in Supplementary Material II (Navarro-González et al., 2023b). Emergent categories were added when necessary, allowing the process to be flexible and dynamic. Some categories were removed or slightly modified due to non-homogeneity of studies and/or irrelevance to our research questions. NVivo (QRS International, 2022) allowed a qualitative data extraction in which the retrieved information of each study was used to categorize the study in accordance with the predefined list of categories. This process was complemented with a quantitative data extraction conducted in SPSS Statistics 22 (IBM Corp., 2013) and jamovi 2.3.26 (The jamovi project, 2022). Templates for both data extractions and extracted data are available in Supplementary Material IV, V, VI and VII (Navarro-González et al., 2023d-g). A first reviewer extracted data from all included studies, whereas an independent reviewer performed a second data extraction for 10% of included studies. Inter-rater agreement was κ = .81 indicating an almost perfect concordance (Landis & Koch, 1977).

Data Items

We extracted data on bibliometric indicators (e.g., the first author's name and country, the year of publication, the journal), methodological aspects of the studies (e.g., the type of study, the characteristics of the participants), quantitative DIF detection/invariance testing analyses (e.g., the techniques used, the pros and cons of these techniques reported by the authors), and qualitative techniques for providing explanations of gender DIF/lack of measurement invariance (e.g., the qualitative methods used, the pros and cons of these methods reported by the authors). All data items and outcomes are fully detailed in the list presented in Supplementary Material II and III (Navarro-González et al., 2023b,c).

Study Risk of Bias Assessment

In order to assess the risk of bias for each study, we used the Quality Assessment Tool (QATSDD) developed by Sirriyeh et al. (2012). By using this tool, we assessed two elements: a) reporting quality, that is the extent to which an article provides detailed information about all research stages; this dimension of quality is related to transparency, accuracy, and completeness; and b) methodological quality indicating how well the study was conducted; this dimension of quality is related to trustworthiness (Hong & Pluye, 2019). The QATSDD allows the assessment of quantitative, qualitative, and mixed-methods studies. It contains 16 criteria scored on a scale from 0 to 3 and showed good (κ = 71.5%) inter-rater reliability (Sirriyeh et al., 2012). An additional item was included to evaluate the level of integration in mixed-methods studies (0 = no integration; 1 = integration on only one level; 2 = integration on two levels; 3 = integration on all three levels), in terms of the levels of integration presented by Fetters et al. (2013). Depending on the scores that studies got on the tool, each of them was classified into three categories of quality: low risk of bias (score ≥ 71%), medium risk of bias (56% ≤ score ≤ 70%) and high risk of bias (score ≤ 55%). A first reviewer applied this tool to all included studies, and a second reviewer applied this tool to 10% of studies. Inter-rater agreement was κ = .61, indicating substantial concordance (Landis & Koch, 1977).

Synthesis Methods

We followed a results-based convergent synthesis design (Hong et al., 2017), based on a data synthesis design flexible and adaptive to the nature of data found in the studies. Qualitative data was qualitatively synthesized by grouping and clustering the included studies into different categories, groups, or themes (Popay et al., 2006), and quantitatively synthesized by using a qualitative meta-summary (Sandelowski et al., 2007). Quantitative data was synthesized by obtaining overall descriptive statistics. Those preliminary syntheses served as a foundation for the final integrative narrative synthesis (Popay et al., 2006), which combines all previous information by exploring relationships within and between studies to develop a theory that provides answers to our research questions. Results from preliminary syntheses were displayed in frequency and cross tables. These representations of results allowed us to explore relationships in the data for our integrative narrative synthesis. Robustness of the final synthesis was assessed by reflecting critically on the synthesis process, taking into account results from the risk of bias and the reporting bias assessments.

Although all studies were eligible for this synthesis design, we categorized studies in three main groups according to DIF detection/invariance testing analysis. Our categorization for the approach to the analysis was based on Li et al.'s (2021), which in turn, is based on Zumbo's (2007) three generations of DIF: Category A includes studies that have not examined sources of DIF; Category B formed by studies that mentioned sources of DIF, but have not examined them; and Category C considers all the studies that have examined sources of DIF. For studies in Category C, further synthesis results are reported, regarding methods used for exploring sources of DIF.

Reporting Bias Assessment

Dissemination/publication bias and outcome reporting bias are known as important biases that can threaten the robustness and confidence of a systematic review (Shamseer et al., 2014). We tried to detect outcome reporting bias in the included studies by comparing the "methods" and "results" sections of the included articles to see if authors reported all outcomes they measured. Studies at risk of showing outcome reporting bias were flagged. Given the scope of this review is focused on the methodological aspects of the studies and the diversity of equivalence/DIF techniques reported, we did not assess dissemination/publication bias.

Certainty Assessment

In order to assess the confidence in cumulative evidence, we used the GRADE-CERQual approach (Lewin et al., 2015) to examine the robustness of our narrative synthesis. We examined the methodological limitations of each study supporting review findings, the relevance of each study, and the coherence and adequacy of the data supporting each finding. As recommended by Lewin et al. (2015), each finding was judged to have very low, low, moderate, or high confidence.

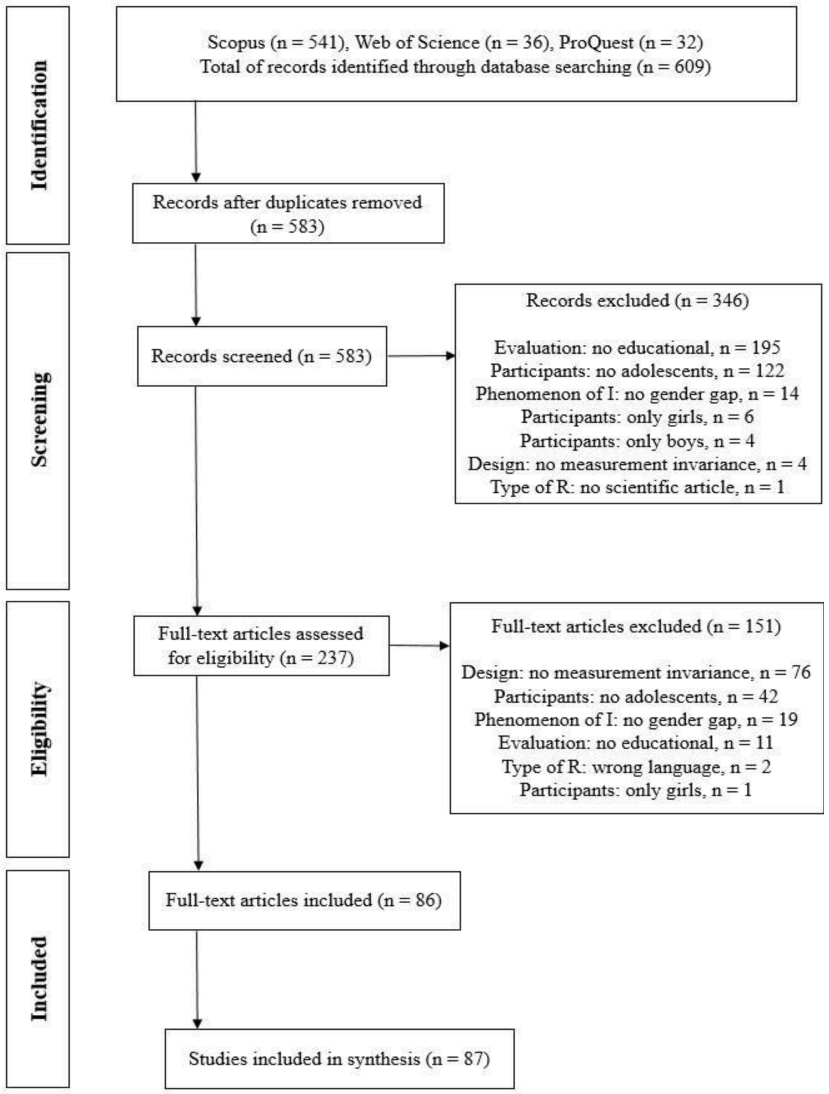
Results

Study Selection

We identified 609 records, of which 26 were duplicates. Therefore, after removing the duplicates, we screened 583 records. Then, we excluded 346 records in the title and abstract screening (reasons for exclusion are listed in Figure 1); 237 full-text articles remained for the assessment of eligibility. Of those, we excluded 151 (reasons for exclusion are listed in Figure 1), and finally included a total of 86 articles in the review. As one of them had two studies, we finally included a total of 87 studies. The flow of the selection process is depicted in Figure 1. A list of references of the included articles is provided in Supplementary Material VIII (Navarro-González et al., 2023h).

Figure 1

Selection of Sources of Evidence: Flow Diagram.


Identification
Scopus (n = 541), Web of Science (n = 36), ProQuest (n = 32)
Total of records identified through database searching (n = 609)
Records after duplicates removed (n = 583)
Screening
Records screened (n = 583)
Records excluded (n = 346)
Evaluation: no educational, n = 195
Participants: no adolescents, n = 122
Phenomenon of I: no gender gap, n = 14
Participants: only girls, n = 6
Participants: only boys, n = 4
Design: no measurement invariance, n = 4
Type of R: no scientific article, n = 1
Eligibility
Full-text articles assessed for eligibility (n = 237)
Full-text articles excluded (n = 151)
Design: no measurement invariance, n = 76
Participants: no adolescents, n = 42
Phenomenon of I: no gender gap, n = 19
Evaluation: no educational, n = 11
Type of R: wrong language, n = 2
Participants: only girls, n = 1
Included
Full-text articles included (n = 86)
Studies included in synthesis (n = 87)

Study Characteristics

Here, we will discuss the bibliometric indicators and methodological aspects of the studies. In Supplementary Material IX (Navarro-González et al., 2023i), two summarizing tables (Tables IX.1 and IX.2) displaying the main characteristics of each study are presented.

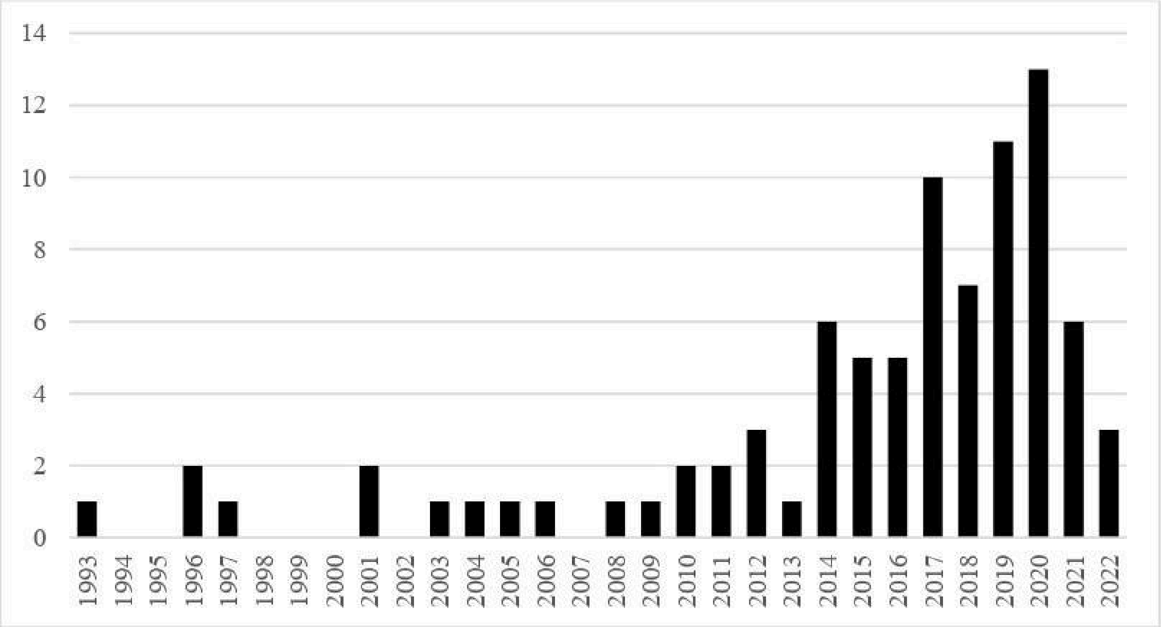
Bibliometric Indicators

Most articles have been published in JCR indexed journals (*n* = 79; 91.86%), with the most frequent journals being *Learning and Individual Differences* (*n* = 5), *Frontiers in Psychology* (*n* = 5), and *Applied Measurement in Education* (*n* = 5). Most articles were published in a Q1 journal (*n* = 30), followed by Q2 (*n* = 22) and Q3 (*n* = 20) ones. Only seven studies were published in a non-indexed journal.

The United States of America (USA) is the most frequent first author's country (*n* = 22) followed by Germany (*n* = 13), having a spectrum of 23 different countries. Figure 2 shows the publication trend from 1993 to 2022, with 2020 being the most fructiferous year.

Figure 2.

Publication Trend from 1993 to 2022.



Methodological Aspects

The vast majority of studies were quantitative (*n* = 83; 95.40%), and only four studies integrated qualitative techniques in a mixed-methods study (Cascella et al., 2020; Ferretti & Giberti, 2020; Mahmud & Nur, 2018; Yildirim & Büyüköztürk, 2018). We did not find any study with only a qualitative approach.

Approximately half of studies used available databases from large-scale studies, whereas the other half collected their own data. One study did both (Ferretti & Giberti, 2020); that is, the authors used available databases for the quantitative first stage and then collected their own data for their qualitative second stage. Among studies that collected their own data, almost half (*n* = 21; 46.67%) used a random recruitment strategy, nine (20.00%) used an incidental strategy, and 15 (33.33%) provided no information about participants' recruitment. Almost all studies (*n* = 82; 94.25%) only used students as participants. The modes of mean ages of students across studies were 16.50 and 16.79 years old (*Mdn* = 15.53), ranging from 12.70 to 17.10 years old. The total sample size is 3,458,853 participants (adolescents), ranging from 62 to 1,063,570. Most studies (73.30%) had more than 1000 participants. When studies had also disaggregated data from participants that did not meet our age requirements (e.g., Cascella et al., 2020), only those participants who met such requirements were considered for the review. The most frequent country of origin of participants was USA (*n* = 15; 17.24%), followed by multi-country (*n* = 13; 14.94%), and Germany (*n* = 12; 13.79%). The rest of the studies were set in the other 19 countries.

For their DIF or invariance analyses, 56.32% of studies picked a cognitive target variable, and 43.68% of studies picked a non-cognitive or context target variable. Table X.1 in Supplementary Material X (Navarro-González et al., 2023j) displays all of the target variables used.

Risk of Bias in Studies

In Supplementary Material XI (Navarro-González et al., 2023k), a summary of the risk of bias assessments for each study is provided. The justifications for assessments and scores are available in Supplementary Material XII (Navarro-González et al., 2023l). Overall, the majority of studies had low risk of bias (62 out of 87; 71.26%), 25 had medium risk of bias, and none of them had high risk of bias.

Results of Individual Studies

DIF/Measurement Invariance Analyses

Table 1 shows the frequency of usage of each DIF detection/measurement invariance method across the studies. Note that some studies used more than one technique.

Table 1.

Frequency of Usage of Each DIF Detection/Measurement Invariance Method.

DIF detection/measurement invariance method	***f***
Parametric	
Multigroup CFA	30
Multigroup SEM	9
IRT Rasch analysis	9
Logistic regression/Ordinal logistic regression	8
IRT log-likelihood test	5
IRT difficulty parameters comparison	4
MIMIC	4
Latent class models	3
Lord’s chi-square statistic	3
Area measures	3
LH method	2
Cognitive diagnosis models	2
Explanatory Item Response Modeling	2
IRT + ANOVA	2
Multilevel analysis	1
Mixture IRT model	1
Logistic discriminant function analysis	1
Generalized Linear Mixed Models	1
Logits of the gender-specific item-difficulty scores	1
Nonparametric	
Mantel-Haenszel	13
SIBTEST/poly-SIBTEST	6
Standardized Mean Difference	2

Note. CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modeling; MIMIC = Multiple-Indicator Multiple-Cause; SIBTEST = Simultaneous Item Bias Test.

The most popular analytic strategy is measurement invariance testing, through Multigroup CFA and/or Multigroup SEM, carried out by 36 out of 87 studies. Then, 13 studies used the Mantel-Haenszel procedure as the DIF detection method, followed by nine that performed an IRT Rasch analysis, eight that performed logistic regressions (LR/OLS), and six that performed a SIBTEST or poly-SIBTEST. Overall, 68 (78.16%) studies used only parametric strategies, whereas 10 (11.49%) used only nonparametric ones (nine used both).

Regarding the characteristics of the instruments analyzed, 36 (41.38%) studies used dichotomously scored instruments, whereas 42 (48.27%) used polytomously scored ones. Four (4.59%) used both types of instruments. Moreover, most (*n* = 65; 74.71%) used only instruments with closed-ended items. Only four (4.60%) used instruments with open-ended items, and 14 (16.09%) used both types of instruments. The majority (*n* = 64; 73.56%) did not specify the administration mode, whereas 10 (11.50%) administered their instruments online, 11 (12.64%) performed a paper-and-pencil administration, and two (2.30%) had a mixed mode of administration. Most studies (86.1%) used instruments with 50 or fewer items, with the mean number of items being 30.40 (*SD* = 35.92).

A summarizing table (Table XIII.1) with all previously described information for each study is presented in Supplementary Material XIII (Navarro-González et al., 2023m). Note that Mahmud and Nur (2018) did not perform any analysis of DIF or measurement invariance, so their study has not been included in this section.

Explanations for DIF/Lack of Measurement Invariance

In Supplementary Material XIV (Navarro-González et al., 2023n), a table is presented (Table XIV.1) summarizing DIF/measurement invariance approaches across all studies. We followed Li et al.'s (2021) categorization to classify studies into three categories based on their approaches to sources of DIF/lack of measurement invariance. First, 48 (56.47%) studies were sorted into Category A (studies that have not examined sources of DIF/lack of measurement invariance). Studies from Category A are those that detected items with DIF but did not perform any further analysis, that eliminated items with DIF from further analyses (e.g., Cheng et al., 2011, p. 205: "Once an item with substantial DIF was identified, it would be removed from further analysis"), and that met measurement invariance or did not encounter items with DIF (e.g., Seo et al., 2016, p. 55: "There were no items indicative of DIF between boys and girls"). Then, six (7.06%) studies were sorted into Category B, because they mentioned sources of DIF/lack of measurement invariance but did not examine them (e.g., Lyons-Thomas et al., 2014, p. 29: "One future direction of this study, or any of those which examine gender DIF, would be to investigate sources of DIF").

Finally, 31 (36.47%) studies were sorted into Category C (studies that have examined sources of DIF/lack of measurement invariance). Table 2 shows all strategies used by studies included in Category C to examine sources of gender DIF/lack of measurement invariance. Note that Chen and Jiao (2014), Cho and Cohen (2010), and Tsaousis et al. (2020) examined sources of latent DIF instead of gender DIF. Even though those three studies are included in Category C, their strategies to examine sources of latent DIF are not displayed in Table 2.

Table 2.

Frequency of Usage of Strategies to Examine Sources of Gender DIF/Lack of Measurement Invariance.

Examination of sources of gender DIF/lack of measurement invariance strategies	***f***
Qualitative	
Content analysis	9
Didactical interpretation of results	1
Interview	1
Focus groups	1
Expert appraisal	1
Delphi technique	1

Examination of sources of gender DIF/lack of measurement invariance strategies	***f***
Quantitative	
Multigroup SEM	4
Multilevel regression analysis	2
ANOVA	1
OLS regression models	1
Multinomial logistic regression	1
Multiple regression analysis	1
Generalized Linear Mixed Models	1
Propensity scores	1
Confirmatory approach proposed by Shealy and Stout (1993)	1
Analytic scoring analysis	1
Poly-BW indices	1
Comparison of distractor response curves	1

Note. SEM = Structural Equation Model; OLS = Ordinal Logistic Regression.

The most frequent strategy used by studies from Category C to examine sources of DIF was the content analysis on DIF items; this procedure was used to categorize them and explore sources of DIF related to item characteristics (nine out of 29) (e.g., Kalaycioğlu & Berberoğlu, 2011). Overall, 14 studies (45.16% of studies from Category C) used at least a qualitative approach. For example, Yildirim and Büyüköztürk (2018) used focus groups to examine bias in DIF items. On the other hand, some studies have used a quantitative approach, such as multigroup SEM (e.g., Nalipay et al., 2019). Although Mahmud and Nur (2018) did not perform any analysis of DIF or measurement invariance and, consequently, this study was not classified into any of the three proposed categories, they used a qualitative strategy for addressing gender differences. Specifically, they carried out interviews to examine girls' and boys' response processes and to explore whether students' learning strategies in English were affected by gender differences.

Regarding causes of gender DIF/lack of measurement invariance reported by the 28 studies that examined such causes, 14 studies have searched for those causes on item domains, such as geometry, probability, reading literacy, etc. (e.g., Doudeen & Annabi, 2008; Innabi & Dodeen, 2018), 13 studies have searched the DIF/lack of measurement invariance causes on individual or cultural variables (e.g., Raufelder et al., 2015; Woitschach et al., 2019), 12 studies have searched them on item characteristics, such as item format (e.g., Taylor & Lee, 2012; Zenisky et al., 2004), and 9 studies have searched on differential cognitive strategies to respond to items (e.g., Doudeen & Annabi, 2008; Kalaycioğlu & Berberoğlu, 2011). Note that some studies have searched for the causes in more than one kind of source.

Results of Syntheses

DIF/Measurement Invariance Analyses

To examine how invariance testing/DIF detection methods were applied, we tried to find patterns and relationships in the data, by synthesizing all previous information extracted from studies. We cross-tabulated between all methods reported and studies' characteristics, such as the sample size or the instruments' characteristics. These tables (Tables XV.1-XV.5) can be found in Supplementary Material XV (Navarro-González et al., 2023o). The most frequent characteristics associated with each method are mentioned below.

Regarding target instruments' characteristics, the DIF detection/invariance testing methods more associated with instruments with polytomous items were multigroup CFA, multigroup SEM, multilevel analysis (MLA), IRT Rasch analysis, logistic discriminant function analysis, and poly-SIBTEST. On the other hand, the methods more associated with instruments with dichotomous items were LR/OLS, LCA, MIMIC, the LH method, Lord's chi-square statistic, area measures, IRT log-likelihood test (IRTLR), mixture IRT model, explanatory item response modeling (EIRM), IRT + ANOVA, IRT difficulty parameters comparison, generalized linear mixed models (GLMM), logits of the gender-specific item-difficulty scores, Mantel-Haenszel, and SIBTEST. The cognitive diagnosis models and the standardized mean difference (SMD) were associated with both types of scoring systems. Almost all methods were associated with instruments with closed-ended items except for the logistic discriminant function analysis. The LH method, Lord's chi-square statistic, IRT + ANOVA, IRT difficulty parameters comparison, GLMM, and SMD are associated with instruments with both closed and open-ended items. As for the administration mode, methods associated with a paper-and-pencil administration were MLA, IRTLR, EIRM, IRT Rasch analysis, and SIBTEST/poly-SIBTEST, whereas methods associated with a computer-based administration were multigroup CFA, multigroup SEM, and IRT difficulty parameters comparison. MIMIC has been associated with both types of administration. Finally, looking at the number of items, almost all methods were associated with instruments with a number of items between 10 and 99. Multigroup SEM, MLA, and IRT + ANOVA were also associated with instruments with less than 10 items, and GLMM and SMD were associated with instruments with 100 or more items.

Regarding sample size, almost all methods were associated with large sample sizes (1000 or more participants) except for logits of the gender-specific item-difficulty scores, which were associated with medium sample sizes (between 100 and 999 participants). EIRM, IRT + ANOVA, cognitive diagnosis models, and SMD were associated with both medium and large sample sizes.

All previously discussed information is summarized in Table XVI.1 from Supplementary Material XVI (Navarro-González et al., 2023p), which can be used as a guide to know in which situations the usage of each technique is more recommended. For example, if one wanted to examine DIF on responses from a large sample to an instrument with polytomous and closed-ended items, said table could be consulted to see which techniques are mostly related to those characteristics and come to the conclusion that an IRT Rasch analysis or a poly-SIBTEST could be carried out in this case.

Finally, we examined the pros and cons of the DIF detection/invariance testing methods. The studies provided pros and cons for almost all methods; they were not provided for GLMM, IRT + ANOVA, IRT difficulty parameters comparison, logits of the gender-specific item-difficulty scores, MLA, and SMD. For example, the studies stressed that multigroup CFA permits to "meaningfully interpret group differences" (Jansen et al., 2014, p. 15), "verify whether the two gender groups shared an identical measurement structure" (Chang, 2019, p. 5), and examine "all aspects of heterogeneity across groups" (Raufelder et al., 2015, p. 4); on the other hand, the studies highlighted that full invariance may be "too strict and unrealistic" (Korpershoek et al., 2019, p. 8) and that this procedure is less parsimonious and "should be supplemented by DIF analyses on the item level" (Hatlevik et al., 2017, p. 23). As another example, the studies commented that the Mantel-Haenszel procedure is "a most frequently applied DIF statistic" (Chen & Jiao, 2014, p. 81) and a "well-defined and well-established" model (Lee & Geisinger, 2014, p. 320), robust to missing data, useful with both large and small samples, "highly consistent with other (DIF detection) methods" (Kalaycioğlu & Berberoğlu, 2011, p. 470), useful with both dichotomous and polytomous items (with Generalized Mantel-Haenszel), and more sensitive than other methods. Nevertheless, they also commented that the Mantel-Haenszel procedure cannot address nonuniform DIF, requires the test to be unidimensional, and is sensitive to the sample size (Chen & Jiao, 2014, p. 81: "The Mantel-Haenszel chi-square is much easier to reject the null hypothesis under a large sample size, even with a small effect size").

Methods for which more pros were reported are MIMIC, Mantel-Haenszel, SIBTEST/poly-SIBTEST, and latent class analysis (LCA), whereas methods for which more cons were reported are multigroup CFA and Mantel-Haenszel. All pros and cons reported for each technique are displayed in Table XVI.1 from Supplementary Material XVI

(Navarro-González et al., 2023p), with the aim to complement the rest of information provided for each technique, in order to guide authors in their decision to use them.

Explanations for DIF/Lack of Measurement Invariance

We have also tried to find patterns by linking each DIF detection/invariance testing method with Li et al.'s (2021) categorization, with the scope to discover whether some DIF detection/invariance testing methods are more associated with the examination of sources of DIF/lack of measurement invariance, which is the cornerstone of the third generation of DIF proposed by Zumbo (2007). A cross table (Table XVII.1) can be found in Supplementary Material XVII (Navarro-González et al., 2023q).

Methods associated with the examination of sources of DIF or lack of measurement invariance (methods mostly used by studies from Category C) were LR/OLS, LCA, MLA, Lord's chi-square statistic, IRTLR, mixture IRT model (MMixIRTM), IRT + ANOVA, Mantel-Haenszel, and SIBTEST/poly-SIBTEST. These findings are consistent with some advantages that studies reported for some of these methods. For example, Tsaousis et al. (2020) mentioned that LCA could "attempt to identify possible sources of DIF across the covariate's levels" (p. 3). On the other hand, some methods were not used by any of the studies from Category C. These are area measures, IRT difficulty parameters comparison, and logits of gender-specific item-difficulty scores.

Regarding the pros and cons of the strategies used to examine sources of DIF/lack of measurement invariance, studies provided only pros for interviews, focus groups, comparison of distractor response curves, didactical interpretations for results and propensity scores. Ferretti and Giberti (2020) highlighted that interviews allow "understanding the cognitive processes adopted by the students" (p. 6). For focus groups, Yildirim and Büyüköztürk (2018) stressed that focus groups "enable deeper and richer information to be reached than from individual views" (p. 451). For a didactical interpretation of results, Cascella et al. (2020) stressed that this strategy "can enrich the meaning of the quantitative datum" (p. 2). For comparison of distractor response curves, Cascella et al. (2021) stated that this analysis is "very informative because it provides a visual interpretation of response patterns" (p. 8). Finally, for propensity scores, Lee and Geisinger (2014) showed that they can "contribute to hypothesizing a causal inference on DIF" (p. 331) and help researchers "to explore the causes of DIF easily with a traditional statistical DIF method" (p. 317). They also stressed that these models "could benefit test fairness for both test developers and test takers" (p. 331), they "do not have any limitations for the inclusion of a large number of covariates, (…) relatively less affected by the inclusion of unassociated (nonsignificant) covariates" (p. 332), they "can lead to less bias than regression models" (p. 332), and they are "quite robust to model misspecifications" (p. 332).

Studies provided pros and cons for the Delphi technique and the confirmatory approach proposed by Shealy and Stout (1993, as cited in Mendes-Barnett & Ercikan, 2006). Yildirim and Büyüköztürk (2018) stated that the Delphi technique was useful in addressing complex problems and highlighted the advantage of having a group opinion, which they considered to be "more effective than individual" (p. 453). Nevertheless, they also commented that "the success of Delphi studies largely depends on the choice of the relevant experts in the field." (p. 453). For the confirmatory approach proposed by Shealy and Stout (1993, as cited in Mendes-Barnett & Ercikan, 2006), Mendes-Barnett and Ercikan (2006) stated that this strategy can test "hypothesized sources of DIF statistically" (p. 291). Nevertheless, they also commented that "although identifying patterns of gender DIF (…) is one step beyond identification of DIF status of these items, such patterns (…) do not provide guidance regarding reasons for DIF or what educators need to do to alter these patterns. More information about the cognitive processes used by boys and girls (…) is needed to explain differential functioning" (p. 302). A summary of those can be found in Table XVIII.1 from Supplementary Material XVIII (Navarro-González et al., 2023r).

Finally, regarding causes of gender DIF/lack of measurement invariance found by the studies, main conclusions are presented: 1) as for item domains, some studies (e.g., Doudeen & Annabi, 2008; Innabi & Dodeen, 2018; Pae, 2012; Taylor & Lee, 2012) show that items tend to favor girls when they assess reading comprehension, interpretation and estimation of data, patterns, and ratio or proportions, and also when items are unapplied (e.g., algebra items with variables or polynomials); while items tend to favor boys when they include content related to graph, grammar, vocabulary, scientific design, probability, logarithms, exponents, decimal numbers, geometry, measurement, the concepts of speed and velocity, technology, and Earth and space science, and also when items are applied real life problems (e.g., problems that include people); 2) as for the cognitive strategies needed to respond to items, some studies (e.g., Doudeen & Annabi, 2008; Kalaycioğlu & Berberoğlu, 2011) found that girls seemed to be favored when they had to give a fixed or algorithmic answer based on their knowledge, while boys seemed to be favored when they had to give some kind of judgment (e.g., predictions, inferences, interpretations, comparisons...), and in items that required a higher level of mental processing; 3) as for item characteristics, some studies (e.g., Taylor & Lee,2012; Zenisky et al., 2004) have found that girls are favored by constructed-response or open-response items, whereas boys are favored by multiple-choice items, and items with visual stimuli such as graphs or tables; and 4) as for individual and cultural variables, some studies (e.g., Ferretti & Giberti, 2020; Raufelder et al., 2015, Woitschach et al., 2019) have shown that girls are more vulnerable to mother pressure regarding test anxiety, more insecure and less self-confident in mathematics, and more likely to answer items correctly when their country has a higher human development index. Boys, on the other hand, benefit more from father support regarding test anxiety.

Reporting Biases

Regarding outcome reporting bias, only one study (Kaye-Tzadok et al., 2017) showed some kind of incoherence between outcomes reported in the “methods” and “results”. That is, the authors claimed in the “methods” section that “some basic questions on children’s age, gender (boy or girl) and living arrangements were included in the study, and details of adults with whom children live with were included in this analysis” (p. 4), but only gender was introduced in the analysis (some information about family structure was presented in the descriptive statistics of the sample, but these data were not included in the main analysis).

Regarding missing data, findings for administration modes have to be interpreted with caution, due to a rate of missing data of 73.56%. The authors may have omitted this information from their articles because 1) they have analyzed available databases from well-known international testing projects like PISA and considered readers already know how assessment instruments are administered; and/or 2) they thought that readers could “guess” that a traditional paper-and-pencil administration had taken place.

Certainty of Evidence

We used the GRADE-CERQual approach (Lewin et al., 2015) to assess the confidence in cumulative evidence. Full judgments were displayed in a CERQual Qualitative Evidence Profile (Table XIX.1) that can be found in Supplementary Material XIX (Navarro-González et al., 2023s). Out of our 29 findings, we graded 11 as high-confidence findings. Among them were the following findings: the most used DIF detection/invariance testing methods, the pros and cons of such methods, the methods associated with instruments with a specific type of scoring system, the methods associated with instruments with closed-ended items, the methods associated with instruments with a medium number of items, the methods associated with large sample sizes, some techniques used to explain sources of DIF/lack of measurement invariance, and where do authors search for such sources.

Nevertheless, we graded eight findings as low-confidence findings, mainly due to the methodological limitations and concerns about the data's adequacy. Among them were the following findings: specific DIF detection/invariance testing methods (e.g., cognitive diagnosis models, GLMM, SMD) associated with certain scoring systems or sample sizes, the methods associated with different administration modes, and the pros and cons of techniques for explaining sources of DIF/lack of measurement invariance, and sources of DIF/lack of measurement invariance related to individual and cultural variables. These findings have to be interpreted with caution.

Discussion

The main aim of the present mixed studies systematic review was to describe how measurement invariance/DIF detection was addressed, quantitatively and qualitatively, when studying the gender gap in national and international educational testing projects involving middle and high school students. Results indicated that the most common invariance testing/DIF detection methods were multigroup CFA and Mantel-Haenszel. This is not surprising because multigroup CFA is a widespread technique used to test measurement invariance in applied studies, and Mantel-Haenszel has been the traditional DIF detection method used, as Chen and Jiao (2014) stated when discussing its advantages.

Choosing a DIF detection/invariance testing method or other should be determined by the research questions and objectives, as well as the research context. We encourage readers to consult Table XVI.1 from Supplementary Material XVI (Navarro-González et al., 2023p) as a guide for opting for the most adequate statistical techniques considering the characteristics of the educational testing project under study. For example, looking at the results of the mixed studies systematic review, we recommend using the following techniques taking item format and test length into account: 1) methods such as multigroup CFA, MLA, IRT Rasch analysis or poly-SIBTEST, when having tests or questionnaires with polytomous items; 2) methods such as LR, LCA, Mantel-Haenszel or SIBTEST, when examining dichotomous items; and 3) methods such as multigroup SEM or MLA, when testing tests or questionnaires with few items. These findings can be complemented with results and guidelines from an extent literature along the history of development of DIF detection and invariance testing techniques (e.g., Hidalgo & Gómez-Benito, 2010; Oliveri et al., 2012).

When researchers especially want to move forward and examine the potential causes of DIF/lack of measurement invariance within the "third generation of DIF" framework, some DIF detection/invariance testing methods can provide additional validity evidence and insights into such causes. By conceiving DIF as a product of item characteristics and/or the testing situation (Zumbo, 2007), researchers could have a more comprehensive understanding of DIF, making it possible to identify sources of DIF that could be endangering validity and leading to mistaken and unfair comparative interpretations of test scores. Identifying such sources is especially relevant when examining gender DIF in educational assessments, given that an unfair comparative interpretation could have negative consequences for girls' decisions and expectations. In this respect, this systematic review could guide on which DIF detection/invariance testing methods are most associated with the examination of sources of DIF/lack of measurement invariance. Researchers interested in uncovering DIF/lack of measurement invariance causes should resort to methods such as LCA, MMixIRTM or SIBTEST, findings in line with the methods recommended for the "third generation of DIF methodology" within the third and four statistical framework (Zumbo et al., 2015).

On the qualitative side, the most common strategy for examining the sources of DIF, not surprisingly, was item content analysis. Content analysis of DIF items has been a usual strategy even in most traditional DIF studies, but these first attempts to explore sources of DIF could now be seen as rather descriptive or superficial. Third generation of DIF studies raises the necessity of more complex and comprehensive explanations, looking at other contextual sources in the testing situation, such as administration modes and the response processes of respondents. Even though we found that some studies used qualitative methods such as interviewing and focus groups, none of them used either cognitive interviewing or web probing. These two methods are widely known in survey research as effective and useful methods to explore response processes (e.g., Benítez et al., 2019). The necessity of using such strategies to examine sources of DIF/lack of measurement invariance is clear. As Mendes-Barnett and Ercikan (2006) said, "more information about the cognitive processes used by boys and girls (…) is needed to explain differential functioning" (p. 302). We encourage future researchers to incorporate cognitive interviewing and web probing as methods to examine sources of DIF/lack of measurement invariance so that they can obtain evidence of differences in the response processes that groups of test-takers follow, derived from their differential experiences and socialization. Mixed-methods DIF/measurement invariance studies integrating quantitative results from DIF detection/invariance testing techniques and qualitative findings from methods such as cognitive interviewing or focus groups can make a difference in investigating gender DIF and/or lack of measurement invariance.

Conducting mixed-methods DIF/measurement invariance research can be a first step to reduce construct-irrelevant variance when assessing girls and boys, because the information obtained through this kind of studies could serve as a base, for example, to adjust item content by changing certain topics that result more familiar for one gender group with other ones that are equally familiar for both groups –e.g., boys are more familiar with solving life-related and challenging mathematics problems because, due to gender stereotypes and roles, they are more encouraged to take risks and try new things than girls (e.g., Innabi & Dodeen, 2018)–, and balancing item format and the kind of cognitive strategies needed to respond to items. By examining and analyzing test items following this approach, fairer and more valid assessments could be developed.

In addition to item-related causes of gender DIF/lack of measurement invariance, we have also identified context-related ones, following the pattern of different levels of causes of DIF that Zumbo et al. (2015) mentioned in the Ecological Model of Item Response. That is why changes can also be made in order to diminish the gender gap at a sociocultural level. The sociocultural expectations that parents, teachers and peers have based on gender roles and stereotypes can act as a self-fulfilling prophecy affecting adolescents' self-concept and motivation, leading girls to be less self-confident and more insecure in mathematics (e.g., Heyder et al., 2020; Innabi & Dodeen, 2018; Yildirim & Büyüköztürk, 2018). Parents' and teachers' influence has been proven (e.g., Cascella et al., 2020; Ferretti & Giberti, 2020; Nalipay et al., 2019), so a non-stereotyping teaching and socialization for children and adolescents is needed. In order to provide a fairer and more equal teaching, teachers should encourage girls to familiarize with problem-solving strategies, and to be confident to explore and try new things and solutions (e.g., Chen & Jiao, 2014; Doudeen & Annabi, 2008).

As for the limitations in this review, this paper outlines the problems we have found with missing data in certain variables (e.g., administration mode). In addition, some findings were graded as low-confidence ones due to the methodological limitations of studies and/or to the fact that they did not have enough studies supporting them. We tried to make these limitations as clear as possible so as to warn readers to interpret those findings with caution.

However, this review is intended to be a sort of compendium of the most used DIF detection/invariance testing methods and approaches, with the scope of describing and linking them to the third generation of DIF, in addition to describing the most used strategies for examining sources of gender DIF/lack of measurement invariance. Thus, it can be useful as a first guide to future researchers who would want to examine sources of DIF and/or lack of measurement invariance in obtaining validity evidence for instruments used in educational assessment and other fields, because this systematic review provides an overview of how DIF and measurement invariance are being addressed at the moment, serving this knowledge as a first base to critically make decisions to advance research on the gender gap in educational testing. Future research could therefore position itself on the right track in developing a more comprehensive understanding of the gender gap and in making fairer comparative interpretations of test scores.

References

American Psychological Association [APA] (2022). *APA Thesaurus of Psychological Index Terms.* Retrieved from https://psycnet.apa.org/thesaurus

Ayuso, N., Murillo, A.C., & Cerezo, E. (2020). Gender Gap in STEM: A cross-sectional study of primary school students' self-perception and test anxiety in mathematics. *IEEE Transactions on Education, 64*(1), 40-49. DOI: 10.1109/TE.2020.3004075

Benítez, I., van de Vijver, F., Padilla, J.L. (2019). A Mixed Methods Approach to the Analysis of Bias in Cross-cultural Studies. *Sociological Methods & Research,* 1-34. DOI: 10.1177/0049124119852390

Cascella, C., Giberti, C., & Bolondi, G. (2020). An analysis of Differential Item Functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. *Studies in Educational Evaluation, 64*, 100819. DOI: 10.1016/j.stueduc.2019.100819

Cascella, C., Giberti, C., & Bolondi, G. (2021). Changing the Order of Factors Does Not Change the Product but Does Affect Students' Answers, Especially Girls' Answers. *Education Sciences*, *11*(5), 201. DOI: 10.3390/educsci11050201

Chang, C.C. (2019). Development of Ocean Literacy Inventory for 16- to 18-Year-Old Students. *SAGE Open, 9*(2), 215824401984408. DOI: 10.1177/2158244019844085

Chen, Y.F. & Jiao, H. (2014). Exploring the Utility of Background and Cognitive Variables in Explaining Latent Differential Item Functioning: An Example of the PISA 2009 Reading Assessment. *Educational Assessment*, *19*(2), 77-96. DOI: 10.1080/10627197.2014.903650

Chen, M.Y., & Zumbo, B.D. (2017). Ecological Framework on Item Responding as Validity Evidence: An Application of Multilevel DIF Modeling Using PISA Data. In B.D. Zumbo, & A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research.* Springer. DOI: 10.1007/978-3-319-56129-5

Cheng, Y.Y., Chen, L.M., Liu, K.S., & Chen, Y.L. (2011). Development and Psychometric Evaluation of the School Bullying Scales: A Rasch Measurement Approach. *Educational and Psychological Measurement*, *71*(1), 200-216. DOI: 10.1177/0013164410387387

Cho, S.J. & Cohen, A.S. (2010). A Multilevel Mixture IRT Model With an Application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336-370. DOI: 10.3102/1076998609353111

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER Tool for Qualitative Evidence Synthesis. *Qualitative Health Research, 22*(10), 1435-1443. DOI: 10.1177/1049732312452938

Doudeen, H.M., & Annabi, H.A. (2008). Sex-Related Differential Item Functioning (DIF) Analysis on TIMSS. *Dirasat, Educational Sciences, 35*, 697-705. Retrieved from https://journals.ju.edu.jo/DirasatHum/article/view/1807

Ferretti, F. & Giberti, C. (2020). The Properties of Powers: Didactic Contract and Gender Gap. *International Journal of Science and Mathematics Education*, *19*(8), 1717-1735. DOI: 10.1007/s10763-020-10130-5

Fetters, M.D., Curry, L.A., & Creswell, J.W. (2013). Achieving integration in mixed methods designs — Principles and practices. *Health Services Research, 48*(6)*,* 2134-2156. DOI: 10.1111/1475-6773.12117

Hatlevik, O.E., Scherer, R., & Christophersen, K.A. (2017). Moving beyond the study of gender differences: An analysis of measurement invariance and differential item functioning of an ICT literacy scale. *Computers & Education*, *113*, 280-293. DOI: 10.1016/j.compedu.2017.06.003

Heyder, A., Weidinger, A.F., & Steinmayr, R. (2020). Only a Burden for Females in Math? Gender and Domain Differences in the Relation Between Adolescents' Fixed Mindsets and Motivation. *Journal of Youth and Adolescence, 50*, 177-188. DOI: 10.1007/s10964-020-01345-4

Hidalgo, M.D., & Gómez-Benito, J. (2010). Differential Item Functioning. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 36-44). Elsevier. DOI: 10.1016/B978-0-08-044894-7.00242-6

Hong, Q.N., & Pluye, P. (2019). A Conceptual Framework for Critical Appraisal in Systematic Mixed Studies Reviews. *Journal of Mixed Methods Research, 13*(4), 446-460. DOI: 10.1177/1558689818770058

Hong, Q.N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews, 6*(61)*,* 1-14. DOI: 10.1186/s13643-017-0454-2

IBM Corp. (2013). *IBM SPSS Statistics for Windows* (Version 22) [Software]. IBM Corp. Retrieved from https://www.ibm.com/es-es/spss

Innabi, H., & Dodeen, H. (2018). Gender differences in mathematics achievement in Jordan: A differential item functioning analysis of the 2015 TIMSS. *School Science and Mathematics, 118*, 127-137. DOI: 10.1111/ssm.12269

Jansen, M., Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences*, *30*, 11-21. DOI: 10.1016/j.lindif.2013.12.003

Kalaycioğlu, D.B. & Berberoğlu, G. (2011). Differential Item Functioning Analysis of the Science and Mathematics Items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment*, *29*(5), 467-478. DOI: 10.1177/0734282910391623

Kan, A., & Bulut, O. (2014). Examining the Relationship Between Gender DIF and Language Complexity in Mathematics Assessments. *International Journal of Testing, 14*(3), 245-264. DOI: 10.1080/15305058.2013.877911

Kaye-Tzadok, A., Kim, S.S., & Main, G. (2017). Children's subjective well-being in relation to gender—What can we learn from dissatisfied children? *Children and Youth Services Review*, *80*, 96-104. DOI: 10.1016/j.childyouth.2017.06.058

Korpershoek, H., King, R.B., McInerney, D. M., Nasser, R.N., Ganotice, F.A., & Watkins, D.A. (2019). Gender and cultural differences in school motivation. *Research Papers in Education*, *36*(1), 27-51. DOI: 10.1080/02671522.2019.1633557

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-74. DOI: 10.2307/2529310

Lee, H. & Geisinger, K.F. (2014). The Effect of Propensity Scores on DIF Analysis: Inference on the Potential Cause of DIF. *International Journal of Testing*, *14*(4), 313-338. DOI: 10.1080/15305058.2014.922567

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C.J., Gülmezoglu, M., Noyes, J., Booth, A., Garside, R., & Rashidian, A. (2015). Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med, 12*(10), 1-18. DOI: 10.1371/journal.pmed.1001895

Li, H., Hunter, C.V., & Bialo, J.A. (2021). A revisit of Zumbo's third generation of DIF: How are we doing in language testing? *Language Assessment Quarterly,* 1-27. DOI: 10.1080/15434303.2021.1963253

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *EĞİTİM VE BİLİM - Education and Science,* *39*(172), 20-32. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/2873

Mahmud, M. & Nur, S. (2018). Exploring Students' Learning Strategies and Gender Differences in English Language Teaching. *International Journal of Language Education*, *2*(1), 51-64. DOI: 10.26858/ijole.v2i1.4346

Mendes-Barnett, S. & Ercikan, K. (2006). Examining Sources of Gender DIF in Mathematics Assessments Using a Confirmatory Multidimensional Model Approach. *Applied Measurement in Education*, *19*(4), 289-304. DOI: 10.1207/s15324818ame1904_4

Nalipay, M.J.N., Cai, Y., & King, R.B. (2019). Why do girls do better in reading than boys? How parental emotional contagion explains gender differences in reading achievement. *Psychology in the Schools*, *57*(2), 310-319. DOI: 10.1002/pits.22330

OECD (2019). *PISA 2018 Results (Volume III): What School Life for Students' Lives.* OECD Publishing. DOI: 10.1787/acd78851-en

Oliveri, M.E., Olson, B.F., Ercikan, K., & Zumbo, B.D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing, 12*(3), 203-223. DOI: 10.1080/15305058.2011.617475

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan — a web and mobile app for systematic reviews. *Systematic Reviews, 5*, 2-10. Retrieved from https://www.rayyan.ai/. DOI: 10.1186/s13643-016-0384-4

Padilla, J.L., Benítez, I., & van de Vijver, F.J.R.. (2018). Addressing Equivalence and Bias in Cross-cultural Survey Research Within a Mixed Methods Framework. In Timothy P. Johnson, Beth-Ellen Pennell, Ineke A.L. Stoop, & Brita Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC).* Wiley. DOI: 10.1002/9781118884997

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29*(4), 533-554. DOI: 10.1177/0265532211434027

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E.,… Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*(71), 1-9. DOI: 10.1136/bmj.n71

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews.* ESRC Methods Programme. Retrieved from https://database.inahta.org/article/2684

QSR International (2022). *NVivo* (Version 13) [Software]. Retrieved from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Raufelder, D., Hoferichter, F., Ringeisen, T., Regner, N., & Jacke, C. (2015). The Perceived Role of Parental Support and Pressure in the Interplay of Test Anxiety and School Engagement Among Adolescents: Evidence for Gender-Specific Relations. *Journal of Child and Family Studies*, *24*(12), 3742-3756. DOI: 10.1007/s10826-015-0182-y

Sakellariou, C., Fang, Z. (2021). Self-efficacy and interest in STEM subjects as predictors of the STEM gender gap in the US: the role of unobserved heterogeneity. *International Journal of Educational Research, 109*, 1-14. DOI: 10.1016/j.ijer.2021.101821

Sandelowski, M, Barroso, J., & Voils, C.I. (2007). Using qualitative metasummary to synthesize qualitative and quantitative descriptive findings. *Research in Nursing and Health, 30,* 99-111. DOI: 10.1002/nur.20176

Schwabe, F., McElvany, N., & Trendtel, M. (2014). The School Age Gender Gap in Reading Assessment: Examining the Influences of Item Format and Intrinsic Reading Motivation. *Reading Research Quarterly, 50*(2), 219-232. DOI: 10.1002/rrq.92

Seo, D., Taherbhai, H., & Frantz, R. (2016). Psychometric Evaluation and Discussions of English Language Learners' Listening Comprehension. *International Journal of Listening*, *30*(1-2), 47-66. DOI: 10.1080/10904018.2015.1065747

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., & the PRISMA-P Group (2014). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ, 349,* 1-25. DOI: 10.1136/bmj.g7647

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: the development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice, 18*, 746-752. DOI: 10.1111/j.1365-2753.2011.01662.x

Taylor, C.S., & Lee, Y. (2012). Gender DIF in Reading and Mathematics Tests With Mixed Item Formats. *Applied Measurement in Education, 25*, 246-280. DOI: 10.1080/08957347.2012.687650

The jamovi project. *jamovi* (Version 2.2.2) [Software]. Retrieved from https://www.jamovi.org

Tsaousis, I., Sideridis, G.D., & AlGhamdi, H.M. (2020). Measurement Invariance and Differential Item Functioning Across Gender Within a Latent Class Analysis Framework: Evidence From a High-Stakes Test for University Admission in Saudi Arabia. *Frontiers in Psychology*, *11*, 622. DOI: 10.3389/fpsyg.2020.00622

UNESCO (2022). *UNESCO Thesaurus.* Retrieved from http://vocabularies.unesco.org/browser/thesaurus/es

van de Vijver, F.J.R., & Leung, K. (2021). *Methods and data analysis for cross-cultural research.* Cambridge University Press. DOI: 10.1017/9781107415188

Woitschach, P., Zumbo, B.D., & Fernández-Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema, 31*(2), 194-203. DOI: 10.7334/psicothema2018.303

Yildirim, H. & Büyüköztürk, S. (2018). Using the Delphi Technique and Focus-Group Interviews to Determine Item Bias on the Mathematics Section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice, 18*(2), 447-470. DOI: 10.12738/estp.2018.2.0317

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2004). DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices. *Educational Assessment, 9*(1-2), 61-78. DOI: 10.1080/10627197.2004.9652959

Zumbo, B.D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly, 4*(2)*,* 223-233. DOI: 10.1080/15434300701375832

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Astivia, O.L.O., & Ark, T.K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12,* 136-151. DOI: 10.1080/15434303.2014.972559

Zumbo, B.D., & Padilla, J.L. (2019). The Interplay Between Survey Research and Psychometrics, with a Focus on Validity Theory. In Paul C. Beatty, Debbie Collins, Lyn Kaye, José-Luis Padilla, Gordon B. Willis, & Amanda Wilmot (Eds.) *Advances in Questionnaire Design, Development, Evaluation and Testing*. Wiley. DOI: 10.1002/9781119263685

Open Science

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.

Open Data: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported results (Navarro-González et al., 2023a-s).

Open Materials: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Navarro-González et al., 2023a-s).

Preregistration of Studies and Analysis Plans: This study was pre-registered on April 9th 2022 with an analysis plan ("Analyzing Measurement Invariance for Studying the Gender Gap in Educational Testing: A Protocol for a Mixed Studies Systematic Review", 2022). DOI: 10.17605/OSF.IO/XZGDQ

Open Analytic Code: The data is available on request from the author(s).

Data, supplementary materials and preregistration are available here: https://osf.io/rvqk7/

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023a). Supplementary Material I. Search strategy. DOI: https://osf.io/6rfms

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023b). Supplementary Material II. Data items. DOI: https://osf.io/yz5mx

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023c). Supplementary Material III. Outcomes. DOI: https://osf.io/6k4r8

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023d). Supplementary Material IV. Template for data extraction (QUAL) [Nvivo]. DOI: https://osf.io/xsfvq

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023e). Supplementary Material V. Template for data extraction (QUAN) [SPSS]. DOI: https://osf.io/acgsk

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023f). Supplementary Material VI. Data extraction (QUAL) [Nvivo]. DOI: https://osf.io/yanhw

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023g). Supplementary Material VII. Data extraction (QUAN) [jamovi]. DOI: https://osf.io/sjfyb

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023h). Supplementary Material VIII. References of included articles. DOI: https://osf.io/h6wpk

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023i). Supplementary Material IX. Study characteristics. DOI: https://osf.io/das35

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023j). Supplementary Material X. Target variables. DOI: https://osf.io/4v7te

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023k). Supplementary Material XI. Risk of bias assessment. DOI: https://osf.io/4nxq9

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023l). Supplementary Material XII. Critical appraisal. DOI: https://osf.io/83thk

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023m). Supplementary Material XIII. DIF/measurement invariance analyses and instruments' characteristics. DOI: https://osf.io/7nb8f

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023n). Supplementary Material XIV. DIF/measurement invariance approaches. DOI: https://osf.io/k4chx

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023o). Supplementary Material XV. Cross tables. DOI: https://osf.io/g6j3t

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023p). Supplementary Material XVI. Characteristics of DIF detection/invariance testing methods DOI: https://osf.io/um2v9

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023q). Supplementary Material XVII. Li et al. 's (2021) categorization across DIF detection/invariance testing methods. DOI: https://osf.io/jv9xf

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023r). Supplementary Material XVIII. Pros and cons of strategies for examination of sources of DIF/lack of measurement invariance. DOI: https://osf.io/by2xf

Navarro-González, M.C., Padilla, J.L., & Benítez, I. (2023s). Supplementary Material XIX. GRADE-CERQual Assessment. DOI: https://osf.io/28gv3