

TESIS DOCTORAL

Programa de Doctorado en Psicología

**DESARROLLO Y EVALUACIÓN DE MÉTODOS PARA
OBTENER EVIDENCIAS DE VALIDEZ SOBRE LOS
PROCESOS DE RESPUESTA A TESTS Y CUESTIONARIOS
EN ESTUDIOS TRANSCULTURALES**

María del Carmen Navarro González

DIRECTOR

José Luis Padilla García



**UNIVERSIDAD
DE GRANADA**

Departamento de Metodología de las Ciencias del Comportamiento

Granada, septiembre de 2025

Editor: Universidad de Granada. Tesis Doctorales.

Autora: María del Carmen Navarro González.

Diseño de portada: David León de la Torre.

La presente tesis doctoral ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España, a través del **Programa de Formación del Profesorado Universitario [FPU]**, en el marco del Subprograma Estatal de Formación del Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+i del Plan Estatal de Investigación Científica y Técnica y de Innovación 2017-2020 (Ref. FPU19/02908). La ayuda fue concedida a la doctoranda, quien ha estado vinculada a la misma desde octubre de 2020 hasta octubre de 2024.

La realización de esta tesis doctoral también ha sido posible gracias a la financiación ofrecida por el **Proyecto de Investigación “Desarrollo de métodos para mejorar la calidad de las medidas de salud y bienestar social en Andalucía: Validación de cuestionarios para encuestas y evaluaciones psicológicas”**, concedido por la Consejería de Conocimiento, Investigación y Universidad de la Junta de Andalucía a través del Fondo Europeo de Desarrollo Regional [FEDER] (Ref. PY20_00898).



Universidad de Granada

Programa de Doctorado en Psicología

Centro de Investigación Mente, Cerebro y Comportamiento (CIMCYC)

Departamento de Metodología de las Ciencias del Comportamiento

Facultad de Psicología

La doctoranda María del Carmen Navarro González y el director de la tesis José Luis Padilla García garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección del director de tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

*A mi padre,
porque sé que le hubiera gustado tener este libro entre sus manos.*

AGRADECIMIENTOS

AGRADECIMIENTOS

No han sido años fáciles.

Realmente, no por “culpa” de la tesis, pero ésta ha supuesto un reto más, un telón de fondo que me ha acompañado —para bien y para mal— en los que han sido, probablemente, los años más difíciles y duros de mi vida hasta ahora.

No es mi intención utilizar los agradecimientos de mi tesis como un “pañito de lágrimas”, pero sí quiero poner en contexto estos años. He visto de cerca la enfermedad y la muerte, y me he enfrentado a la realidad de la pérdida, todo esto dos veces. No tengo claro que la madurez y fortaleza que he ganado hayan compensado el sufrimiento y la indefensión que se apoderaron de mi familia, pero reconozco que son virtudes que me han ayudado y me ayudarán a relativizar otros retos y a afrontarlos con entereza.

Ese ha sido el caso de la tesis. Creo que, a pesar de que ha habido momentos de estrés y algo de miedo por si no conseguíamos “sacar las cosas” a tiempo, he podido mantener la calma casi sin problemas y disfrutar del proceso. Debido a mis circunstancias personales, ha sido una tesis “atípica”: sin estancia ni mención internacional, hecha en cinco años..., pero el aprendizaje que me ha permitido llevarme, tanto de mi área de estudio como de cómo funcionan la Ciencia y la Universidad, es de un valor incalculable.

El hecho de que esté escribiendo estas palabras es señal de que este viaje está llegando a su fin, aunque todavía, sinceramente, no termino de asimilarlo. Pero ya puedo decir que *lo hemos conseguido*. Y lo digo en plural porque esta tesis doctoral no habría sido posible sin mi director, **José Luis Padilla García**, que es la primera persona a la que me gustaría agradecer, puesto que su trabajo y mentoría han supuesto uno de los pilares fundamentales en el desarrollo de la tesis. Sé que no ha sido fácil estar pendiente de mí y dedicarle tiempo a mi tesis en estos últimos tiempos, por eso quiero que sepas que valoro realmente el tiempo que me has dedicado. Siempre aportando una visión nueva, *dándole una vuelta* a las cosas. Siempre resolviendo parsimoniosamente las dudas y los problemas que me surgen, *dándome paz*. Siempre dando la cara por mí. Siempre alentándome a que *me venda mejor* y a que de alguna manera “me crea” que las ideas que tengo que aportar son valiosas. Siempre preocupándote por mi bienestar y respetando mis ritmos y mis manías. Siempre poniéndome nerviosa por si no llegamos a tiempo, pero siempre acabándolo todo en plazo. Siempre recomendándome exposiciones de arte y otras actividades culturales que *sabes que me van a gustar*. Siempre enseñándome y guiándome en la investigación y en la docencia. Gracias por todo eso, y por todo lo que seguramente se me habrá olvidado mencionar, porque los siete años que llevo trabajando contigo —desde cuarto de carrera, con las prácticas externas del Grado y el TFG— dan para mucho, tanto a nivel profesional como a nivel personal.

Quiero agradecer también a todas las personas que he conocido en la Universidad —tanto en la de Granada como en otras—, y que me han acompañado a lo largo de este proceso. Muchas gracias al **Departamento de Metodología de las Ciencias del Comportamiento de la UGR**, me habéis acogido como nadie y habéis conseguido que me sienta como en casa en el departamento. En especial, muchas gracias a **Isabel** por ayudarme muchas veces a conceptualizar mis trabajos y aportarme ideas estupendas, a **Andrés** por ser tan increíblemente amable y echarme siempre una mano cuando he estado en apuros, y a **Luisma** por hacerme

siempre reír (*¡Luisma, socorro, grapadora!*) y darme consejos vitales muy valiosos. Los tres sois personas que, además de ser inteligentísimas y de aportarme mucho a nivel investigador, habéis sido claves a nivel personal y habéis hecho que me sienta de verdad como una más. También gracias a **Nacho** y a **M^a José**, por ser los mejores directores de departamento. No me olvido de **Carolina** y **Macarena**, a quienes siempre les agradeceré haber confiado en mí para trabajar con ellas en su Proyecto de Innovación Docente, con aquella beca Ícaro. En definitiva, muchas gracias a todos por haberme acogido con los brazos abiertos.

Muchas gracias también a mis “doctorandos *seniors*”, que evidentemente —y merecidísimamente—, ya son todos doctores. **Celia**, de quien heredé escritorio y despacho y fue la primera “doctoranda mayor” que me acogió; **Jorge**, que es literalmente un ser de luz; **Álvaro**, con quien me he reído más que con ninguno; y el amigo **Tao**, con quien comparto orígenes (seguro que el profe Ángel se pone contentísimo de que los dos seamos doctores). Muchas gracias a los cuatro por tantos momentos buenos y tantos consejos útiles.

Muchísimas gracias a quienes han sido además de mis compañeros, mis amigos en este camino. **Mariela**, igual que me dijiste tú a mí, *has sido más que una compañera de despacho*. A lo largo de estos años, nos hemos convertido en confidentes la una de la otra, y hemos ido transitando este camino mano a mano. Gracias por todo eso. **Belén**, aunque no hemos sido compañeras de despacho, como si lo hubiéramos sido, porque sí que hemos sido compañeras de camino, y todo lo que le digo a Mariela podría decírtelo a ti también. Hemos confiado la una en la otra y hemos compartido tanto desgracias como alegrías. Gracias por ser como eres. ¡La próxima vez que coincidamos las tres tenemos que volver a ir a un *karaoke*! **David**, que llegaras más tarde no ha sido un inconveniente para que tú también te convirtieras en amigo mío, rápidamente te cogimos cariño y fuiste un “bro” más. *Grazie per il tuo senso*

dell'umorismo e la tua sensibilità. Que no se te olvide que tenemos pendiente una revancha a las Magic.

También gracias al resto de compañeros doctorandos —de mi departamento y también de otros—, con quienes he compartido espacio muchas veces en el CIMCYC. Especialmente, gracias a **Dörte** por esas conversaciones en las que tú y yo nos entendemos perfectamente —aunque muchas hayan sido en la distancia—. **Rocío**, ya te queda poquito para terminar a ti también, ¡qué bien! **Lucía**, bienvenida al departamento, si te tratan aunque sea la mitad de bien que a mí, te garantizo que te sentirás como en casa.

No puedo terminar mis agradecimientos a las personas de la academia sin agradecer enormemente a **Lola Hidalgo**, ya que, a pesar de que no llegué a Murcia en el mejor momento, se encargó de acogerme con un cariño increíble. Los ratos de trabajo y creatividad en tu despacho fueron todo un honor para mí.

De algún u otro modo, todos vosotros habéis sido imprescindibles para que esta tesis pudiera desarrollarse y nacer, pero, sobre todo, los estudios de la tesis no hubieran sido posibles sin **las chicas y los chicos participantes** que compartieron conmigo algunas de sus experiencias más dolorosas. Muchísimas gracias por abrirme vuestro corazón y por hablar con sinceridad de temas que para nada son agradables de recordar, de vivir, e incluso de presenciar. Gracias también a **Beatriz y Luis** por todo el esfuerzo.

También quiero agradecer a muchas personas que no forman parte del mundo académico y que han contribuido de forma más “indirecta” a la elaboración de esta tesis, pero no por ello su influencia ha sido menos importante.

Gracias a las personas que me han acompañado académicamente, desde mis compañeros de la carrera, esos “**Y PIN**”, hasta los compañeros que tengo actualmente en el

Máster de Psicología General Sanitaria. Gracias en especial a estos últimos por hacer más llevaderas las clases y los momentos de tensión cuando había mil trabajos que entregar y por la comprensión que habéis tenido conmigo, sabiendo que yo llevaba la recta final de mi tesis a la par que el máster.

Quiero agradecer también a algunas personas que han sido amigas mías desde hace ya muchos años y son figuras de apoyo muy importantes para mí. **Cris, Azu**, ya sabéis lo que os quiero y lo que agradezco que hayáis estado siempre apoyándome y animándome cuando la vida me sobrepasaba. Muchas gracias por la calidez y ternura de Cris cuando escucha mis problemas, y por lo natural que le resulta el uso del humor para levantarme el ánimo. Muchas gracias también por las experiencias compartidas con Azu en el inicio del doctorado, y las conversaciones kilométricas en las que nos desahogábamos de todo. **Pablo**, aun en la distancia, siempre he sentido que podía contar contigo y eres uno de mis mejores amigos. Fuiste una de esas personas que estuvieron conmigo cuando mi padre ingresó por última vez, junto con Cris, Azu y Ángel, y eso no lo voy a olvidar nunca. Gracias. **Alberto, Noe**, muchas gracias por estar siempre ahí y por seguir siendo tan “*chismosos*” como siempre. **David**, gracias por acceder a hacer la portada de esta tesis y sobre todo porque da igual los años que pasen, cada vez que nos vemos es como si no hubiera pasado el tiempo.

Ángel, tú has sido mi piedra angular. No me quiero extender mucho contigo porque podría escribir libros enteros enumerando las cosas por las que quiero darte las gracias, pero creo que tú ya las sabes. Has sido y eres mi mayor apoyo y la persona que cada día se estruja los sesos para hacerme la vida un poquito más fácil. Has sido también la persona que más ha conseguido darme las fuerzas necesarias para que no me venzan la desesperación o la ansiedad, y para que yo siguiera con mi tesis hasta el final, pero siempre dándome a mí la agencia de

tomar la última decisión. Me faltan palabras para agradecer toda la confianza que tienes depositada en mí. Muchas gracias por querer ser mi compañero en la vida.

Quiero agradecer también a **la familia de Ángel**, porque desde el principio me han acogido como si fuera una hija más y se esfuerzan por transmitirme siempre todo el cariño que me tienen, y por supuesto también a **mi propia familia** que, aunque no sea muy numerosa, rebosa amor. **Abuela**, sé que me estás cuidando desde donde quiera que estés. **Abuelo**, la complicidad que tenemos tú y yo poca gente la entiende. **Dante**, aunque seas un perro, eres el único que siempre me saca una sonrisa.

Mamá, gracias por haberme hecho ser como soy. Gracias por enseñarme que los fracasos y las equivocaciones no tienen porqué ser un problema, y que las personas sabias son las que aprenden de esas vivencias y se hacen más capaces. Gracias por inculcarme la importancia de la independencia y por enseñarme que el conocimiento te hace más libre. Gracias por hacerme saber que el límite está donde yo lo quiera poner. Gracias por haber inculcado en mí la bondad. Gracias por todas tus buenas intenciones y por tu incondicionalidad. Gracias por acompañarme en el camino de aprender a ser más fuertes.

Papá, aunque ya no estés aquí, tengo claro que no morirás hasta que yo me muera, porque tu presencia y tu influencia siguen vivas dentro de mí. Gracias por haber existido y gracias por haber sido mi padre, mi “*compinche*”. Gracias por haber sido la persona con la que abría largos debates, que despertó en mí el sentido del humor, la preocupación social y el amor por la música. Gracias por haberte sentido tan orgulloso de mí. Probablemente, sin tus ánimos ni siquiera hubiera empezado a hacer un doctorado, y sé que te quedaste con las ganas de ver mi tesis terminada, como te quedaste con las ganas de ver otras tantas cosas. Por eso, te la dedico a ti.

Termino de escribir esta tesis llorando, pero son lágrimas agridulces, llenas de nostalgia y de recuerdo, llenas de gratitud y de amor, y llenas también de satisfacción por el cierre de una etapa que me ha visto crecer como persona más que ninguna otra.

Gracias por haberme acompañado a transitarla.

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS

PLANTEAMIENTO GENERAL	25

CAPÍTULOS TEÓRICOS

CAPÍTULO I: INTRODUCCIÓN	37

1. Validez y justicia en las mediciones	39
1.1. La validez y el proceso de validación	39
1.2. La equivalencia y su relación con la validez	44
1.3. La justicia en las mediciones	49
1.4. Evidencias de validez basadas en los procesos de respuesta para abordar la falta de invarianza y la justicia en las mediciones	52
2. El <i>gender gap</i> como problema de equivalencia y justicia en las mediciones	55
3. El <i>gender gap</i> en medidas educativas y del clima escolar	60
3.1. Equivalencia y justicia en medidas educativas	60
3.2. El caso del <i>bullying</i> : contextualización	61
3.3. El caso del <i>bullying</i> : diferencias entre chicos y chicas	63

CAPÍTULO II: OBJETIVOS	67

CAPÍTULO III: METODOLOGÍA	73

1. Metodología mixta para el estudio de los procesos de respuesta	75
1.1. Introducción a la metodología mixta	75
1.2. Diseños mixtos y las diferentes formas de integración	77
2. Un diseño mixto secuencial explicativo como estrategia de validación	80
2.1. Fase preliminar: revisión sistemática mixta de la evidencia	81
2.2. Fase 1: obtención y análisis de la información cuantitativa	83
2.3. Fase 2: obtención y análisis de la información cualitativa	85
2.4. Fase 3: integración de los resultados cuantitativos y cualitativos	86

CAPÍTULOS EMPÍRICOS

CAPÍTULO IV: ESTUDIO I. ANALYZING MEASUREMENT INVARIANCE FOR STUDYING THE GENDER GAP IN EDUCATIONAL TESTING: A MIXED STUDIES SYSTEMATIC REVIEW 91

Abstract	95
Rationale	97
Objectives	99
Methods	100
Results	107
Discussion	122
Declarations	127
References	129

CAPÍTULO V: ESTUDIO II. TESTING COMPARABILITY OF GENDER BULLYING VICTIMIZATION ASSESSMENTS THROUGH A LATENT CLASS APPROACH 139

Abstract	143
Introduction	145
Methods	154
Results	158
Discussion	164
Declarations	171
References	173

CAPÍTULO VI: ESTUDIO III. DO DIFFERENT EXPERIENCES OF BULLYING IMPACT RESPONSE PROCESSES? A MIXED-METHODS APPROACH TO OBTAIN VALIDITY EVIDENCE OF BULLYING SCALES' SCORES 183

Resumen	187
Abstract	189
Introduction	191

Methods	197
Results	200
Discussion	209
Declarations	215
References	217
CAPÍTULO VII: DISCUSIÓN GENERAL	223
1. Discusión general e integración de hallazgos	225
2. Limitaciones e investigaciones futuras	237
3. Implicaciones prácticas	239
4. Conclusiones	242
REFERENCIAS	245
MATERIAL SUPLEMENTARIO	265

PLANTEAMIENTO GENERAL

PLANTEAMIENTO GENERAL

Garantizar la validez y justicia de las interpretaciones en las que se basan las decisiones tomadas a partir de las mediciones aportadas por tests, escalas y cuestionarios que evalúen problemas y fenómenos importantes para la vida de las personas, como el *bullying*, requiere aportar evidencias de validez. Los *Standards for Educational and Psychological Testing* (*Estándares para Pruebas Educativas y Psicológicas*), publicados por la American Educational Research Association [AERA], la American Psychological Association [APA], y el National Council of Measurement in Education [NCME] (2014), contemplan la existencia de cinco fuentes de evidencias de validez, siendo la basada en los procesos de respuesta una de ellas y la que menos atención ha recibido hasta hace poco, tanto en la investigación como en el ámbito profesional (e.g., Padilla y Benítez, 2014; Sireci y Benítez, 2023). A través de dicha fuente de evidencias de validez se pretende abordar “la adecuación entre el constructo y la naturaleza detallada del desempeño o respuesta efectivamente empleada por los examinandos” (AERA et al., 2014, p. 16). Obteniendo evidencias de validez sobre los procesos de respuesta, se puede explorar la que podemos denominar “ruta” cognitiva que las personas llevan a cabo mientras contestan a los ítems de un test o cuestionario, permitiendo esto analizar si los procesos cognitivos “previstos” en la definición semántica y la tabla de especificaciones (e.g., interpretación de los conceptos o expresiones clave, recuperación de la información en la memoria, elaboración de la respuesta, etc.), son los finalmente puestos en práctica por las personas que contestan al instrumento de evaluación.

A menudo han surgido dificultades y limitaciones a la hora de examinar estos procesos de respuesta y de aportar evidencias basadas en esta fuente de evidencias de validez (Benítez et al., 2019; Sireci y Benítez, 2023), ya que tanto en los *Standards for Educational and Psychological Testing* (AERA et al., 2014), como en la literatura disponible faltan indicaciones claras sobre cómo llevar a cabo estudios de validación centrados en los procesos de respuesta. Además, debido a la complejidad que entraña obtener información sobre dichos procesos, estas dificultades surgen especialmente cuando se trata de abordar esta cuestión sólo a través de métodos cuantitativos. Para solucionar este problema, se ha propuesto el uso de métodos mixtos que combinen los métodos cuantitativos con métodos cualitativos que permitan profundizar y abordar de manera más comprehensiva los fenómenos de estudio, superando las limitaciones que acarrear los métodos cuantitativos y cualitativos cuando se aplican solos (Creswell y Plano Clark, 2011). Así, diversos estudios (e.g. Castillo y Padilla, 2013; Benítez et al., 2019; Padilla y Benítez, 2014), han hecho uso de los métodos mixtos para aportar evidencias de validez basadas en los procesos de respuesta.

Este tipo de evidencias de validez son especialmente útiles cuando las interpretaciones de las puntuaciones se utilizan para comparar subgrupos de personas definidos en función del género, la lengua, la cultura, etc. Si los procesos de respuesta de varios grupos de personas difieren, se podría dar lugar a un problema de equivalencia a nivel de constructo (He y van de Vijver, 2012), comprometiendo la comparabilidad de las interpretaciones de las puntuaciones entre esos grupos, lo que supone un problema de validez (Zumbo y Padilla, 2019). Las evidencias de validez basadas en los procesos de respuesta contribuyen a “responder preguntas sobre diferencias en el significado o interpretación de puntajes de pruebas entre subgrupos relevantes de examinandos” (AERA et al., 2014, p. 16), permitiendo explorar los procesos de respuesta de las personas evaluadas. Por esta misma razón son también evidencias de validez muy útiles a la hora de asegurar la justicia de las mediciones, puesto que estas evidencias

permiten comprobar si los ítems reflejan el mismo constructo o fenómeno para todas las personas que los contestan, requisito para que una medida sea justa y no beneficie o perjudique a uno de los grupos de personas debido a características irrelevantes al constructo (AERA et al., 2014), que cuestionen la validez de las interpretaciones comparativas en que se basan las decisiones. Por tanto, no sería ni ética ni psicométricamente correcto asumir, o de algún modo “dar por hecho”, que no existen diferencias entre los grupos de comparación en los procesos de respuesta a instrumentos de evaluación del constructo objetivo, y es siempre conveniente corroborar si existen o no dichas diferencias para asegurar que se está “midiendo” el mismo constructo en todos los grupos de comparación. En ese sentido, deberíamos evitar hacer interpretaciones culturalmente homogéneas, puesto que “las diversas formas de ser humano” necesariamente interactúan con la medición y la evaluación (Zumbo, 2021; Zumbo, 2023; Zumbo, en prensa). Por ejemplo, históricamente se ha estudiado el “género” en términos binarios que correspondían con el sexo biológico, atribuyendo las diferencias de género en los procesos de respuesta a esas diferencias biológicas. Sin embargo, la experiencia ha demostrado que “las diversas formas de ser humano”, en palabras del profesor Bruno D. Zumbo (Zumbo, 2021), hacen que la visión binaria del género ligada al sexo biológico, pierda sentido, y que las causas de las diferencias de género en los procesos de respuesta a los ítems de los instrumentos de evaluación deban también ser buscadas en otras variables contextuales o situacionales ligadas a las diferencias en los procesos de socialización (Zumbo, 2021).

El Modelo Ecológico de Respuesta a los Ítems propuesto por Zumbo et al. (2015), puede dar cuenta de cómo las “diversas formas de ser humano” interactúan con la medición. Según dicho modelo, numerosos factores personales y contextuales, cuya influencia puede variar en los diferentes grupos evaluados, pueden moldear los procesos de respuesta a los ítems y amenazar la justicia de la medición y la comparabilidad de las interpretaciones de las puntuaciones. Por ejemplo, la socialización diferencial de género, mediada por estereotipos,

roles y “mandatos” de género, podría constituir uno de estos factores. Así, es ética y metodológicamente crucial asegurar la comparabilidad y la justicia de las mediciones entre mujeres y hombres, descartando la aparición de sesgos en el proceso de medición que puedan, de alguna manera, beneficiar o perjudicar a uno de los grupos (e.g., García-Calvente et al., 2010).

Especialmente vital es asegurar la justicia de las mediciones que se hacen en el contexto educativo a chicos y chicas, ya que numerosas decisiones importantes para la vida del estudiantado derivan de dichas mediciones. Por ejemplo, la detección de casos de *bullying* o acoso escolar podría derivarse de mediciones aportadas por tests, cuestionarios o *screenings*, así como la intervención posterior, compuestos por ítems a los que chicas y chicos respondieran siguiendo “rutas” cognitivas diferentes. Esto podría ser un “arma de doble filo”, puesto que si las interpretaciones de las puntuaciones están sesgadas, las intervenciones en una cuestión tan importante para la salud mental y física del estudiantado, también estarán sesgadas y podrían no ser lo suficientemente efectivas.

El *bullying* o acoso escolar es un problema que requiere de rigor y compromiso a la hora de ser abordado, ya que importantes consecuencias para la vida de quienes lo sufren están en juego. Ser víctima de *bullying* puede acarrear problemas como depresión, ansiedad, baja autoestima, soledad, absentismo escolar, consumo de drogas, comportamientos sexuales de riesgo o incluso, el suicidio (e.g., Holt et al., 2015; Kim y Leventhal, 2008; Lee, 2021; Swearer y Hymel, 2015; UNESCO, 2019). En España, según un estudio realizado por la Unidad de Psicología Preventiva de la Universidad Complutense de Madrid y la Fundación ColaCao (Díaz-Aguado et al., 2023), un 6,2% del estudiantado encuestado se identifica como “víctima de bullying”, aunque el 19,2% admite haber sufrido algún tipo de abuso por parte de sus iguales, dando a entender que el problema es más común de lo que a priori podría parecer.

Cuando se hacen interpretaciones comparativas de las puntuaciones a escalas de *bullying* de chicas y chicos, podemos encontrarnos con los problemas de validez y justicia anteriormente expuestos. La literatura previa ha demostrado que existen diferencias sustanciales en cuanto a las experiencias de victimización que viven ambos géneros (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Hanci y Akça, 2020; Lee et al., 2022; Rivers y Smith, 1994; Rosen y Nofziger, 2021; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022): los chicos suelen recibir más *bullying* en general que las chicas y más agresiones directas, las cuales son las que están dirigidas a causar un daño inmediato (agresiones físicas, como pegar o empujar, y agresiones verbales, como insultar o amenazar); las chicas, por su parte, suelen recibir más agresiones indirectas, cuyo objetivo es dañar la reputación o las relaciones interpersonales de la víctima (agresiones relacionales como criticar, difundir rumores o excluir). Estas diferencias en las experiencias vitales con la victimización de *bullying* entre chicas y chicos podrían suscitar diferentes representaciones cognitivas del *bullying*, que a su vez podrían dar lugar a diferencias en los procesos de respuesta a los ítems de victimización de *bullying*. En ese caso, no se estaría midiendo el mismo fenómeno en ambos grupos sociodemográficos, amenazando, como se apuntaba anteriormente, la equivalencia, la validez de las interpretaciones comparativas de las puntuaciones, y la justicia en la medición.

Por tanto, el **objetivo general** de la tesis doctoral ha sido desarrollar y mostrar la aplicación de estrategias “mixtas” de validación basadas en los procesos de respuesta, que integren métodos cuantitativos y cualitativos para evaluar la validez de las interpretaciones comparativas de las respuestas de chicas y chicos a ítems sobre victimización de *bullying*, con el fin de asegurar la comparabilidad de la interpretación de las puntuaciones y la justicia de la medición entre chicos y chicas. A pesar de haber centrado el trabajo en un contexto y una escala en concreto, el propósito es que la estrategia mixta de validación que se desarrolla se pueda generalizar y trasladar a diferentes contextos de evaluación en los que se desee realizar

interpretaciones comparativas de las puntuaciones de diferentes grupos sociodemográficos, lingüísticos y/o culturales.

La tesis doctoral se estructura en un total de siete capítulos. En el **primer capítulo** se introducen los conceptos de validez y justicia en las mediciones, se discute cómo la falta de invarianza o equivalencia podría poner en peligro dicha validez y dicha justicia de las interpretaciones comparativas de las puntuaciones de un test, se expone cómo podría explorarse esta falta de invarianza a través de aportar evidencias de validez basadas en los procesos de respuesta, y se discute concretamente cómo esta problemática puede entorpecer la detección de casos de *bullying* en adolescentes, debido a que el *gender gap* que se encuentra al evaluar las experiencias de victimización de *bullying* podría reflejar una falta de invarianza. En el **segundo capítulo** se presentan los objetivos específicos de investigación. En el **tercer capítulo** se detalla la metodología que se ha seguido a lo largo de la tesis doctoral. En los **capítulos cuarto, quinto y sexto** se incluyen los tres estudios que conforman la tesis doctoral. Por último, en el **séptimo capítulo** se desarrolla una discusión general derivada de los hallazgos de los estudios realizados así como la síntesis de los resultados cuantitativos y cualitativos, y se exponen las principales limitaciones, se proponen líneas de investigación futuras e implicaciones prácticas de los hallazgos, y se comentan las principales conclusiones que se derivan de la tesis doctoral.

Para finalizar, resulta relevante destacar que, a pesar de que se ha tratado de caracterizar a cada capítulo con una identidad y finalidad propias, podrá haber información reiterada a lo largo de los mismos. Esto se debe a que los estudios de los capítulos tercero, cuarto y quinto fueron escritos como artículos independientes para ser publicados en diferentes revistas científicas, por lo que es posible que alguna información referente a teorías o conceptos se

repita entre ellos. Cabe también mencionar que, a lo largo de la tesis doctoral, los términos “instrumento”, “prueba”, “test” y “cuestionario” se usarán a menudo de forma intercambiable.

CAPÍTULOS TEÓRICOS

CAPÍTULO I: INTRODUCCIÓN

1. VALIDEZ Y JUSTICIA EN LAS MEDICIONES

1.1. La validez y el proceso de validación

Asegurar la validez de las interpretaciones de las mediciones aportadas por tests y cuestionarios es crucial para poder tomar decisiones importantes a partir de dichas interpretaciones. Por ejemplo, es necesario comprobar, para el fenómeno objetivo de los estudios realizados en la tesis —con los que se pretende además “ilustrar” la aproximación metodológica “mixta” para obtener evidencias de validez sobre los procesos de respuesta—, si las puntuaciones que aporta el test o cuestionario empleado, representan realmente y de manera fidedigna las experiencias de victimización de *bullying* de las personas que lo contestan, —siendo esto especialmente importante si se van a utilizar las puntuaciones de dicho test para detectar casos de *bullying* en un centro escolar, o para evaluar un programa de intervención, etc.—, y actuar en consecuencia.

Sin embargo, la conceptualización de la validez no fue siempre como la que mayoritariamente presentan los manuales y artículos actualmente, que han puesto el foco de la validación en la interpretación que sostiene el uso previsto de las mediciones. Las primeras recomendaciones formales que se elaboraron para tratar de establecer un consenso en cuanto al concepto de validez fueron las *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, publicadas por la American Psychological Association [APA] (1954). En dichas *Recommendations*, se estableció que la validez era “el grado en el que el test es capaz de alcanzar ciertos objetivos” (APA, 1954, p. 13), y se incluyeron cuatro tipos de validez (APA,

1954): validez de contenido (en qué medida el contenido del test representa fielmente las situaciones objeto de medición), validez predictiva (en qué medida las predicciones que hace el test se confirman con la evidencia posterior), validez concurrente (en qué medida las puntuaciones del test se corresponden con criterios o medidas obtenidas al mismo tiempo), y validez de constructo (en qué medida el test representa adecuadamente la conceptualización teórica del constructo objeto de medición).

Más adelante, en 1966, la APA, la American Educational Research Association [AERA], y el National Council of Measurement in Education [NCME] publicaron la siguiente versión de las *Recommendations*, con el título de *Standards for Educational and Psychological Tests and Manuals* (AERA et al., 1966). En esta edición de los *Standards*, la validez concurrente y la validez predictiva se agruparon en el concepto de validez de criterio y los “tipos” de validez pasaron a llamarse “aspectos”. La siguiente edición, llamada *Standards for Educational and Psychological Tests* (AERA et al., 1974), mantuvo la misma conceptualización. También se mantuvo en la versión siguiente, los *Standards for Educational and Psychological Testing*, (AERA et al., 1985), aunque en dichos *Standards* se estableció por primera vez que la validez es un concepto unitario.

La concepción actual de la validez, en términos generales, empezó a fraguarse a partir de los trabajos de Samuel J. Messick (1931-1998), y los de Michael T. Kane (1992) y su propuesta del “*argument-based approach*”, que defiende que es necesario establecer cuáles son las interpretaciones propuestas para las puntuaciones de los tests y guiar la validación en función de ellas. Así, los *Standards for Educational and Psychological Testing* de 1999 (AERA et al., 1999) introdujeron ya una nueva visión de la validez que sigue vigente en la versión actual de los *Standards for Educational and Psychological Testing* (AERA et al., 2014). La validez actualmente se conceptualiza como unitaria y se define como el “grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos

propuestos de las pruebas” (AERA et al., 2014, p. 11). Así, no podría decirse que un test está “validado”, ya que la validación es un proceso constante y continuo, en el que se vuelve necesaria la obtención de evidencias de validez que apoyen o desmientan las interpretaciones que se hacen de las puntuaciones de los tests, para cada uso concreto que se le dé a dicho test. Además, el objeto de validación ya no es el instrumento en sí, sino la interpretación propuesta de las puntuaciones en dicho instrumento. Por ejemplo, el objetivo no sería ya determinar si un test que evalúa las experiencias de victimización de *bullying* es válido o no, sino reunir evidencias que avalen que las interpretaciones de las puntuaciones de personas concretas en ese test, obtenidas en un momento y contexto concreto, apoyan el uso previsto o toma de decisiones para las que se utilizarán dichas puntuaciones (e.g., comprobar si, en esa administración del test, mayores puntuaciones en el test significan mayores experiencias de victimización de *bullying*, y que esas personas son las que recibirán la intervención prevista). Esta nueva interpretación de la validez implica que el proceso de validación es plural, en diversos sentidos: un solo tipo de evidencia de validez no es suficiente para determinar la validez de una interpretación de unas puntuaciones, sino que se necesitan múltiples fuentes de evidencias para apoyar una interpretación dada; es necesario obtener evidencias de validez de forma continua cada vez que el test se administre y en cada contexto; y obtener las evidencias de validez ya no es solo responsabilidad de quienes elaboran los tests, sino también de quienes los administran, interpretan las mediciones y toman decisiones a partir de las mismas.

Esta nueva “idea” sobre la validez que trasladaron los *Standards for Educational and Psychological Testing* de 1999 (AERA et al., 1999) y que sigue vigente con los *Standards* actuales (AERA et al., 2014), viene acompañada de cinco fuentes de evidencias de validez para agrupar los estudios de validación en función de qué se obtiene y cómo. Estas fuentes no representan diferentes tipos de validez, sino que son diversas formas de aportar evidencias,

desde diferentes aristas o aspectos de la situación de aplicación de un test o cuestionario. Estas fuentes de evidencias son las siguientes:

- ❖ **Evidencias de validez basadas en el contenido.** Mediante esta fuente de evidencias, se obtiene información sobre “la relación entre el contenido de una prueba y el constructo que se intenta medir” (AERA et al., 2014, p. 14). Es decir, se obtienen evidencias sobre en qué medida el contenido de la prueba (e.g., comportamientos representados en los ítems, redacción de los ítems), se corresponde con la especificación del contenido que establecieron las o los profesionales que desarrollaron dicha prueba, tanto en términos de representación como de relevancia.
- ❖ **Evidencias de validez basadas en la estructura interna.** En esta fuente de evidencias de validez se analiza “el grado en que las relaciones entre ítems de la prueba y componentes de la prueba se ajustan al constructo sobre el que se basan las interpretaciones propuestas de puntajes de la prueba” (AERA et al., 2014, p. 17). Así, se comprueba si los ítems de la prueba reflejan las dimensiones en las que se espera que se divida el constructo objeto de medición.
- ❖ **Evidencias de validez basadas en relaciones con otras variables.** En esta fuente de evidencias se explora “la relación de los puntajes de la prueba con variables externas a la prueba”, teniendo en cuenta que “la interpretación prevista para un uso dado implica que el constructo debería relacionarse con algunas otras variables” (AERA et al., 2014, p. 17). En otras palabras, se comprueba en qué medida la relación existente entre las puntuaciones de la prueba y variables externas se corresponde con la relación que se espera que se dé entre dichas variables y el constructo objeto de medición.
- ❖ **Evidencias de validez basadas en los procesos de respuesta.** Esta fuente de evidencias de validez aborda “la adecuación entre el constructo y la naturaleza detallada del

desempeño o respuesta efectivamente empleada por los examinandos” (AERA et al., 2014, p. 16). De esta manera, se obtienen evidencias sobre en qué medida el proceso cognitivo que han llevado a cabo las personas que contestan la prueba es el esperado para el desempeño en la misma.

- ❖ **Evidencias de validez basadas en las consecuencias del uso de las pruebas.** Desde esta fuente de evidencias de validez, se pretende tener en cuenta de qué manera las consecuencias del uso de las pruebas “pueden ser parte de una afirmación que se extiende más allá de la interpretación o el uso de puntajes previsto” (AERA et al., 2014, p. 20). Consecuencias del uso de las pruebas que pueden poner en peligro la validez de las interpretaciones de las puntuaciones de dichas pruebas pueden ser las expectativas del uso de las pruebas, los juicios de valor asociados a términos presentes en los ítems, o la realización de afirmaciones sobre las puntuaciones de las pruebas que van más allá de las interpretaciones que se hacen de las puntuaciones.

A pesar de que esta nueva concepción de la validez ha alcanzado un consenso entre las y los profesionales de la Psicometría, varias revisiones de la literatura (Cizek et al., 2008; Flake et al., 2022; Hogan y Agnello, 2004; Shear y Zumbo, 2014; Sireci y Benítez, 2023; Zumbo y Chan, 2014), han evidenciado que pocos artículos proporcionan alguna evidencia de validez en sus mediciones, y que los que las proporcionan, lo hacen desde modelos de validez desfasados. En ese sentido, muy pocas veces se aportan o se exploran las evidencias de validez basadas en los procesos de respuesta y las basadas en las consecuencias del uso de los tests. Tal y como ya pusieron de manifiesto Padilla et al. (2006), existe una separación entre la teoría de la validez y la práctica, porque quizá, el hecho de entender la validación como un proceso continuo y prolongado hace que se perciba como demasiado exigente. No obstante, aportar evidencias de validez basadas en diversas fuentes es esencial para asegurar una interpretación de las puntuaciones fidedigna y justa.

1.2. La equivalencia y su relación con la validez

Una de las problemáticas que puede comprometer la validez de las interpretaciones de las puntuaciones de los tests es la falta de equivalencia o invarianza métrica. La invarianza o equivalencia métrica es el nivel de comparabilidad de las puntuaciones entre varios grupos sociodemográficos, lingüísticos y/o culturales, y supone una ausencia de sesgos en el proceso de medición (He y van de Vijver, 2012). Dicha equivalencia puede alcanzarse (He y van de Vijver, 2012): 1) a nivel de constructo (se está midiendo el mismo constructo a nivel teórico en todos los grupos), 2) a nivel métrico (la unidad de medida es la misma en todos los grupos en cuanto al intervalo o ratio, pero no en origen), o 3) a nivel escalar (equivalencia completa). Una falta de invarianza o equivalencia supondría no estar midiendo el mismo constructo o fenómeno en diferentes grupos de personas, con lo que la validez de las interpretaciones comparativas de las puntuaciones de dichos grupos de personas estaría comprometida dependiendo del nivel logrado de equivalencia. En ese sentido, una falta de invarianza o equivalencia se puede conceptualizar como un problema de validez (Zumbo y Padilla, 2019). Por ejemplo, inferir que las personas extranjeras (e.g., con diferente lengua materna, cultura, familiaridad con la evaluación, etc.), son menos inteligentes que las personas de nuestro país porque han puntuado menos en un test de inteligencia cuyos ítems son dependientes de la cultura sería erróneo, puesto que las diferencias en las puntuaciones se deberían realmente a que los ítems no evalúan lo mismo y de la misma manera en ambos grupos culturales. Así, es necesario comprobar si la invarianza se da o no cada vez que se administre una prueba a grupos de personas diferentes que posteriormente deseen compararse en función de sus puntuaciones en dicha prueba. El análisis del nivel de invarianza o equivalencia y la detección de posibles sesgos en el proceso de medición pueden plantearse como formas de buscar y aportar evidencias de validez para apoyar las interpretaciones comparativas que se hacen de las puntuaciones de los diferentes grupos sociodemográficos, lingüísticos y/o culturales.

Una de las estrategias que se han usado en Psicometría para detectar una falta de invarianza es el análisis del funcionamiento diferencial de los ítems o *differential item functioning* (DIF). De hecho, los *Standards for Educational and Psychological Testing* (AERA et al., 2014) proponen el DIF como estrategia para aportar evidencias de validez basadas en la estructura interna, y también se ha visto que es una herramienta que puede resultar útil para aportar evidencias de validez basadas en otras fuentes (Gómez-Benito et al., 2018). El DIF ocurre cuando personas de diferentes grupos sociodemográficos, lingüísticos y/o culturales, con el mismo nivel del constructo objeto de medición, presentan diferencias en la forma de contestar a los ítems (e.g., Chen y Zumbo, 2017). La presencia de DIF puede “dar pistas” acerca de la existencia de sesgos que estén influyendo en la manera que han tenido las personas de responder a los ítems, dando lugar a diferencias entre grupos que en realidad serían artefactos del proceso de medición. Bruno D. Zumbo (2007) hace un análisis de lo que llama las “generaciones del DIF”, para tomar consciencia de cómo se ha abordado la problemática del DIF a lo largo de la historia. Estas generaciones son:

- ❖ **Primera generación del DIF.** En esta generación comenzó a hablarse de la problemática de los sesgos en los ítems o *item bias*, y surgió la preocupación por el hecho de que las diferencias de puntuaciones entre grupos diferentes de personas se debiera realmente al sesgo en la medida y no a diferencias reales en el constructo objeto de medición. El término DIF empezó a usarse para reemplazar al de *item bias* y se empezaron a aplicar técnicas estadísticas para detectarlo.
- ❖ **Segunda generación del DIF.** Esta segunda generación empezó cuando el término DIF se extendió y se introdujeron nuevas y mejores técnicas estadísticas para su detección, tales como el Mantel-Haenszel (Holland y Thayer, 1988), los modelos de Teoría de Respuesta al Ítem (TRI), o el *simultaneous item bias test* o SIBTEST (Shealy y Stout,

1993). Así, esta generación se caracterizó por el énfasis en el desarrollo de dichas técnicas con el objetivo de detectar los ítems con DIF de la manera más precisa posible y no confundir el DIF con el “impacto” para, normalmente, eliminar dichos ítems de los tests finales por resultar problemáticos.

❖ **Tercera generación del DIF.** Esta generación del DIF comenzó cuando surgió la preocupación por explorar y averiguar las causas del DIF en las investigaciones (e.g., Angoff, 1993; Zumbo y Gelin, 2005), surgiendo la primera “semilla” de preocupación por las causas del DIF en los trabajos de Muthén y colaboradores (Muthén, 1985, 1988, 1989; Muthén et al., 1991; Muthén y Lehman, 1985). La presencia de DIF puede deberse a sesgos que están presentes en los ítems, pero también podría deberse a diferencias en variables que afectan a las respuestas a los ítems después de igualar los grupos de comparación en la medida objeto del instrumento de evaluación, por lo que en esta generación se empezó a considerar esencial explorar las causas del DIF encontrado para comprobar si las diferencias entre los grupos son reales o son artefactos del proceso de medición. Además, en el caso de que el DIF se debiera a sesgos, también se empezó a considerar importante conocer los factores, tanto del ítem como de la situación de evaluación, que han influido diferencialmente en los grupos de personas a la hora de contestar para, por un lado, hacer interpretaciones comparativas de las puntuaciones correctas y válidas, y por otro lado, estudiar de qué manera puede mejorarse la medida para tener en cuenta dichos factores y garantizar medidas más equivalentes y justas.

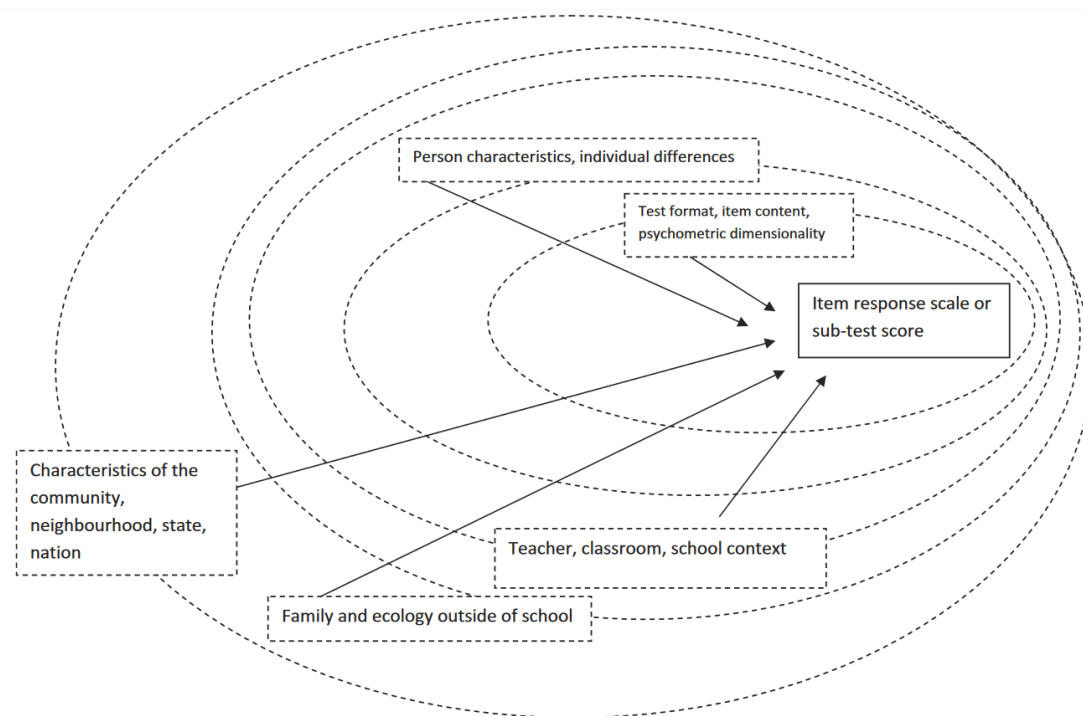
❖ **Cuarta generación del DIF.** Recientemente, Zumbo (en prensa) ha propuesto una nueva generación que está emergiendo en el momento actual. En esta generación se empieza a entender el DIF ya no como un problema a resolver, sino como “una ventana

hacia la diversidad humana” que, si se explora concienzudamente, aporta información sobre cómo las diferentes experiencias vitales y otras variables culturales influyen en la forma de interactuar con los tests. Así, en esta generación el objetivo no sería tanto señalar qué ítems tienen DIF o son problemáticos, sino entender mejor cómo son las personas que contestan a esos ítems y cómo sus “idiosincrasias” moldean su forma de contestar a los ítems, para diseñar instrumentos de evaluación que respeten esas idiosincrasias y las tengan en cuenta para que se realice una interpretación de las puntuaciones justa —y también válida, ya que esta interpretación sería más precisa y más ajustada a la realidad de la diversidad humana—. El estudio del DIF, por tanto, sería valioso para estudiar y conocer mejor, no solamente los tests o los ítems, sino el propio contexto de evaluación, explicando cómo funcionan los instrumentos de evaluación *en el mundo real* y respetando las diversas formas que tienen las personas de interactuar con los tests.

En línea con la filosofía de la tercera y la cuarta generación del DIF, Zumbo et al. (2015) proporcionan un marco teórico en el que buscar las causas del DIF y la falta de invarianza con su Modelo Ecológico de Respuesta a los Ítems o *Ecological Model of Item Responding*. En este modelo se defiende la idea de que existen diversos factores o variables, pertenecientes a diversos niveles (e.g., variables relativas a los ítems, variables relativas al contexto próximo de evaluación, variables de la persona que contesta al test, variables familiares, variables socioculturales o estructurales...), que pueden influir en la forma de responder a un test o cuestionario en un momento determinado. Zumbo et al. (2015) proporcionan un ejemplo de modelo ecológico (Figura 1) en el que se realiza una evaluación educativa. En este ejemplo, representan diversos niveles: 1) las características psicométricas del test, el formato del test, o el contenido de los ítems; 2) las características psicológicas y cognitivas de la persona que contesta; 3) el contexto escolar (profesores, clase, tipo de colegio...); 4) el contexto familiar; y

5) las características de la comunidad a la que pertenece la persona (barrio, ciudad, país...). Por tanto, las respuestas a los ítems no dependen solamente de la interacción entre las personas que contestan a los ítems y los ítems en sí, sino que esta interacción también refleja la influencia de múltiples factores entre los que está el contexto en el que está envuelta esta persona (Zumbo, en prensa).

Figura 1. Ejemplo de Modelo Ecológico de Respuesta a los Ítems.



Nota. Tomado de *An ecological model for item responding* (p. 139), por Zumbo et al., 2015, *Language Assessment Quarterly*, 12.

En ese sentido, diferentes variables pertenecientes a diversos niveles pueden influir en la manera en la que las personas contestan a los tests y cuestionarios, modificando sus procesos de respuesta y generando así diferencias que después pueden traducirse en DIF o falta de invarianza. Por tanto, una manera efectiva de explorar estas variables origen del DIF o de la

falta de invarianza sería indagar en los propios procesos de respuesta de las personas (e.g., Padilla y Benítez, 2014). Así, teniendo como referencia el modelo, es posible aportar evidencias de validez basadas en los procesos de respuesta contando con factores en varios niveles de análisis, con el objetivo de validar las interpretaciones comparativas (e.g., Chen y Zumbo, 2017), al tiempo que se encuentran posibles explicaciones para las diferencias entre las puntuaciones de personas pertenecientes a diversos grupos sociodemográficos, lingüísticos y/o culturales.

1.3. La justicia en las mediciones

La falta de equivalencia o invarianza métrica, además de suponer un problema de validez, supone también un problema de justicia o imparcialidad en la medición. Según los *Standards for Educational and Psychological Testing* (AERA et al., 2014, p. 54), una prueba es justa o imparcial cuando “refleja los mismos constructos para todos los examinandos, y los puntajes de esta tienen el mismo significado para todos los individuos en la población prevista; una prueba imparcial no favorece ni desfavorece a algunos individuos debido a características irrelevantes para el constructo previsto”. En ese sentido, la falta de invarianza supone un problema para la justicia porque, si no se está midiendo el mismo constructo en los diversos grupos (no hay invarianza métrica), las puntuaciones de los diferentes grupos no pueden tener el mismo significado entre sí, con lo que las interpretaciones comparativas de esas puntuaciones no serían justas (ni válidas). De esta definición se deriva también que solamente unas interpretaciones “válidas” de las puntuaciones de un test pueden ser también justas, poniendo de manifiesto que validez y justicia son dos caras de la misma moneda. Si no se tienen en cuenta las numerosas variables contextuales que pueden influir en la forma de contestar un instrumento de evaluación, no se podrán llegar a asegurar que las interpretaciones de las puntuaciones de los tests realmente son válidas y justas, y tampoco se contará con medidas realmente inclusivas, puesto que no se estarían respetando las diversas formas que tienen las

personas de interactuar con los tests y los ítems, o en otras palabras, las “diversas formas de ser humano” (Zumbo, 2021; Zumbo, 2023; Zumbo, en prensa).

Siguiendo el mismo ejemplo que se usó en el subapartado anterior, al comparar las puntuaciones en un test de inteligencia dependiente de la cultura de personas extranjeras y de personas de nuestro país, —las primeras con diferente lengua materna, o cultura, o familiaridad con la evaluación, etc.—, se podría inferir que son menos inteligentes si tienen puntuaciones promedio menores. Esa interpretación de las puntuaciones, tal y como se discutió en el subapartado anterior, no es válida, y tampoco es justa, puesto que se basa en una medida que no cumplía con los requisitos de justicia e imparcialidad, ya que no evaluó lo mismo en ambos grupos de personas, claramente perjudicando al grupo de personas extranjeras al presentarles ítems dependientes de un “contexto cultural” o “lingüístico” que no es el suyo. Así, la diferencia encontrada en las puntuaciones de ambos grupos no se debería a una diferencia real en inteligencia, sino a que las personas extranjeras se encontraban en desventaja a la hora de contestar a los ítems de la prueba.

Beneficiar o perjudicar a un grupo de personas con respecto a otro, por razón de una característica que no es relevante para la evaluación, supone una serie de conflictos éticos a nivel tanto teórico como práctico. Por un lado, a nivel teórico, no preocuparse por realizar interpretaciones válidas y justas de las puntuaciones de diversos grupos de personas, implica un mal desempeño a la hora de hacer evaluaciones psicológicas y educativas y no estar llevando a cabo prácticas científicas éticas, al desvirtuar la calidad metodológica de la evidencia empírica que se obtendría a través de pruebas que no se han revisado a través de este prisma. Esta cuestión entraría en conflicto con los Artículos 6º, 17º y 18º del Código Deontológico del Psicólogo elaborado por el Colegio Oficial de Psicólogos [COP] (1987), que mencionan la importancia del rigor metodológico y científico en el ejercicio y la competencia profesional, y también con los Principios Éticos de Evaluación Psicológica también elaborados por el COP a

partir del Código Ético de la APA (1992), que enfatizan la importancia del rigor metodológico en la construcción de los tests y de la aplicación adecuada de los mismos para asegurar interpretaciones fidedignas.

Por otro lado, a nivel práctico, se podrían estar perpetuando ideas erróneas sobre algunos grupos de personas, que pudieran fortalecer o dar lugar a ciertos estereotipos, que a su vez llevaran a las y los profesionales a emitir profecías autocumplidas, a caer en sesgos confirmatorios en la investigación científica, o a tomar decisiones injustas, equivocadas y/o discriminatorias sobre, por ejemplo, otorgar o no un puesto de trabajo o aplicar o no un tratamiento psicológico. Esto no respetaría el Artículo 10º del Código Deontológico del Psicólogo elaborado por el COP (1987), el cual menciona la importancia de mantener una postura no discriminatoria y no sesgada para con las personas, ni los Principios Éticos de Evaluación Psicológica que la Comisión de Test del COP tomó del Código Ético de la APA (1992), que, a pesar de no hacer alusión directa a la justicia y su relación con la validez, subrayan la necesidad de no hacer un uso inapropiado de las técnicas de evaluación, de “ser conscientes de los límites de la certidumbre con la que pueden hacerse los diagnósticos, juicios y predicciones acerca de las personas”, de identificar en qué situaciones y para qué grupos de personas concretos es necesario hacer un ajuste en la aplicación o interpretación de las pruebas, y de tener en cuenta los aspectos del test y de la persona evaluada que pueden influir en la forma que tomen las interpretaciones de las mediciones que haga cada profesional.

Así, asegurar la justicia en las mediciones es necesario tanto ética como profesionalmente, ya que la justicia va ligada a la validez de las interpretaciones comparativas de las puntuaciones y a una toma de decisiones justa y precisa para cada una de las poblaciones involucradas.

1.4. Evidencias de validez basadas en los procesos de respuesta para abordar la falta de invarianza y la justicia en las mediciones

Aportar evidencias de validez basadas en los procesos de respuesta puede ser una estrategia útil para explorar las causas de la falta de invarianza (e.g., Padilla y Benítez, 2014; Sireci y Benítez, 2023), debido a que, según el Modelo Ecológico de Respuesta a los Ítems (Zumbo et al., 2015), diversas variables pueden influir diferencialmente en los procesos de respuesta de las personas pertenecientes a diferentes grupos sociodemográficos, lingüísticos y/o culturales, generando diferencias en sus formas de contestar a los ítems. Además, aportar evidencias de validez de los procesos de respuesta con el objetivo de averiguar qué variables son las que influyen de forma diferencial en los grupos de personas, puede, a su vez, ayudar a elaborar instrumentos de medida más justos a la hora de evaluar a personas pertenecientes a dichos grupos.

A fin de proponer estrategias para aportar evidencias de validez basadas en esta fuente, es necesario primero entender qué son y en qué consisten los procesos de respuesta. Estos son las fases o “rutas cognitivas” que las personas llevan a cabo cuando procesan los ítems, y elaboran y comunican sus respuestas a los mismos. Este proceso de pregunta-y-respuesta tendría, en concreto, cuatro fases (Tourangeau, 2018):

- ❖ **Comprensión del ítem.** En esta fase, las personas leen el ítem que deben responder y hacen una interpretación del mismo de la cual derivarán un significado concreto. En esta fase, ya que para la comprensión es necesario hacerse una representación cognitiva de las relaciones lógicas entre las palabras, pueden darse innumerables errores de comprensión o idiosincrasias en la comprensión (e.g., que un mismo término implique significados diferentes para grupos de personas distintos, en función de sus experiencias idiosincrásicas).

- ❖ **Recuperación de la información requerida.** Una vez comprendido el ítem, la siguiente fase consiste en recuperar en la memoria la información que se requiere para poder contestar a la pregunta que se formula en el ítem. En esta fase, las personas son susceptibles de cometer errores de memoria, debidos a la posibilidad de que la formulación del ítem no funcione como una buena clave de memoria que active la información precisa, a distorsiones en la memoria, o a que simplemente no se había codificado en su memoria la información que se les requiere.
- ❖ **Juicio y estimación.** La siguiente fase consiste en, con la información que se ha recuperado de la memoria, realizar un juicio o estimación sobre cuál sería la respuesta a la pregunta que se formula en el ítem. Estas respuestas a los ítems estarán además moduladas por juicios, actitudes, u otros sesgos como el *priming* (e.g., preguntas o términos anteriores podrían suscitar valoraciones en un sentido determinado), o el efecto de contraste (ítems anteriores podrían servir de punto de comparación para elaborar la respuesta).
- ❖ **Comunicación de la respuesta.** Por último, las personas comunican la respuesta que han estimado en la fase anterior. Dependiendo de las características de la prueba, la forma requerida para comunicar la respuesta tendrá un formato u otro, el cual las personas deberán comprender y seguir. Esta cuestión puede presentar también problemas, puesto que algunos formatos de respuesta pueden resultar ambiguos o, dependiendo de las idiosincrasias de las personas, pueden interpretarse de formas no esperadas (e.g., para algunas personas, “bastante” puede tener un significado más cercano a “mucho”, mientras que para otras personas puede tener un significado más cercano a “poco”). Además de estas dificultades, sesgos como la deseabilidad social juegan también un papel clave en esta fase, ya que muchas personas pueden tratar de

enmascarar su verdadera respuesta porque consideren que no es socialmente deseable, a fin de salvaguardar su imagen o por temor a ciertos juicios de valor emitidos por parte de observadores externos.

Como explicita Tourangeau (2018), numerosos sesgos y factores pueden manifestarse en cada una de las cuatro fases del proceso de pregunta-y-respuesta, poniendo en peligro la invarianza y la validez de las interpretaciones de las puntuaciones. Otras perspectivas consideran además que la motivación también puede jugar un papel importante a la hora de llevar a cabo un proceso de respuesta más o menos elaborado. Por ejemplo, Krosnick (1991) propone la teoría de la satisfacción o del *satisficing*, en la que defiende que las personas que no están suficientemente motivadas para responder el test o cuestionario puede que recurran a estrategias que les requieran poco esfuerzo mental y contesten de manera “satisfactoria”, es decir, entendiendo las preguntas de manera superficial, omitiendo las fases de recuperación de la información y de juicio y estimación, y dando lo que “parece” una respuesta razonable pero que no refleja ningún proceso psicológico subyacente (e.g., tendencias de respuesta como la aquiescencia o respuestas de una sola palabra en preguntas abiertas).

Numerosas son, pues, las amenazas a la validez de las interpretaciones de las puntuaciones de los tests, y en consecuencia, a la invarianza y la justicia en las mediciones, que tienen que ver con los procesos de respuesta. Existen algunas estrategias para indagar en los procesos de respuesta, que tradicionalmente se han usado en el campo de la Metodología de Encuestas para asegurar la calidad de las respuestas, y que puede resultar útil introducir en el campo de la Psicometría y de la Evaluación Psicológica y Educativa para aportar evidencias de validez basadas en los procesos de respuesta (e.g., Benítez y Padilla, 2014; Benítez et al., 2019; Padilla y Benítez, 2014). Estas estrategias son la entrevista cognitiva (Beatty y Willis, 2007) y la aplicación de las técnicas propias de la entrevista cognitiva en entornos web, el *web probing* (Behr et al., 2017). Ambos son métodos cualitativos que pueden proporcionar

información sobre las interpretaciones que hacen las personas de los ítems, los procedimientos que han llevado a cabo para estimar sus respuestas, las dificultades que han encontrado en el proceso, o incluso las experiencias vitales que han podido modular dichas respuestas (Beatty y Willis, 2007; Behr et al., 2017). A pesar de existir dichos métodos, estos no se han explotado aún lo suficiente para desarrollar protocolos concretos y definidos específicamente para explorar la invarianza métrica y sus causas aportando evidencias de validez basadas en los procesos de respuesta, cuestión a la que se desea responder con la presente tesis doctoral.

2. EL GENDER GAP COMO PROBLEMA DE EQUIVALENCIA Y JUSTICIA EN LAS MEDICIONES

El término *gender gap* o “brecha de género” se acuñó por primera vez en 1980 de mano de la feminista estadounidense Eleanor Smeal, cuando se quejó de las diferencias que había en el voto entre hombres y mujeres y en cómo afectarían esas diferencias a la elección de un nuevo presidente para E.E.U.U. (Carroll, 2018). A partir de ahí, el término se popularizó en diversos contextos (e.g., política, economía, salud, educación...), para designar las diferencias que existen entre mujeres y hombres en un determinado dominio (e.g., Bertocchi y Bozzano, 2020; Merone et al., 2021; Perrin, 2021).

En la cuestión que nos ocupa, el *gender gap* en la medición —las diferencias de género en las puntuaciones en un test, cuestionario o escala— podría reflejar un problema de falta de equivalencia o invarianza métrica en las mediciones, si esas diferencias de género que se encuentran fueran artefactos del proceso de medición y no diferencias “verdaderas” en los constructos objeto de medición. En ese caso, siguiendo el Modelo Ecológico de Respuesta a los Ítems (Zumbo et al., 2015), diversos factores contextuales podrían haber influido de forma

diferencial en los procesos de respuesta de mujeres y hombres, haciendo que sus respuestas varíen en función de elementos irrelevantes para el constructo objeto de medición, y que, en definitiva, sus diferencias en las puntuaciones no estén reflejando diferencias reales en dicho constructo.

En el caso del *gender gap*, algunas variables contextuales que pueden influir diferencialmente en los procesos de respuesta de hombres y mujeres pueden ser los estereotipos y roles de género, que suponen una socialización diferencial de género de las personas desde que estas nacen (Eagly, 1987; Eagly y Wood, 2012). Esta socialización diferencial de género provoca que mujeres y hombres vivan experiencias muy diferentes, e incluso que interactúen con una misma realidad de forma distinta. Por ejemplo, debido a que las mujeres son socializadas en ser amables y sociables (e.g., Diekman y Eagly, 2008), pueden tener manifestaciones muy diferentes de ciertas condiciones como el Trastorno del Espectro Autista (TEA), teniendo más motivación social, más coordinación entre comunicación verbal y no verbal y en definitiva, más enmascaramiento de las manifestaciones “clásicas” del TEA que los hombres (e.g., Hervás, 2022). Esas diferencias en las experiencias y esas formas idiosincrásicas de vivir y entender ciertas realidades pueden moldear de manera distinta los procesos de respuesta a los ítems de los instrumentos de evaluación de unos y otras, generando diferencias en las fases del proceso de pregunta-y-respuesta (Tourangeau, 2018), que darían lugar, a su vez, a diferencias en las puntuaciones que no se deben a diferencias reales en el constructo objeto de medición. A pesar de esto, las diferencias de género en los procesos de respuesta se han estudiado tradicionalmente como diferencias biológicas ligadas al sexo biológico (Zumbo, 2021), lo cual supone una aproximación simplista, tautológica, injusta, determinista y profundamente errada, puesto que la investigación y la experiencia humana han demostrado que la concepción binaria del sexo biológico no tiene sentido para explicar las diferencias en los procesos de respuesta (Zumbo, 2021; Zumbo, en prensa), ya que las causas

de estas diferencias residen, probablemente, en la socialización diferencial de género que establece los estereotipos y roles de género que moldean la conducta (e.g., Diekman y Eagly, 2008; Eagly, 1987; Eagly y Wood, 2012), tal y como se ha argumentado anteriormente. Por tanto, es necesaria una aproximación ecológica que tenga en cuenta de forma rigurosa la influencia de estas variables en los procesos de respuesta, tal y como proponemos en esta tesis doctoral.

Otra variable que puede poner en peligro la validez y la justicia de las interpretaciones comparativas de las puntuaciones de mujeres y hombres podría ser el “sesgo androcéntrico” que existe en la ciencia a la hora de llevar a cabo estudios y de sacar conclusiones a partir de ellos, a la hora de desarrollar procedimientos e intervenciones, y a la hora de desarrollar tests y cuestionarios, fruto de la falta de perspectiva de género en la ciencia y la ceguera de género que existe en la investigación (e.g., Ferrer-Pérez, 2019; García-Calvente et al., 2010; Menéndez-Monte et al., 2022). Este sesgo androcéntrico representa la consideración de que lo masculino es la norma, teniendo sólo en cuenta las experiencias de los hombres e invisibilizando las de las mujeres (e.g., Bonilla-Campos, 2010; García-Calvente et al., 2010), propiciando así una falta de investigación sobre las realidades femeninas (e.g., Menéndez-Monte et al., 2022). Esta ceguera e “invalidación” de las realidades femeninas es un ejemplo del no respeto por las “diversas formas de ser humano” de las que habla Bruno D. Zumbo (2021; 2023; en prensa) cuando se refiere a la “cuarta generación del DIF”.

Por ejemplo, anteriormente se hizo referencia a las manifestaciones “clásicas” de TEA, y precisamente estas son las manifestaciones típicas en hombres, que se han convertido en la norma, pero las realidades femeninas con respecto al TEA son muy diferentes, lo que ha conllevado al infradiagnóstico sistemático de mujeres TEA (Hervás, 2022). En el caso del desarrollo de tests y cuestionarios, este sesgo se traduciría en elaborar los instrumentos desde el punto de vista masculino, considerando únicamente indicadores que reflejen

comportamientos que se asume que, como son normales o anormales en hombres, lo son también en mujeres. Hay algunos indicios de que esto podría estar ocurriendo en el contexto de la evaluación psicológica, como se ha visto con el infradiagnóstico de mujeres TEA (Hervás, 2022), y se sospecha que puede estar ocurriendo también en el diagnóstico de otras condiciones (Garb, 2021), como el Trastorno por Déficit de Atención con Hiperactividad o TDAH (e.g., Hurtado-García, 2019). Debido a que es una inquietud relativamente reciente, que nació en parte de las impresiones de los y las profesionales al llevar a cabo sus evaluaciones, es necesaria más investigación al respecto que indague la naturaleza de estos sesgos —con procedimientos como el que se propone en esta tesis doctoral—, con el fin último de desarrollar instrumentos libres de sesgo que sean justos y válidos, ya que, por el momento, no se cuenta con ellos en la práctica clínica.

Otro ejemplo podría ser el relativo a la interpretación que se ha dado en ocasiones de que las mujeres presentan mayor deseabilidad social que los hombres, derivada de haber encontrado en algunos estudios que las mujeres puntúan más alto en deseabilidad social que sus contrapartes masculinos (e.g., Camerini y Schulz, 2018; Gutiérrez et al., 2016). Sin embargo, uno de los ítems de la escala más utilizada para evaluar la deseabilidad social, la escala de Deseabilidad Social de Marlowe-Crowne (Crowne y Marlowe, 1960; versión adaptada al español por Ferrando y Chico, 2000), dice así: “Siempre soy muy cuidadoso con mi forma de vestir”. Responder afirmativamente a dicho ítem incrementa la puntuación en deseabilidad social, ya que, según la lógica de los ítems de esta escala, el comportamiento que refleja este ítem sería un comportamiento muy poco habitual, pero muy deseable socialmente, por lo que las personas con alta deseabilidad social tenderían a manifestar que lo llevan a cabo. No obstante, al elaborar este ítem no se tuvo en cuenta que para las mujeres, por cuestiones de socialización y de asimilación de estereotipos de género tradicionales que ponen sobre ellas una gran presión estética, ser cuidadosas con la forma de vestir es algo habitual, y que las

mujeres van a tender a responder afirmativamente a dicho ítem, pero no porque sean más “socialmente deseables”. Este tipo de sesgos hacen que las diferencias de puntuaciones que se encuentran no se deban a diferencias reales en el constructo objeto de medición, en este caso la “deseabilidad social” como tendencia de respuesta, y pueden derivar en interpretaciones comparativas de las puntuaciones injustas y no válidas, que guíen un proceso de toma de decisiones sesgado y equivocado (e.g., etiquetado innecesario y estigmatizante de personas, o aplicación de un tratamiento psicológico erróneo).

También existe este sesgo en el campo de la metodología de encuestas. Por ejemplo, se ha visto que encuestas demográficas o sobre el uso del tiempo, al recopilar información sobre la ocupación laboral, toman como referencia las realidades laborales masculinas, invisibilizando las femeninas (Carrasco et al., 2004).

Por todo lo expuesto anteriormente, es preciso como psicómetras, tomar conciencia siguiendo una perspectiva de género en la ciencia para explorar las realidades femeninas y analizar cómo los estereotipos y roles de género influyen en los diferentes constructos objetos de medición. Así, una perspectiva de género permite una comprensión del fenómeno de estudio más exhaustiva, comprehensiva, precisa y acertada, aprehendiendo los fenómenos de estudio en toda su complejidad (e.g., Comisión Europea, 2009). Además, debemos asegurar el abordaje de los intereses de mujeres y hombres, así como de las consecuencias positivas y negativas que tendría para ambos grupos cada desarrollo científico (e.g., Bonilla-Campos, 2010). Estas cuestiones están estrechamente relacionadas con los principios de validez y justicia en las mediciones, tan importantes a la hora de realizar interpretaciones comparativas de las puntuaciones de pruebas en las que existe *gender gap*.

3. EL GENDER GAP EN MEDIDAS EDUCATIVAS Y DEL CLIMA ESCOLAR

3.1. Equivalencia y justicia en medidas educativas

En el caso de los estudios con medidas educativas y del clima escolar, es común hallar ejemplos de un posible *gender gap*. Por ejemplo, en el ámbito educativo, es frecuente encontrar que, a pesar de tener un nivel académico similar, las chicas tienen menos “autoeficacia” que los chicos, sobre todo cuando se trata de estimar sus habilidades en matemáticas, tendiendo a creer que no son buenas en matemáticas con más frecuencia de la que creen eso los chicos sobre sí mismos (e.g., Ayuso et al., 2020; Organización para la Cooperación y el Desarrollo Económico [OCDE], 2019). Esta diferencia de género en el nivel de autoeficacia podría actuar como profecía autocumplida para las chicas, dando lugar a resultados en los que las chicas de verdad tuvieran un peor desempeño que los chicos en las pruebas educativas de matemáticas. Esto, a su vez, podría traducirse en que existan diferencias en la motivación para perseguir en el futuro carreras universitarias del ámbito STEM (*Science, Technology, Engineering and Mathematics*, o Ciencias, Tecnología, Ingeniería y Matemáticas), generando y perpetuando así un *gender gap* en el número de chicas y chicos que se decantan por dichas carreras (e.g., Sakellariou y Fang, 2021). De hecho, en el año 2022 en España, solo el 5.5% de las mujeres ocupadas trabajaba en carreras profesionales STEM, mientras que en hombres este porcentaje subía hasta el 13% (Cobrerros et al., 2024).

En cuanto a las medidas sobre el clima escolar, por ejemplo, se han encontrado ciertos resultados en el Programa para la Evaluación Internacional de los Estudiantes o *Programme for International Student Assessment* [PISA] 2018 que podrían parecer contradictorios: los chicos tienden a reportar que tienen un mayor sentimiento de pertenencia a su centro escolar

que las chicas, mientras que también admiten haber sufrido más episodios de acoso escolar o *bullying* que ellas (OCDE, 2019; OCDE, 2023). Este tipo de resultados podrían levantar sospechas sobre la existencia de algún fenómeno que influye de algún modo en la forma de contestar de los chicos a los ítems relativos al clima escolar, que en chicas podría no estar operando, o de algún tipo de sesgo en el proceso de medición, como el anteriormente mencionado sesgo androcéntrico. De ser así, la detección de casos de acoso escolar o *bullying* podría verse perjudicada y cabría plantearse algunas preguntas: ¿podría haber casos que se etiquetaran como *bullying* pero que los chicos no los consideren como tal por haber interiorizado y normalizado la violencia a consecuencia de sus roles de género?, o ¿podría haber experiencias de *bullying* en chicas que no se etiqueten como tal, y, por tanto, no se detecten, como causa de que ellas sufren formas de violencia más sutiles que se alejan de la norma masculina?

Estas situaciones ilustran cómo una interpretación de las puntuaciones de un instrumento de evaluación que no haya tenido en cuenta las diversas variables que pueden influir en las respuestas según el Modelo Ecológico de Respuesta a los Ítems (Zumbo et al., 2015), puede no ser válida ni justa, y conllevar una serie de consecuencias que desemboquen en un acceso desigual a ciertos contextos educativos o en problemas a la hora de detectar casos en los que existe acoso escolar o *bullying*.

3.2. El caso del *bullying*: contextualización

El acoso escolar o *bullying* es un comportamiento agresivo y no deseado que implica un desequilibrio de poder entre la persona perpetradora y la víctima (Olweus, 1978, 2013), y entraña una serie de consecuencias negativas para la salud mental y el bienestar de las y los adolescentes que lo sufren. En concreto, la victimización de *bullying* se relaciona con problemas tales como el absentismo escolar, el mal desempeño académico, la depresión, la baja

autoestima y auto-eficacia, la ansiedad, la insatisfacción vital, la soledad, o el consumo de sustancias y otras conductas de riesgo (e.g., Lee, 2021; OCDE, 2019; Swearer y Hymel, 2015; Organización para las Naciones Unidas para la Educación, la Ciencia y la Cultura o *United Nations Educational, Scientific and Cultural Organization* [UNESCO], 2019), y tiene consecuencias tan fatales como el suicidio (e.g., Díaz-Aguado et al., 2023; Holt et al., 2015; Kim y Leventhal, 2008). Es por ello por lo que la presente tesis doctoral centra su parte más sustantiva en la problemática del acoso escolar o *bullying*, ya que no detectar correctamente y de forma acertada los casos de *bullying* puede acarrear consecuencias muy severas y negativas (e.g., pasar por alto casos de *bullying* que requieran una intervención y no proporcionarla, perpetuando así el sufrimiento de la víctima).

Además, la prevalencia de este fenómeno varía ampliamente en función del estudio, desde menos del 10% hasta más del 65% (Chung y Lee, 2020). Esta es una cuestión que resulta llamativa y que podría estar poniendo de manifiesto la dificultad que entraña capturar dicho fenómeno. Así, se ha visto que en ocasiones la prevalencia varía, al menos en parte, en función de la redacción o presentación de las preguntas o ítems. Por ejemplo, mientras que sólo el 6.2% de estudiantes se identificaron como víctimas de *bullying*, el 19.3% admitieron haber sufrido algún tipo de comportamiento violento o abusivo por parte de sus iguales (Díaz-Aguado et al., 2023). A la luz de datos como estos, parece que las y los estudiantes son reticentes a reconocerse como víctimas, mientras que los comportamientos abusivos entre iguales son, desgraciadamente, bastante comunes. En ese sentido, detectar correctamente casos de *bullying* con el objetivo de aplicar programas de intervención apropiados y ajustados, plantea un reto en términos de validez de las interpretaciones de las puntuaciones de las pruebas, que requerirá para su abordaje explorar las variables que, según el Modelo Ecológico de Respuesta a los Ítems (Zumbo et al., 2015), influyen en los procesos de respuesta de los y las estudiantes, y analizar la invarianza o equivalencia métrica de las respuestas de chicas y chicos a las pruebas.

3.3. El caso del *bullying*: diferencias entre chicos y chicas

Las experiencias de *bullying* de chicas y chicos son muy diferentes, y tienen que ver con las dos formas principales que se han encontrado en la literatura (e.g., Björkqvist et al., 1992; Gladden et al., 2014; Lee et al., 2022; Olweus, 1994; Rivers y Smith, 1994; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022), en las que las conductas de acoso escolar o *bullying* pueden manifestarse. Estas son:

- ❖ **Bullying directo.** Esta categoría engloba, a su vez, las formas de *bullying* físico y *bullying* verbal, ya que ambas incluyen agresiones que causan un daño inmediato y son de naturaleza abierta, directa y evidente. Se consideran agresiones de tipo físico golpear, empujar, patear, o robar o dañar las pertenencias de la víctima. Se consideran agresiones de tipo verbal insultar, amenazar, o poner sobrenombres ridículos.
- ❖ **Bullying indirecto.** También puede llamársele *bullying* relacional, e incluye agresiones de naturaleza indirecta y sutil, destinadas a dañar las relaciones sociales de la víctima y/o su reputación. Este tipo de agresiones no requieren que la persona perpetradora y la víctima interactúen física y directamente. Algunas de estas agresiones son esparcir rumores falsos sobre la víctima, o la exclusión social.

Teniendo esa tipología de agresiones en cuenta, la literatura previa ha encontrado que, por un lado, los chicos son más propensos a perpetrar y ser víctimas de las tipologías de *bullying* directo, tanto físico como verbal; y por otro lado, las chicas son más propensas a perpetrar y ser víctimas de *bullying* indirecto o relacional (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Rivers y Smith, 1994; Rosen y Nofziger, 2019; Shu et al., 2023). Además, también se ha encontrado que los chicos tienden a sufrir más *bullying* que las chicas, independientemente de la tipología (e.g., Carrera-Fernández et al., 2013; Hanzi & Akça, 2020; Lee et al., 2022; Wang et al., 2010; Zhao et al., 2022).

Estas diferencias, como se apuntó en el apartado anterior sobre el *gender gap*, podrían estar causadas, al menos en parte, por la socialización diferencial de género que se les proporciona a chicas y chicos, en la que se les educa en seguir los roles y estereotipos de género tradicionales (e.g., Carrera-Fernández et al., 2013; Rosen y Nofzinger, 2019). Por un lado, en el caso de las chicas, se las socializa para ser sensibles, amables y serviciales (e.g., Diekman y Eagly, 2008), lo que podría concordar con la preferencia a perpetrar modalidades de *bullying* más sutiles e indirectas. Por otro lado, a los chicos se les socializa en los principios de la masculinidad tradicional o hegemónica, como son el poder, control y estatus, la fortaleza y dureza, o la agresividad (e.g., Diekman y Eagly, 2008; Vescio et al., 2010). Esto también casa con la preferencia que parecen tener los chicos a perpetrar modalidades de *bullying* más directas y con un componente más físico.

Tal y como se expuso en apartados anteriores, esta diferencia en la socialización y experiencias de chicos y chicas con respecto a la violencia y, en concreto, al acoso escolar o *bullying*, pueden afectar de forma diferencial a la forma de entender y experimentar el fenómeno, lo cual, a su vez, podría suscitar procesos de respuesta diferentes a la hora de contestar a los ítems que evalúen victimización de *bullying*. Así, las diferencias que pudieran encontrarse al hacer interpretaciones comparativas de las puntuaciones de ambos grupos (e.g., la interpretación de que los chicos sufren más *bullying* que las chicas), podrían deberse tanto a diferencias reales en el constructo objeto de medición como a artefactos del proceso de medición, surgiendo así la necesidad de aportar evidencias de validez basadas en los procesos de respuesta para averiguar cuál sería el caso, establecer si se cumple o no el principio de invarianza o equivalencia métrica, descartar cualquier sesgo androcéntrico que pudiera haber en la evaluación y asegurar una interpretación de las puntuaciones válida y justa.

En ese sentido, explorar los procesos de respuesta de chicas y chicos cuando responden a ítems de victimización de *bullying* podría arrojar luz sobre hasta qué punto entienden o

interpretan los enunciados y/o términos de los ítems de forma idiosincrásica, hasta qué punto calibran de diversa forma las opciones de respuesta, o hasta qué punto el entendimiento diferencial que tienen sobre el *bullying* les lleva a estimar sus respuestas de una forma u otra. Así, se podría esbozar una descripción precisa de la problemática, ajustada debidamente a cada grupo sociodemográfico, y que permita realizar evaluaciones y consecuentes intervenciones acertadas, personalizadas si se requiere, y sobre todo, justas.

CAPÍTULO II: OBJETIVOS

OBJETIVOS

En el capítulo anterior se presentó cómo la falta de invarianza podría poner en peligro la validez y la justicia de las interpretaciones comparativas de las puntuaciones de un test, y en concreto, cuando se comparan las experiencias de victimización de *bullying* en chicas y chicos. Para abordar esta problemática, la presente tesis doctoral plantea una serie de objetivos de investigación, detallados a continuación.

El **objetivo general** de la tesis doctoral ha sido desarrollar y examinar la aplicación de una estrategia “mixta” de validación, en la que se integran métodos cuantitativos y cualitativos, para obtener evidencias de validez basadas en los procesos de respuesta con las que apoyar las interpretaciones comparativas de las puntuaciones de chicos y chicas en ítems sobre victimización de *bullying*, así como asegurar la justicia en la medición. El desarrollo y examen de esta estrategia mixta de validación pretende que se pueda aplicar en cualquier contexto de evaluación en el que se deseen realizar interpretaciones comparativas de puntuaciones de diferentes grupos de personas, ya sean grupos sociodemográficos, lingüísticos, o culturales, a pesar de que la tesis doctoral se ha centrado en un contexto concreto: la evaluación de victimización de *bullying* en chicas y chicos adolescentes.

Para alcanzar el objetivo general se plantearon cinco objetivos específicos, centrados en el contexto de evaluación de la victimización de *bullying*. Son los siguientes:

- ❖ **Objetivo específico 1:** Evaluar el estado de conocimientos sobre los resultados, las condiciones de aplicación, y las ventajas e inconvenientes de los métodos cuantitativos

y cualitativos para el análisis de la equivalencia, y de las fuentes de inequivalencia. Para ello, se realizó una revisión sistemática mixta de la literatura científica sobre los métodos cuantitativos y cualitativos utilizados para abordar el estudio de la equivalencia o invarianza métrica en medidas educativas (Estudio I), comenzando así a acotar el contexto específico en el que se han desarrollado los estudios de la tesis doctoral.

- ❖ **Objetivo específico 2:** Examinar, por medio de estudios empíricos, las condiciones de aplicación de métodos cuantitativos para estudiar la equivalencia o invarianza métrica y sus resultados, identificando una posible ejecución condicionada diferencial entre diferentes grupos de personas evaluadas, y de apuntar a fuentes potenciales de inequivalencia. Para ello, se realizó un análisis de clases latentes multigrupo a las respuestas de más de 11,000 adolescentes españoles a la escala de “Experiencias de *bullying*” usada en el Cuestionario del Estudiante del PISA 2018 para, por un lado, identificar diferentes grupos de estudiantes a través de sus patrones de respuesta, y por otro lado, analizar si la invarianza métrica entre chicas y chicos se mantenía (Estudio II).
- ❖ **Objetivo específico 3:** Examinar, a través de estudios empíricos, las condiciones de aplicación de los métodos cualitativos y sus potenciales aportaciones de evidencias de validez sobre los procesos de respuesta, asociadas a los resultados de los métodos cuantitativos. Para ello, se realizaron entrevistas cognitivas (Estudio III) a estudiantes de Educación Secundaria Obligatoria (E.S.O.), para obtener evidencias de validez basadas en los procesos de respuesta de chicos y chicas a la escala de “Experiencias de *bullying*” usada en el Cuestionario del Estudiante del PISA 2018 que pudieran apoyar

las interpretaciones comparativas de las puntuaciones de chicas y chicos en dicha escala, o apuntar a fuentes de inequivalencia.

- ❖ **Objetivo específico 4:** Aportar explicaciones a las diferencias en percepción del clima escolar y las experiencias de *bullying* entre chicos y chicas, mediante la aplicación de las estrategias metodológicas mixtas desarrolladas en los objetivos específicos anteriores. Este objetivo se alcanza de forma transversal, mediante los Estudios II y III.
- ❖ **Objetivo específico 5:** Proponer diferentes estrategias de integración de los métodos cuantitativos y cualitativos para la validación de las interpretaciones comparativas en estudios con diferentes grupos sociodemográficos, lingüísticos y/o culturales. Este objetivo se alcanza de forma transversal, mediante los Estudios II y III.

CAPÍTULO III: METODOLOGÍA

1. METODOLOGÍA MIXTA PARA EL ESTUDIO DE LOS PROCESOS DE RESPUESTA

1.1. Introducción a la metodología mixta.

En la presente tesis doctoral se desarrolla una estrategia “mixta” de validación centrada en aportar evidencias de validez de los procesos de respuesta, y en asegurar la justicia de las interpretaciones de las mediciones aportadas por los instrumentos de evaluación. En concreto, y con el fin último de que esta estrategia de validación se pueda generalizar y aplicar a cualquier contexto de evaluación, se propone y muestra el uso de la metodología mixta para indagar en los procesos de respuesta de chicas y chicos al contestar ítems de victimización de *bullying*, con el objetivo de detectar una posible falta de invarianza y sus causas, y así asegurar la validez de la interpretación comparativa de sus puntuaciones.

Propuesta como alternativa a la “exclusividad metodológica” —uso de metodologías cuantitativas o cualitativas como aproximaciones incompatibles—, la metodología mixta es aquella que combina e integra, en todas las fases del proceso de investigación, métodos cuantitativos y cualitativos para lograr un conocimiento más “compreensivo” de fenómenos complejos y obtener hallazgos más sofisticados (Cresswell y Plano-Clark, 2018; Johnson et al., 2007; Tashakkori y Teddlie, 1998). Esta metodología hace posible tener en cuenta múltiples puntos de vista y perspectivas que no se podrían obtener utilizando solamente métodos cuantitativos o métodos cualitativos (Johnson et al., 2007).

El estudio de los sesgos que pueden estar presentes en las mediciones a menudo ha resultado difícil, ya que tradicionalmente se han usado casi exclusivamente métodos cuantitativos para explorarlos, y estos métodos tienen limitaciones a la hora de buscar o determinar las causas de dichos sesgos (e.g., Benítez et al., 2019). Así, el uso de una metodología mixta que integre métodos cuantitativos con cualitativos podría solventar este problema al aportar explicaciones más comprehensivas de estos fenómenos, superando las limitaciones de los métodos cuantitativos y cualitativos por sí solos (Benítez et al., 2019; Creswell y Plano-Clark, 2018).

Esa es la razón por la que cada vez más investigaciones en Psicometría optan por esta aproximación a la hora de explorar, por ejemplo, los sesgos en la medición indagando en los procesos de respuesta de las personas que contestan a los tests o cuestionarios. Por ejemplo, Koskey y Stewart (2013) utilizaron métodos mixtos para explorar los significados que las personas atribuían a los valores numéricos de las opciones de respuesta, concluyendo que ese significado e interpretación que las personas hacían de las opciones de respuesta no era universal ni compartido entre personas de diferentes culturas, dando lugar a diferencias en las respuestas al cuestionario. Otros estudios han utilizado los métodos mixtos para aportar evidencias de validez basadas en los procesos de respuesta (e.g., Benítez et al., 2019; Castillo-Díaz y Padilla, 2013). Por ejemplo, Benítez et al. (2019) llevaron a cabo un estudio en el que compararon las respuestas de personas españolas y holandesas en una escala sobre calidad de vida, analizando, por una parte, la presencia de funcionamiento diferencial de los ítems (DIF) a través de estrategias estadísticas cuantitativas, y por otra, analizando, a través de entrevistas cognitivas, los procesos de respuesta de dichas personas al contestar los ítems, integrando finalmente ambos tipos de resultados para aportar una visión más completa del fenómeno del DIF, de acuerdo con las últimas concepciones que se tienen del mismo (Zumbo, 2007; Zumbo, 2021; Zumbo, en prensa).

Aplicar métodos mixtos para aportar evidencias de validez basadas en los procesos de respuesta parece, no solamente factible, sino deseable, puesto que enriquece el tipo de información que se obtiene acerca de posibles “amenazas” a la comparabilidad de las interpretaciones que pueden estar comprometiendo tanto la validez como la justicia de las mediciones (e.g., Padilla y Benítez, 2014).

1.2. Diseños mixtos y las diferentes formas de integración.

La integración en la metodología mixta es la piedra angular sobre la que se cimenta esta metodología. Consiste en la combinación intencional y completa de los métodos cuantitativos y cualitativos en todas las fases de la investigación, para conseguir un entendimiento “cohesionado” sobre el fenómeno de estudio, que no se podría obtener a partir de la aplicación de los métodos cuantitativos y cualitativos por separado (e.g. Bazeley, 2024; Benítez y Padilla, 2014; Moseholm y Fetters, 2017).

Fetters et al. (2013) distinguen tres niveles a los que se puede realizar la integración, siendo lo ideal llevar a cabo la integración en todos esos niveles y no solamente en un único momento temporal (Bazeley, 2024). Son los siguientes (Fetters et al., 2013):

❖ **Nivel del diseño.** La integración se realizaría en el momento de diseñar y conceptualizar el estudio. Los tres diseños más utilizados que varían en función del planteamiento del estudio, son los siguientes:

- **Diseño secuencial exploratorio.** En este diseño, primero se recoge y se analiza la información cualitativa, y esa información sirve para guiar la recogida de datos cuantitativos posterior. Por ejemplo, Serrano-Montilla et al. (2025) extrajeron información cualitativa procedente de grupos focales sobre las actitudes de las fuerzas policiales ante las actuaciones en casos de violencia de

género, y después utilizaron esa información para realizar un análisis de redes semánticas para conectar la información cualitativa con una estructura de datos cuantitativa, y así poder ver si hay diferencias entre las actitudes de policías especializados en violencia de género y de policías no especializados.

- **Diseño secuencial explicativo.** En este diseño, primero se recoge y se analiza la información cuantitativa, y esa información guía la recogida de datos cualitativa posterior. Por ejemplo, Benítez et al. (2019) en su estudio mencionado anteriormente, analizaron en primer lugar la presencia de DIF de forma cuantitativa y esos resultados guiaron la elaboración de las entrevistas cognitivas posteriores, encaminadas a indagar en los procesos de respuesta para explicar las causas del DIF encontrado.
- **Diseño convergente.** En este diseño, también llamado diseño concurrente, la información cualitativa y cuantitativa se recoge y se analiza en el mismo momento. Esto puede hacerse, o bien llevando a cabo un procedimiento interactivo mediante el cual los hallazgos cuantitativos y cualitativos se influyen constantemente entre sí, o bien llevando a cabo los procedimientos cuantitativos y cualitativos al mismo tiempo en paralelo para integrarlos al final. Un ejemplo de la primera aproximación sería el estudio de Crabtree et al. (2005), en el cual, con el objetivo de comprobar cómo estaban funcionando los servicios de medicina preventiva, utilizaron los hallazgos cuantitativos para ir informando la recogida y el análisis de datos cualitativos, los cuales se realizaron varias veces, de forma iterativa. Un ejemplo de la segunda aproximación sería el estudio de Saint-Arnault y Feters (2011), en el que llevaron a cabo, al mismo tiempo, una recogida de datos cuantitativa y una

recogida de datos cualitativa para después analizar los datos conjuntamente, con el objetivo de averiguar cómo afectan las experiencias de enfermedad, las interpretaciones culturales y los factores socioculturales a la conducta de búsqueda de ayuda de mujeres japonesas.

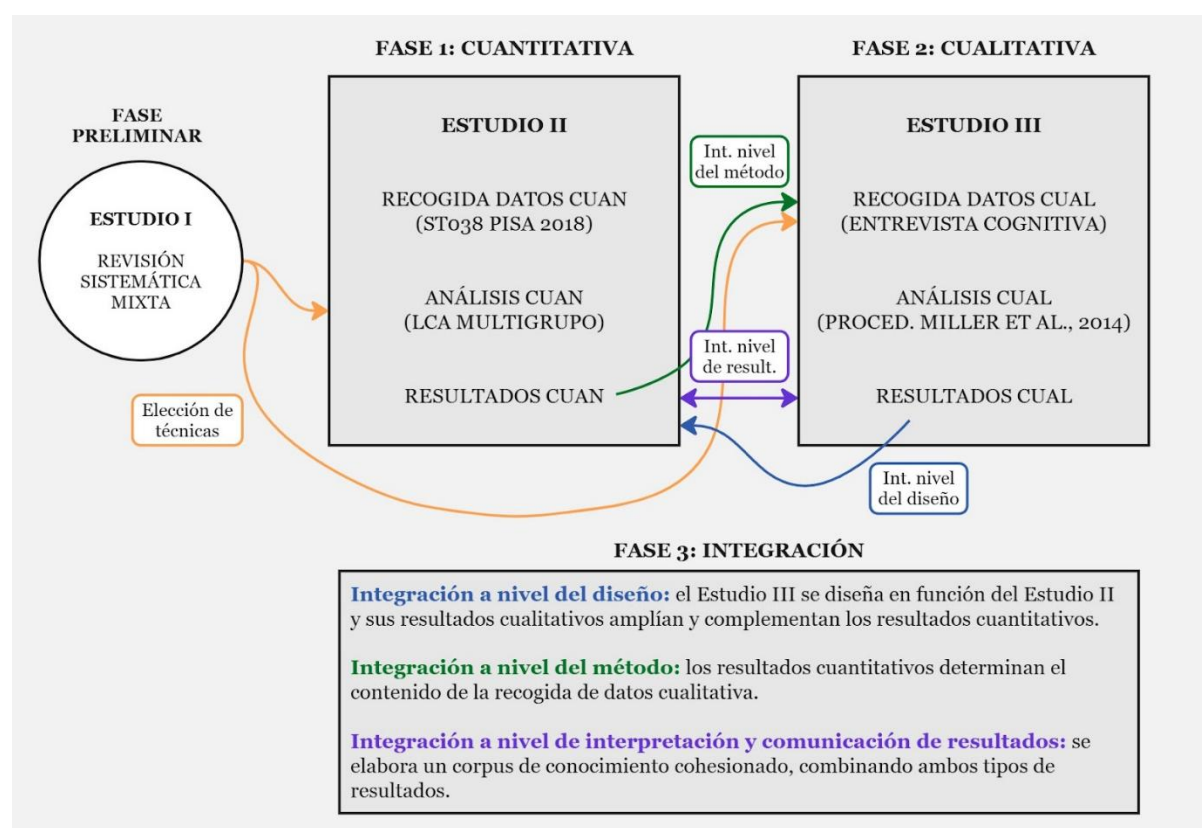
- ❖ **Nivel del método.** La integración se realizaría en el momento de recoger y analizar los datos. Puede hacerse a través de diferentes estrategias: 1) *connecting*, que consiste en que una tipología de datos se une con la otra a través del muestreo; 2) *building*, que es cuando los resultados de una recogida de datos informan la otra recogida de datos; 3) *merging*, que sería unir los resultados de ambas tipologías de datos para hacer los análisis de forma conjunta; y 4) *embedding*, cuando la información cualitativa y cuantitativa se unen en múltiples puntos, tanto de la recogida de datos como del análisis.
- ❖ **Nivel de la interpretación y la comunicación de resultados.** La integración se realizaría en el momento de interpretar los resultados y transmitirlos de forma conjunta y cohesionada. Esto puede llevarse a cabo a través de diferentes procedimientos: 1) integrar a través de la “narrativa”, describiendo los resultados de ambos métodos, bien de forma conjunta, o bien de forma secuencial; 2) integrar a través de la “transformación de los datos”, convirtiendo los datos cualitativos en cuantitativos o viceversa; y 3) integrar a través de *joint displays*, que son maneras visuales de plasmar, por ejemplo en tablas, ambos tipos de información.

En definitiva, es necesario alcanzar una integración de la información cuantitativa y cualitativa completa, que combine las dos tipologías de información en diversos puntos a lo largo de todas las fases del estudio, y no solo en un único momento, asumiendo que existe una interdependencia entre lo cualitativo y lo cuantitativo (Bazeley, 2024).

2. UN DISEÑO MIXTO SECUENCIAL EXPLICATIVO COMO ESTRATEGIA DE VALIDACIÓN

En la presente tesis doctoral, se propone una estrategia de validación que aporte evidencias basadas en los procesos de respuesta a través de un diseño mixto secuencial explicativo, el cual se ha ido llevando a cabo a lo largo de todos los estudios que componen la tesis doctoral. Dicha estrategia mixta de validación (Figura 1) se detalla a continuación.

Figura 1. Estrategia Mixta de Validación: Esquema.



2.1. Fase preliminar: revisión sistemática mixta de la evidencia.

Como paso previo a la aplicación del diseño mixto secuencial explicativo que conforma la estrategia mixta de validación, se realizó una revisión sistemática de la literatura científica para recopilar la evidencia sobre las técnicas cuantitativas y cualitativas que se utilizan para examinar la invarianza en la medida y el DIF en el contexto de la evaluación educativa, así como sus causas. Los resultados de esta revisión sirvieron para elaborar la estrategia mixta de validación y escoger los métodos cuantitativos y cualitativos a aplicar en cada fase de la misma. Esta revisión sistemática fue mixta —también llamada “de estudios mixtos”—, para permitir la inclusión en la misma revisión de estudios cuantitativos, cualitativos y mixtos, y conseguir así un corpus de conocimiento más completo y comprehensivo (e.g., Lizarondo et al., 2020; Pearson et al., 2015; Pluye y Hong, 2014; Stern et al., 2020). La metodología de la revisión sistemática, por tanto, se ajustó para adecuarla a una metodología mixta, en todas sus fases, aún siempre siguiendo las recomendaciones del *Preferred Reporting Items for Systematic Review and Meta-analysis* [PRISMA] (Page et al., 2021). La información detallada sobre este estudio se desarrolla en el Capítulo IV de la tesis doctoral, aunque a continuación se señalan las adecuaciones metodológicas que se realizaron para adaptar la revisión sistemática a una metodología mixta.

En primer lugar, se utilizó la aproximación SPIDER (*Sample, Phenomenon of Interest, Design, Evaluation, Research type*; Cooke et al., 2012), para desarrollar las preguntas de investigación y la estrategia de búsqueda, debido a que resulta una alternativa a la estrategia PICO (*Population, Intervention, Comparison, Outcome*) tradicional (e.g., Thomas et al., 2024) que se ajusta más a nuestros objetivos de investigación. Se optó por esta aproximación puesto que, tanto la estrategia PICO como otras alternativas a ella como la SPICE (*Setting, Perspective, Intervention, Comparison, Evaluation*; Booth, 2006), guían la planificación de la revisión sistemática hacia una comparación de la eficacia de diferentes intervenciones, lo cual

supone una visión muy ligada al contexto biomédico, que no se correspondía con los objetivos de nuestra revisión. Además, estrategias como la PICO, aunque son muy efectivas para elaborar ecuaciones de búsqueda que encuentren estudios cuantitativos, no lo son tanto para encontrar estudios cualitativos o mixtos (Cooke et al., 2012). La aproximación SPIDER, por su parte, no habla de intervención sino de “fenómeno de interés”, permitiendo así preguntas de investigación más amplias sobre temáticas diversas, y permite también especificar el tipo de diseño empleado en los estudios que se buscan, así como el tipo de investigación, pudiéndose explicitar en la ecuación de búsqueda si se desean encontrar estudios cualitativos, cuantitativos y/o mixtos (Cooke et al., 2012).

El proceso de cribado de artículos se realizó siguiendo el procedimiento habitual. Sin embargo, en cuanto a la extracción de datos de los estudios incluidos, se siguió un modelo convergente basado en los resultados o “*result-based convergent design*” (Hong et al., 2017), en el que se llevaron a cabo de manera paralela y en un mismo momento temporal, por un lado, una extracción de datos cuantitativos y, por otro lado, una extracción de datos cualitativos. Posteriormente, estos resultados cuantitativos y cualitativos fueron integrados a través de una síntesis narrativa (Popay et al., 2006). Este procedimiento se describe con detalle en el Capítulo IV de la tesis doctoral.

Además, para llevar a cabo la evaluación del riesgo de sesgo y la calidad de los estudios incluidos en la revisión, se optó por utilizar una herramienta diseñada para evaluar tanto estudios cuantitativos, como cualitativos y mixtos. Esta herramienta fue la *Quality Assessment Tool* (QATSDD), desarrollada por Sirriyeh et al. (2012), la cual evalúa la calidad metodológica y del reporte de los estudios, comprobando que los estudios tengan la suficiente robustez metodológica para ser creíbles, y que sean transparentes, claros, completos y precisos a la hora de aportar información sobre la forma de llevar a cabo el estudio. La QATSDD cuenta con ítems generales y también con ítems específicos para evaluar solamente estudios cuantitativos

y solamente estudios cualitativos, permitiendo un sistema de puntuaciones específico para cada tipo de estudio (Sirriyeh et al., 2012).

Por último, para evaluar el grado de confianza en la evidencia acumulativa de la revisión sistemática también se optó por usar una herramienta específicamente diseñada para evaluar hallazgos cualitativos, la GRADE-CERQual (Lewin et al., 2015), ya que la totalidad de los hallazgos finales de nuestra revisión sistemática fueron cualitativos, al haber “*cualificado*” los datos cuantitativos mediante el procedimiento de síntesis narrativa anteriormente mencionado (Popay et al., 2006). Esta herramienta es una adaptación de la tradicional GRADE (*Grading of Recommendations Assessment, Development and Evaluation*; Schünemann et al., 2013), la cual evalúa hallazgos cuantitativos. Con la GRADE-CERQual se evalúan las limitaciones metodológicas, la relevancia, la coherencia y la adecuación o suficiencia de cada evidencia cualitativa, dando una puntuación del grado de confianza de cada hallazgo de la revisión sistemática.

2.2. Fase 1: obtención y análisis de la información cuantitativa.

En primer lugar, se analizó de forma cuantitativa la invarianza métrica entre chicas y chicos españoles en sus respuestas a la escala de experiencias de *bullying* (ST038) del Cuestionario del Estudiante de la edición del año 2018 del Programa para la Evaluación Internacional de Estudiantes (*Programme for International Student Assessment*), también conocido como PISA. El PISA es un estudio de carácter internacional, promovido por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), que se lleva a cabo cada tres años desde el año 2000. Su objetivo principal es discernir y explicar lo que chicos y chicas de 15 años saben y pueden hacer en términos de rendimiento escolar, aunque también se exploran otras variables de tipo “no-cognitivo”, en las que se engloban dimensiones como la calidad de vida o el clima escolar. Así, el PISA consta de una prueba principal de carácter

“cognitivo”, en la que se evalúan las competencias académicas de los y las estudiantes en dominios como matemáticas, lengua o ciencias, y del Cuestionario del Estudiante, que es un compendio de escalas que evalúan estas variables “no-cognitivas”, con el objetivo de discernir cómo impactan en el rendimiento académico del estudiantado (OCDE, 2018; OCDE, 2019a; OCDE, 2019b).

Se optó por analizar datos ya recabados al tratarse de un estudio internacional de un amplio recorrido e impacto, puesto que en él participan 79 sistemas educativos, tanto de países pertenecientes a la OCDE como de países asociados, y las muestras de cada país son amplias y representativas, obtenidas a través de un procedimiento de muestreo robusto y riguroso, al tratarse de un estudio sólido y consolidado. Esto nos aseguró que los resultados obtenidos podrían generalizarse a la población del estudiantado español de 15 años. Además, el PISA ejerce una gran influencia a la hora de que los países tomen decisiones sobre las políticas educativas que dirigen sus currículos educativos y sus centros escolares, y la escala ST038 sobre las experiencias de *bullying* podría utilizarse como medida de *screening*, tanto para evaluar la prevalencia de la problemática, como para empezar a detectar casos en las aulas. Se utilizó la edición del 2018 porque la siguiente, que se publicó en 2022, se llevó a cabo durante los años de la pandemia del COVID-19 y la adaptación a la “nueva normalidad” posterior, siendo este un período en el que el estudiantado tuvo que alternar la educación presencial con la telemática, alterando posiblemente los resultados de variables tan dependientes de la fisicidad de las interacciones, como son las del clima escolar.

Así, para analizar la invarianza métrica de forma cuantitativa, se utilizó un análisis de clases latentes (LCA), que además también permitió obtener diferentes perfiles de respuestas al cuestionario, y se utilizaron estos resultados para informar y guiar el desarrollo de la fase cualitativa. La información detallada sobre este estudio se desarrolla en el Capítulo V de la tesis doctoral.

2.3. Fase 2: obtención y análisis de la información cualitativa.

Tal y como se ha mencionado, los resultados obtenidos en la primera fase cuantitativa sirvieron para informar y guiar el desarrollo de la segunda fase cualitativa. En concreto, a partir de dichos resultados, se realizó un protocolo de entrevista cognitiva que indagó en los procesos de respuesta del estudiantado y exploró las posibles causas de la falta de invarianza encontrada en la primera fase, completando de esa manera el conocimiento sobre la problemática desde otra perspectiva, y añadiendo el contexto necesario para eventualmente interpretar de forma válida y justa las puntuaciones de chicas y chicos en la escala de experiencias de *bullying*.

Al elaborar las pruebas de indagación o *probes* para la entrevista cognitiva, se trató de priorizar aquellas que “provocaran” el discurso espontáneo de chicos y chicas al responder, formulando así preguntas abiertas y generales que apelaran a los procesos de respuesta de las y los participantes, pero que no guiaran de ninguna manera sus respuestas (e.g., “*Cuéntame, ¿qué has pensado mientras respondías a las preguntas? ¿Qué has recordado? ¿Qué situaciones se te han venido a la mente?*”).

Así, se llevaron a cabo una serie de entrevistas cognitivas a estudiantes que primero contestaron a la escala ST038 sobre experiencias de *bullying* del Cuestionario del Estudiante del PISA 2018, en un entorno que simuló la recogida de datos del estudio original, esto es, mediante un cuestionario por ordenador en el centro escolar. La información de dichas entrevistas se analizó siguiendo el procedimiento propuesto por Miller et al. (2014), en el que a través de un razonamiento inductivo, se realiza un análisis temático de categorías emergentes cuya información se va sintetizando en diferentes pasos establecidos. De esta manera, se sintetizó en primer lugar el texto de las transcripciones, se realizó un análisis temático de categorías emergentes, y a partir de este análisis se elaboraron esquemas temáticos de los que

después se extrajeron conclusiones. La información detallada sobre este estudio se desarrolla en el Capítulo VI de la tesis doctoral.

2.4. Fase 3: integración de los resultados cuantitativos y cualitativos.

Por último, se integran los hallazgos cuantitativos y cualitativos en un corpus de conocimiento comprensivo. Dicha integración se ha realizado a lo largo de todo el procedimiento, en varios momentos diferentes, y a distintos niveles (Fetters et al., 2013).

En primer lugar, a nivel del diseño (Fetters et al., 2013), se ha utilizado, como ya se ha apuntado, un diseño mixto secuencial explicativo en el que los hallazgos cuantitativos se han usado para elaborar y desarrollar el propio diseño de la fase cualitativa de la estrategia mixta. Así, la conceptualización del Estudio III estuvo fundamentada en los hallazgos cuantitativos del Estudio II, con el objetivo de explicarlos mejor y complementarlos.

En segundo lugar, a nivel del método (Fetters et al., 2013), se ha utilizado la estrategia de *building* (Fetters et al., 2013) para que los resultados de la recogida de datos cuantitativa determinaran el contenido y la información a recolectar de la recogida de datos cualitativa. Es decir, el protocolo de entrevista cognitiva de la segunda fase se elaboró a partir de los hallazgos de la primera fase, incluyendo información específica determinada por dichos hallazgos.

Por último, debido a que las fases 1 y 2 se llevaron a cabo en estudios que se publicaron de forma independiente, la integración a nivel de la interpretación y comunicación de resultados (Fetters et al., 2013) se realizará en la propia discusión de la tesis doctoral, en el Capítulo VII (Discusión general). Esta integración se llevará a cabo a través de la narrativa, siguiendo una *weaving approach* (Fetters et al., 2013), en la que se describen de forma conjunta los resultados cuantitativos y cualitativos, estructurando la información que se presenta en función de los temas o conceptos que se han estudiado. Así, con la información cualitativa de las entrevistas

cognitivas, se irá tratando de dar sentido completo a los perfiles de respuestas encontrados con el LCA y se tratarán de identificar las causas de la falta de invarianza encontrada, discutiendo las diferentes implicaciones que acarreen estos hallazgos.

CAPÍTULOS EMPÍRICOS

CAPÍTULO IV: ESTUDIO I

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}, and Isabel Benítez^{1,2}

¹Department of Methodology for Behavioral Sciences, University of Granada, Granada, Spain

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

This chapter is published as:

Navarro-González, M. C., Padilla, J. L., & Benítez, I. (2024). Analyzing measurement invariance for studying the gender gap in educational testing: A mixed studies systematic review. *European Journal of Psychological Assessment*, 40(5), 1-15.
<https://doi.org/10.1027/1015-5759/a000820>

Abstract

Gender differential item functioning (DIF) is essential to address if fair and valid test score interpretations are desired. Despite inequivalence having been addressed previously, there are no 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 the causes of the gender gap in educational assessments with adolescents. We used the Quality Assessment Tool (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 Confirmatory Factor Analysis (CFA) and Mantel-Haenszel were the most used strategies to test measurement invariance/detect DIF. Certain methods, such as latent class analysis (LCA), mixture item response theory model (MMixIRTM), or Simultaneous Item Bias Test (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

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 valid 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 valid 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 a 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 valid evidence for comparative group 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:

- ❖ **Research Question 1 (RQ1):** 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?
- ❖ **Research Question 2 (RQ2):** 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?
- ❖ **Research Question 3 (RQ3):** How have these quantitative techniques and qualitative methods been applied (e.g., research designs)?
- ❖ **Research Question 4 (RQ4):** 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 on 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 11, 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., 2023). 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 $\kappa = .55$, 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 $\kappa = .66$, 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 (QSR 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., 2023). Emergent categories were added when necessary, allowing the process to be flexible and dynamic. Some categories were removed or slightly modified due to the non-homogeneity of studies and/or irrelevance to our research questions. NVivo (QSR 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., 2023). 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 $\kappa = .81$ 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., 2023).

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 ($\kappa = 71.5\%$) 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 $\geq 71\%$), medium risk of bias (56% $\leq$ 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 $\kappa = .61$, 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. The 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 into 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 the 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 for the included articles is provided in Supplementary Material VIII (Navarro-González et al., 2023).

Study Characteristics

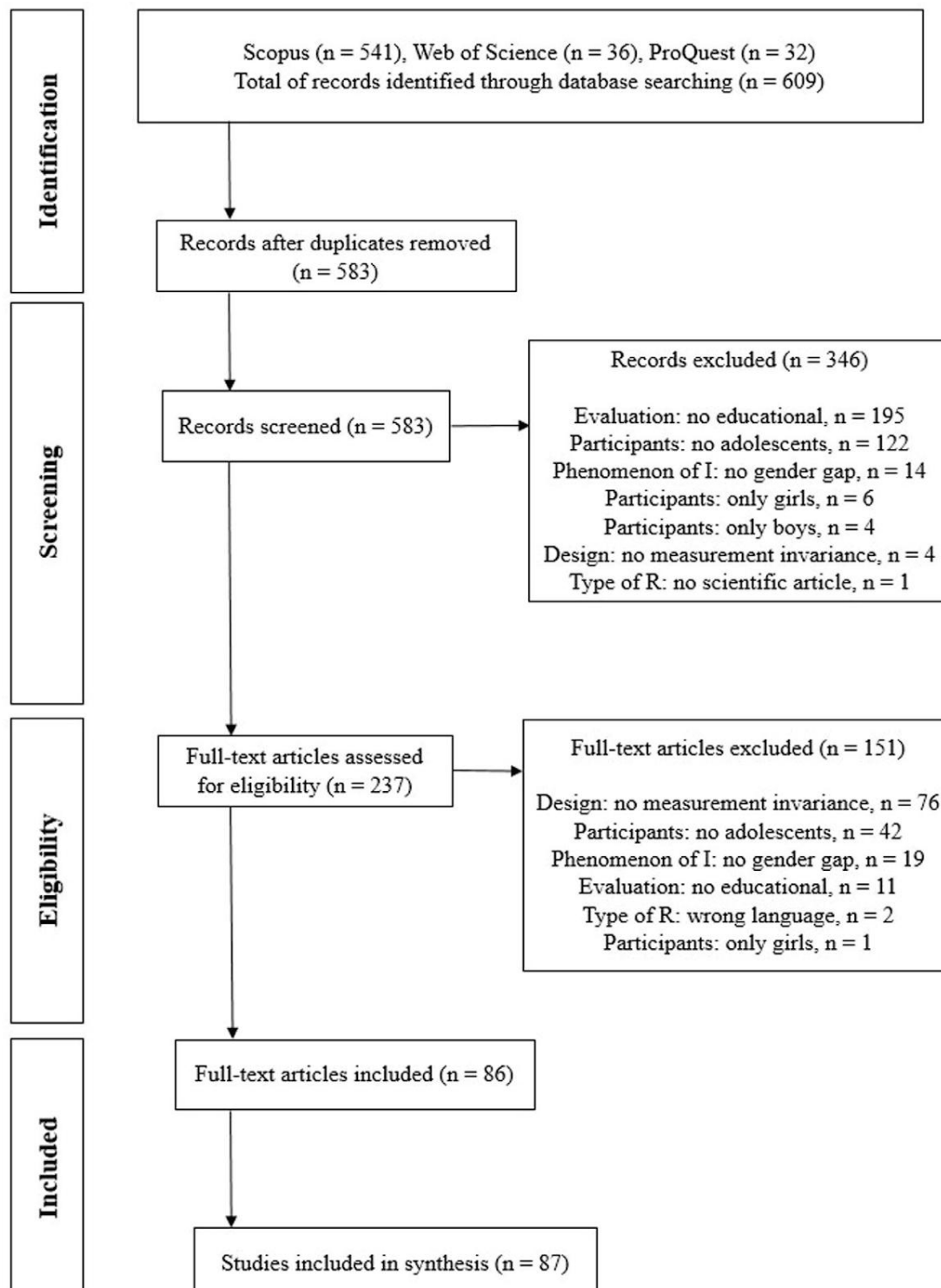
Here, we will discuss the bibliometric indicators and methodological aspects of the studies. In Supplementary Material IX (Navarro-González et al., 2023), two summarizing tables (Tables IX.1 and IX.2) displaying the main characteristics of each study are presented.

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 ($n = 20$). 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 1. Selection of sources of evidence: Flow diagram.



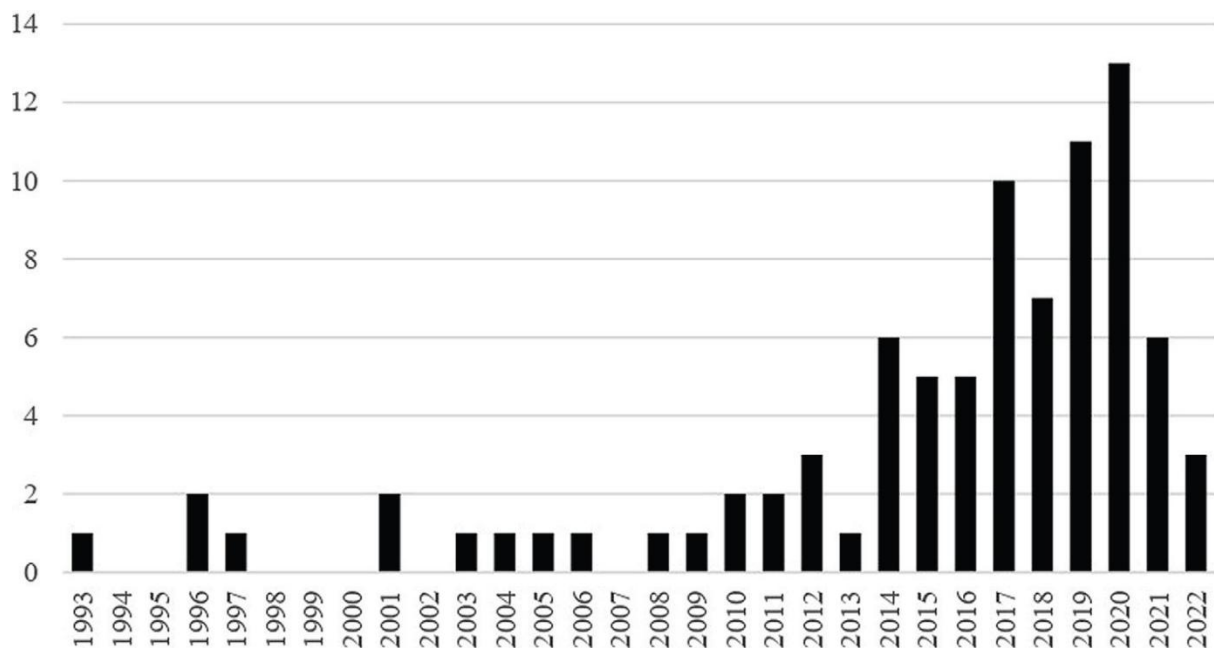
Methodological Aspects

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 the 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 the 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 1,000 participants. When studies had also disaggregated data from participants who 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 the 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., 2023) displays all of the target variables used.

Figure 2. Publication trend from 1993 to 2022.



Risk of Bias in Studies

In Supplementary Material XI (Navarro-González et al., 2023), 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., 2023). Overall, the majority of studies had a low risk of bias (62 out of 87; 71.26%), 25 had a medium risk of bias, and none of them had a 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. The most popular analytic strategy is measurement invariance testing, through Multigroup Confirmatory Factor Analysis (CFA) and/or Multigroup Structural Equation Modeling (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 item response theory (IRT) Rasch analysis, eight

that performed logistic regressions (LR/OLS), and six that performed a Simultaneous Item Bias Test (SIBTEST) or poly-SIBTEST. Overall, 68 (78.16%) studies used only parametric strategies, whereas 10 (11.49%) used only nonparametric ones (nine used both).

Table 1. Frequency of usage of each DIF Detection/Measurement Invariance Method.

DIF detection/measurement invariance method	<i>f</i>
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.

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., 2023). 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., 2023), 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.

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.

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	<i>f</i>
Qualitative	
Content analysis	9
Didactical interpretation of results	1
Interview	1
Focus groups	1
Expert appraisal	1
Delphi technique	1
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.

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, and so forth (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., 2023). 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, latent class analysis (LCA), Multiple-Indicator Multiple-Cause (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 (1,000 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., 2023), 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 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., 2023), with the aim to complement the rest of the 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., 2023).

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 (non-significant) 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 Büyükoztürk (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., 2023).

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 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., 2023). 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., 2023) 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 of literature along the history of the 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 fourth 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. The 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., 2022). 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 reducing construct-irrelevant variance when assessing girls and boys, because the information obtained through this kind of study 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 –for example, 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 socio-cultural level. The socio-cultural 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 themselves 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 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 valid 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.

Declarations

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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 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. Template data collection forms and Supplementary Material are available here: <https://osf.io/rvqk7/>

Open Materials: I confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology: <https://osf.io/rvqk7/>

Pre-registration of Studies and Analysis Plans: This study was pre-registered on April 9, 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): <https://osf.io/xzgdq>

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

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).

References

- American Psychological Association. (2022). *APA thesaurus of psychological index terms*. APA. <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. <https://doi.org/10.1109/TE.2020.3004075>
- Benítez, I., Van de Vijver, F., & Padilla, J. L. (2022). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research*, 51(1), 237–270. <https://doi.org/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, Article 100819. <https://doi.org/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), Article 201. <https://doi.org/10.3390/educsci11050201>
- Chang, C. C. (2019). Development of ocean literacy inventory for 16- to 18-year-old students. *SAGE Open*, 9(2), 1-15. <https://doi.org/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.
<https://doi.org/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* (pp. 53-68). Springer. <https://doi.org/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.
<https://doi.org/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.
<https://doi.org/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.
<https://doi.org/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.
<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.
<https://doi.org/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.
<https://doi.org/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.
<https://doi.org/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.
<https://doi.org/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* (3rd ed., pp. 36–44). Elsevier. <https://doi.org/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.
<https://doi.org/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. <https://doi.org/10.1186/s13643-017-0454-2>
- IBM Corp. (2013). *IBM SPSS statistics for windows* (Version 22) [Software]. <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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1177/0734282910391623>
- 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. <https://doi.org/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. <https://doi.org/10.1080/02671522.2019.1633557>
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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*, 19(1), 1–27. <https://doi.org/10.1080/15434303.2021.1963253>
- Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Eğitim ve Bilim – Education and Science*, 39(172), 20–32. <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. <https://doi.org/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. https://doi.org/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. <https://doi.org/10.1002/pits.22330>
- Navarro-González, M. C., Padilla, J. L., & Benítez, I. (2023). *Analyzing measurement invariance for studying the gender gap in educational testing* [Supplementary Materials, Data, Output]. <https://osf.io/rvqk7/>
- OECD. (2019). *PISA 2018 results (volume III): What school life for students' lives*. OECD Publishing. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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 T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: multinational, multiregional, and multicultural contexts (3MC)* (pp. 45-64). Wiley. <https://doi.org/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. <https://doi.org/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. *British Medical Journal*, 372(71), 1–9. <https://doi.org/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*. <https://database.inahta.org/article/2684>
- QSR International. (2022). *NVivo* (Version 13) [Software]. <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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1002/nur.20176>

- 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. <https://doi.org/10.1080/10904018.2015.1065747>
- Shamseer, L., Moher, D., Clarke, M., Ghera, 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. *British Medical Journal*, 349, 1–25. <https://doi.org/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(4), 746–752. <https://doi.org/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(3), 246–280. <https://doi.org/10.1080/08957347.2012.687650>
- The Jamovi Project. (2022). *jamovi* (Version 2.2.2) [Software]. <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, Article 622. <https://doi.org/10.3389/fpsyg.2020.00622>
- UNESCO. (2022). *UNESCO Thesaurus*. <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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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 P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing*. Wiley. <https://doi.org/10.1002/9781119263685>

CAPÍTULO V: ESTUDIO II

TESTING COMPARABILITY OF GENDER BULLYING VICTIMIZATION ASSESSMENTS THROUGH A LATENT CLASS APPROACH

M. Carmen Navarro-González^{1,2}, José-Luis Padilla^{1,2}, and Álvaro Postigo³

¹Department of Methodology for Behavioral Sciences, University of Granada, Granada, Spain

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

³Faculty of Psychology, Department of Psychology, University of Oviedo, Spain

This chapter is published as:

Navarro-González, M. C., Padilla, J. L., & Postigo, A. (2025). Testing comparability of gender bullying victimization assessments through a latent class approach. *Journal of Interpersonal Violence*, 00(0) [online first], 1-25.
<https://doi.org/10.1177/08862605251325926>

Abstract

Bullying is a social and health problem that requires appropriate interventions based on valid and fair evaluations of bullying experiences. The validity of interpretations of bullying victimization scores can be compromised by measurement artifacts or biases that may arise during the assessment process. Boys' and girls' bullying experiences could lead to differences in their response processes when they answer the bullying scale items and compromise validity if such differences come from measurement artifacts. The study is intended to illustrate how to obtain validity evidence of response processes for the *Students' Experience of Being Bullied Scale* from the Programme for International Student Assessment (PISA) 2018 Student Questionnaire and test measurement invariance across gender by latent class analysis (LCA). The sample was taken from the PISA 2018 study and consisted of 11,599 Spanish high school students (50.3% female, 49.7% male). Response patterns were examined through LCA. Four profiles were found: (a) *Not Bullied*, (b) *Bullied (All Types)*, (c) *Relational Bullying*, and (d) *Potential Friendly Teasing*. Measurement invariance across gender was analyzed by a multigroup LCA. LCA results do not guarantee equivalence of measurements. Class prevalence per group and a multinomial logistic regression were calculated to further examine gender differences across latent classes. Boys were more likely to belong to the *Bullied (All Types)* class and the *Potential Friendly Teasing* class, whereas girls were more likely to belong to the *Relational Bullying* class. These findings illustrate the possibilities LCA can offer to provide validity evidence of response processes and suggest that different bullying experiences of girls and boys could compromise the fairness and validity of comparative interpretations of test scores, which is an important issue to address to develop bias-free classroom interventions.

Keywords: bullying, gender differences, adolescents, measurement invariance, latent class analysis, validity.

Introduction

According to Olweus (1978, 2013), bullying is an unwanted and aggressive behavior that involves a real or perceived imbalance of power between perpetrator and victim. Bullying can take various forms: direct bullying (physical and verbal bullying; aggressions that cause immediate harm) or indirect bullying (relational bullying; aggressions that damage relationships/reputation; e.g., Björkqvist et al., 1992; Gladden et al., 2014; Olweus, 1994; Rivers & Smith, 1994). The definition of bullying has remained the same throughout the years (e.g., Gladden et al., 2014; Lee, 2021), but the variety of ways in which those “aggressive behaviors” can emerge have expanded recently through new forms of bullying that take place in social networks (cyberbullying), like revenge porn and cyberstalking (e.g., Lee et al., 2017). Bullying is a relevant social problem that can compromise school climate and endanger students’ well-being. The prevalence of bullying among children and adolescents ranges widely across studies (from less than 10% to over 65%; Chung & Lee, 2020), with a mean prevalence of 35% in a metaanalysis by Modecki et al. (2014). Based on data from the United Nations Educational, Scientific, and Cultural Organization (UNESCO, 2019), about 32% of the global student population reported having been bullied at school in the past month. Moreover, according to the Programme for International Student Assessment (PISA) 2018 (Organization for Economic Cooperation and Development [OECD], 2019b), a triennial educational international testing project involving 15-year-olds in 79 countries, between 9% and 35% of students suffered some kind of bullying at least a few times a month. The OECD mean was 23%, having increased from 18.7% since the PISA 2015 (OECD, 2017b), indicating

that the number of bullying victims seems to have increased in OECD countries between 2015 and 2018.

In Spain, the bullying data seem to follow the same pattern. The prevalence of bullying victimization revealed in the PISA 2018 was 17% (OECD, 2019a), higher than the prevalence estimated in 2015, which was 14% (OECD, 2017a). Additionally, in Spain, the National Observatory on School Climate found a prevalence of 6.2% of bullying victimization (Observatorio Estatal de la Convivencia Escolar, 2010), which is the exact same percentage of students who identified as bullying victims more recently shared by the Preventive Psychology Unit of the Complutense University of Madrid and ColaCao Foundation (Díaz-Aguado et al., 2023). The latter study (Díaz-Aguado et al., 2023) also reported that 19.2% of students admitted to having suffered some kind of abusive behavior from their peers, such as name calling and teasing. Moreover, the Ayuda a Niños y Adolescentes en Riesgo (ANAR) and Mutua Madrileña Foundations showed that 11.8% of Spanish students reported having detected at least one case of bullying in their class (Fundación ANAR & Fundación Mutua Madrileña, 2023). The diversity of study results suggests that even though students do not tend to identify as bullying victims, abusive behaviors from peers in school are quite common. These variations in prevalence estimated by surveys could occur, at least partially, due to differences in the survey questions' wording and formats. For instance, when students are directly asked about being bullying victims, prevalence is lower than when they are asked indirectly, or the word "bullying" is not used.

From a measurement perspective, fair and valid assessments of bullying victimization are crucial to providing a precise description of the problem and to properly detecting cases that require intervention. According to the *Standards of Educational and Psychological Testing*, published by the American Educational Research Association (AERA), the American Psychological Association, and the National Council of Measurement in Education (NMCE,

2014, p. 50), an assessment is fair when it reflects the same construct for all test takers without advantaging or disadvantaging some individuals because of characteristics irrelevant to the intended construct, and when tests or questionnaire scores have the same meaning or interpretation for all individuals. Only valid interpretations of test scores entail fair comparisons between groups of individuals in levels of bullying victimization.

Having that in mind, studying girls' and boys' assessments is particularly relevant to ensure fair comparisons. Boys tend to suffer more direct aggressions (physical and verbal aggressions), whereas girls tend to suffer more indirect or relational aggressions (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Rivers & Smith, 1994; Rosen & Nofziger, 2019; Shu et al., 2023). Additionally, some studies indicate that boys are more likely to be situated within a bullying victim-perpetrator overlap, whereby a perpetrator is also a victim (Falla et al., 2020). The variations in boys' and girls' experiences with bullying could entail differences in their response processes when answering bullying victimization items. Boys and girls could interpret the items' stems and/or response options differently due to their different understanding of the bullying phenomenon. When such differences in response processes do not represent true differences in the intended construct (bullying victimization) but come from measurement artifacts, validity and fairness are compromised by what the psychometric literature calls "sources of construct-irrelevant variance" (AERA et al., 2014). A first step to examine whether the same construct is being assessed for girls and boys, and to know the extent to which boys' and girls' scores are comparable, is examining measurement invariance. Therefore, this study is intended to illustrate how to obtain validity evidence of response processes of Spanish boys and girls in a bullying victimization assessment using latent class analysis (LCA).

Gender and Typologies of Bullying Victimization

According to the literature (e.g., Björkqvist et al., 1992; Gladden et al., 2014; Olweus, 1994; Rivers & Smith, 1994), three types of bullying can be distinguished: physical, verbal, and social or relational. Similar patterns have been found in more recent studies that included LCA (e.g., Lee et al., 2022; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022). Physical bullying includes overt behaviors of physical abuse (e.g., hitting, pushing, kicking, robbing, breaking the victim's belongings). Verbal bullying includes overt behaviors of verbal abuse (e.g., name calling, threatening, teasing, insulting). These two types of bullying are considered direct bullying because all the related behaviors are overt and open and are intended to cause immediate harm (Wójcik & Flak, 2021). The third type of bullying, considered indirect bullying because it does not require an encounter between the bully and victim, is relational bullying. This type of bullying comprises behaviors intended to damage the victim's relationships, reputation, or acceptance (e.g., social exclusion, spreading false gossip about the victim).

The literature (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Rivers & Smith, 1994; Rosen & Nofziger, 2019; Shu et al., 2023) also shows that boys are more likely to perpetrate and suffer direct bullying (physical and verbal bullying), whereas girls are more likely to perpetrate and suffer indirect bullying (relational bullying). In general, boys seem to suffer more bullying than girls, regardless of the type (e.g., Carrera-Fernández et al., 2013; Hanzi & Akça, 2020; Lee et al., 2022; Wang et al., 2010; Zhao et al., 2022). In fact, the PISA 2018 (OECD, 2019b) revealed that in most countries (including Spain), boys were more likely to report having been bullied than girls. Some studies (e.g., Carrera-Fernández et al., 2013; Rosen & Nofzinger, 2019) have suggested that these findings could be a consequence of the different socialization process mediated by gender stereotypes and roles for girls and boys,

linking bullying perpetration to characteristics of hegemonic masculinity, such as heteronormativity, physical dominance, acceptance of violence, control, and instrumentality.

The Consequences of Bullying Victimization

The consequences of bullying victimization can be extreme. Problems such as trauma, depression, anxiety, poor psychosocial and educational functioning (partially caused by increased absenteeism), substance misuse, risky sexual behaviors, and low levels of self-esteem, resilience, and self-efficacy have been associated with bullying victimization (e.g., Lee, 2021; Swearer & Hymel, 2015; UNESCO, 2019). For example, the PISA 2018 results (OECD, 2019b) show that students who reported being bullied had lower scores on reading tests and reported being lonelier at school, sad, scared, and less satisfied with their lives. Suicidal ideations and suicide rates are also higher among bullying victims (e.g., Holt et al., 2015; Kim & Leventhal, 2008), being the ultimate most severe consequence of bullying. In Spain, 20.4% of students who identified as bullying victims admitted having attempted suicide at least once (Díaz-Aguado et al., 2023). Moreover, also in Spain, according to the ANAR Foundation, 70% of students with suicidal ideations reported having been bullied at school, and 906 children and adolescents attempted to commit suicide in 2022, which is the largest number of attempts in the last 10 years (Fundación ANAR, 2022). Exploring the nature of bullying is of vital importance to prevent those consequences.

Threats to Fairness, Validity, and Measurement Invariance

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), validity evidence based on response processes is one of the five sources of validity evidence that can ensure valid and fair interpretations of test and questionnaire scores. This source of validity evidence addresses “the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers” (AERA et al., 2014, p. 15).

Response processes refer to the routes or phases respondents go through to come up with and communicate the answers to the test and scale items in an assessment context. Tourangeau (2018) outlined the four stages of that route: (a) comprehension of item stem, (b) retrieval of information requested, (c) judgment and estimation, and (d) reporting the answer. Several group differences in the response processes can occur in all four stages, which can compromise the validity of comparative test scores interpretations and the fairness of the assessments if such differences are due, for instance, to measurement artifacts. If the response processes are different between several groups of respondents due to measurement artifacts, it is likely that the same construct is not being measured across those groups and that the meaning of the groups' scores is not the same. Validation studies focused on response processes are needed to examine such potential sources of bias (Padilla & Benítez, 2014). In fact, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) proposed that “questioning test takers from various groups making up the intended test-taking population about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct” (p. 15).

Boys' and girls' diverse experiences in life and school with bullying victimization could somehow affect their unique understanding of bullying. These different understandings of bullying could elicit different cognitive representations of the phenomenon for girls and boys, which may lead to different response processes when they answer items intended to assess their bullying victimization. These differences could be the result of measurement artifacts or biases (He & van de Vijver, 2012) that could compromise the validity of comparative interpretations of girls' and boys' bullying victimization scores, leading to mistaken and unfair comparisons. Therefore, when biases are present, the equivalence—the level of comparability of scores across groups—is threatened (He & van de Vijver, 2012), which in turn compromises the assessment's fairness (AERA et al., 2014). Contradictory results are found in the PISA 2018

(OECD, 2019b), which could hint at the presence of some kind of bias: Boys reported having been bullied more than girls but also reported having a stronger sense of belonging to their school, contradicting previous studies that showed that bullied students felt more lonely and less attached to school (e.g., OECD, 2019b; UNESCO, 2019). Therefore, it is of vital importance to distinguish bias from true score differences to precisely and fairly assess bullying victimization in girls and boys and eventually develop effective, efficient intervention and prevention programs.

Examining measurement invariance can be the first step to achieving that. In the assessment of measurement invariance, equivalence—or lack of bias—is also being tested (e.g., He & van de Vijver, 2012). Several analytic strategies are often applied to test measurement invariance, such as confirmatory factor analysis (CFA; e.g., Brown, 2015), differential item functioning analysis (e.g., Zumbo, 2007), and LCA (e.g., Kankaraš et al., 2018). Recent approaches have highlighted the importance of considering contextual factors to explain the lack of measurement invariance, such as via the Ecological Model of Item Response (Zumbo et al., 2015). Following these approaches, it becomes necessary to search for possible contextual explanations or sources of the lack of measurement invariance. Therefore, understanding gender roles/stereotypes as distal contextual factors that can influence response processes is key in addressing measurement invariance of bullying victimization assessments.

A Person-Centered Approach to Understanding Response Patterns

LCA is a statistical technique that follows a person-centered approach and is based on the assumption that a sample contains several subsamples—or classes—with various scores and response patterns (e.g., Lanza et al., 2013; Weller et al., 2020). In several studies, researchers have resorted to LCA to uncover subgroups of bullying involvement and experiences. For example, some studies (e.g., Chung & Lee, 2020; Gao et al., 2023; Goldweber

et al., 2013) have revealed subgroups among victims and perpetrators that tend to follow the patterns Nansel et al. (2001) proposed: *victim*, *bully*, *bully-victim*, and *uninvolved*. Other studies (e.g., Lee et al., 2022; Nylund et al., 2007; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022) have focused only on uncovering the patterns of bullying victimization. These studies have shown between three and five classes of bullying victimization, always including *uninvolved* and *all types of victimization* classes. Some of them have also revealed classes that represented the bullying types described previously, such as *verbal/relational victims* and *traditional/direct victims*.

LCA is a useful tool to properly identify subgroups because it relies on response patterns to identify latent classes, assuming that an underlying latent variable determines the individuals' class membership (Nylund et al., 2007). In addition, it allows for the inclusion of predictor and outcome variables in the model, making it possible to associate latent class membership with relevant variables and examine, for example, whether class membership varies depending on sociodemographic characteristics (e.g., gender). Moving forward, LCA can also be used to assess measurement invariance (Kankaraš et al., 2018; Rudnev, 2018), comparing latent structures across groups to examine how similar or different they are. In other words, LCA can also be used to examine whether latent classes have the same meaning across groups, which could mean that response processes were similar among groups and that the same construct has been fairly measured in all groups of comparison. Therefore, using LCA to reject the possibility of measurement bias existing and to check whether bullying victimization measures have been producing fair and valid comparative interpretations of boys' and girls' scores can be a promising contribution to precisely understanding and addressing bullying victimization.

The Current Study

This study is intended to illustrate how to obtain validity evidence of response processes of boys' and girls' bullying victimization test scores to reject possible biases that could undermine the validity and fairness of comparative interpretations through LCA and testing measurement invariance across gender. Specifically, this study represents an attempt to disentangle Spanish adolescents' response processes to the *Students' Experience of Being Bullied Scale* used in the PISA 2018 Student Questionnaire by identifying groups of students that exhibit different response patterns and to determine whether measurement invariance between girls and boys is reached. We opted for the Spanish version of the PISA 2018 *Students' Experience of Being Bullied Scale* for several reasons. First, the PISA data sets are obtained by a rigorous sampling procedure that allows results to be generalized to national adolescent populations. Second, the bullying experience scale can be seen as a screener instrument developed following accepted psychometric best practices. The analytic approach will allow us to check whether the different response patterns adolescents adopt when responding to the items are related to various types of bullying and whether those patterns are the same among boys and girls. Differences in bullying experiences between girls and boys can provoke differences in their response processes. Testing measurement invariance is a wide consensus psychometric approach to check if those differences in response processes compromise the validity of group comparisons. If measurement invariance is maintained, the response processes of boys and girls do not undermine validity, despite being different.

Two main research questions guide the study:

1. Could differences in bullying victimization experiences among boys and girls cause differences in their response processes to bullying victimization items?

2. Is measurement invariance between boys' and girls' measures of bullying victimization reached?

Methods

Participants

The database for this study was taken from the PISA 2018 Spanish subdataset, available at <https://www.oecd.org/pisa/data/2018database/>. We carried out a data cleaning process, in which a median imputation of missing values was performed when participants answered all items but one of the *Students' Experience of Being Bullied Scale* from the PISA 2018 Student Questionnaire, and participants with two or more missing values (any two or more items unanswered) in that same scale were removed. After that cleaning process, responses from 11,599 Spanish high school students (50.3% female, 49.7% male) from 930 schools were analyzed. All participants completed the questionnaire in Spanish. The participants were between 15 and 16 years old ($M = 15.84$ years, $SD = 0.28$). The majority (99.3%) attended compulsory secondary education (known in Spain as Educación Secundaria Obligatoria) at the moment of the PISA 2018 administration.

Regarding the PISA 2018 study sampling, students for the whole PISA 2018 study were selected from all participant countries but one (Russia carried out a three-stage design) by a two-stage stratified sample design. During the first stage, schools were selected from a national list of PISA-eligible schools by using a systematic probability proportional to size sampling, based on the estimated number of eligible students (15-year-old students attending educational institutions in Grade 7 or higher). Other variables, such as location and funding, were also taken into account to select the schools. During the second stage, each sampled school developed a list of eligible students, and around 42 students were selected with equal probability in each

sampled school. Those composed the final PISA 2018 sample (OECD, 2018). More details about the technical aspects of the PISA 2018 sampling can be found in the PISA 2018 Technical Report (OECD, 2018).

Procedure

The 2018 PISA cycle was a computer-based administration in most countries (including Spain). The computer questionnaires were presented adopting a paging design, mimicking the kind of interactions students would have with a paper-based assessment (e.g., students could move from page to page; OECD, 2018). The assessment was conducted between March and December 2018. Students responded at their schools. They had 2 hours to complete the cognitive section and about 35 minutes (about an hour in Spain) to complete the background questionnaires (which included the Student Questionnaire). Detailed information is available in the PISA 2018 Technical Report (OECD, 2018).

Measures

The instrument under examination was the scale ST038, intended to measure the students' experience of being bullied at school. The scale comprises six items, with four response categories (1 = *never or almost never*, 2 = *a few times a year*, 3 = *a few times a month*, 4 = *once a week or more*). The first item refers to being excluded from peers ("*Other students left me out of things on purpose*"), the second item describes being laughed at ("*Other students made fun of me*"), the third item refers to being threatened ("*I was threatened by other students*"), the fourth item refers to being robbed ("*Other students took away or destroyed things that belonged to me*"), the fifth item refers to being physically assaulted ("*I got hit or pushed around by other students*"), and the sixth item describes having gossip spread about oneself ("*Other students spread nasty rumors about me*").

Cronbach's alpha for the analyzed Spanish sample was .83, and McDonald's omega was .84, both indicating good internal consistency. The unidimensionality assumption was tested through a cross-validation analysis by splitting the sample into two random subsamples: a categorical exploratory factor analysis (EFA) was conducted with the first subsample, and a categorical CFA was conducted with the second subsample. Due to the nonnormality of the data distributions, an unweighted least squares (ULS) estimator was used (Ferrando et al., 2022). In the EFA, factors were extracted by parallel analysis (Brown, 2015; Ferrando et al., 2022). Because large sample sizes affected the Chi-square statistics, other model fit statistics were calculated. For the EFA, a GFI of .99 and an RMSR of .04 were obtained, which represented a good/reasonable fit (Lloret-Segura et al., 2014). For the CFA, a GFI of .99, a CFI of .99, a TLI of .99, and an RMSEA of .06 were obtained, which also represented a good/reasonable fit (Hu & Bentler, 1999; Lloret-Segura et al., 2014). Therefore, the dataset was considered unidimensional.

Due to the low frequency of response categories for Options 2 (*a few times a year*), 3 (*a few times a month*), and 4 (*once a week or more*), we collapsed those three response categories, as some studies have done (e.g., DiStefano et al., 2020).

Data Analysis

We examined measurement invariance between boys and girls by conducting an LCA in several steps. In the first step, various latent class models were explored to determine which model best represented the data to identify latent classes based on the bullying indicators. This step was carried out for the total sample and with the girls' and boys' samples separately. Models with one to five latent classes were tested, using the maximum likelihood estimation method (Muthén et al., 2015).

To decide on the number of latent classes, multiple indices of parsimony were considered: Bayesian information criterion (BIC), sample-adjusted BIC (SABIC) to adjust for the effect of large sample sizes, and Akaike information criterion (AIC). Smaller BIC, SABIC, and AIC values indicate superior model parsimony. The Lo-Mendell-Rubin log-likelihood ratio test (LMR) and the bootstrapped likelihood ratio test (BLRT) were also used. The LMR and BLRT compare the improvement in fit between two nested class models (e.g., comparing k and $k-1$ class models). Significant LMR and BLRT imply that adding a class improves model fit. As for classification diagnostics, entropy was considered, which has a value between 0 and 1. A higher entropy value indicates better separation between latent classes (Lanza & Cooper, 2016; Nylund-Gibson & Choi, 2018), with a recommended value greater than .70. The size of the smallest class was also considered to select the model that had the greatest parsimony and whose classes each represented more than 5% of the total sample. In addition, we considered substantive theoretical criteria based on the literature to finally select the optimal model. Once the optimal model was selected, each latent class was identified based on the probability of its members having experienced each bullying indicator.

In the second step of the analysis, a multigroup LCA was conducted to assess measurement invariance between boys and girls. Two models were tested, and their model fits were compared. In the heterogeneous model, no restrictions were set and classes were allowed to vary across groups, whereas conditional response probabilities were fixed to be equal across groups in the homogeneous model. If the homogeneous model fits better, measurement invariance is met, meaning that latent classes have the same size across groups and share the same meaning between girls and boys. If the heterogeneous model fits better, latent classes are not distributed equally between boys and girls. These results could mean that classes have different meanings for girls and boys, which would endanger fair comparative interpretations of scores.

Multiple indices of fit were considered: log-likelihood, AIC, BIC, SABIC, and the likelihood-ratio Chi-square (L^2). A relative model fit between the two models was calculated. To examine deeper potential sources of measurement invariance and obtain a more comprehensive understanding, the class prevalence for girls and boys was obtained, and gender's influence on latent class membership was examined by conducting a multinomial logistic regression analysis with gender as a predictor. Thus, gender was added as a covariate to the LCA model selected in the first step. The latent class that represented fewer experiences of bullying victimization was selected as the reference group in the model. All analyses were conducted in Mplus (Muthén & Muthén, 1998–2017). The database and scripts used for all analyses described in this section are available here: https://osf.io/uqje6/?view_only=2c78aa31295a4429a8033dccaace62c2.

Results

Optimal Latent Class Solution and Profiles

In the LCA, models with two to five latent classes were estimated for the total sample and for boys and girls separately (Table 1). The best-fitting models were the solutions with four latent classes. Based on the best-fit model and substantive criteria based on the literature, the four-class model was chosen as the final solution for the total sample, for girls, and for boys.

As Figure 1 shows, individuals belonging to Latent Class 1 (10.34%, $n = 1,199$) were characterized by high probabilities of experiencing indirect or relational bullying acts (such as exclusion from peers, being laughed at, and having gossip spread about them). As a result, this class was labeled *Relational Bullying*. Individuals in Latent Class 2 (6.29%, $n = 729$) had high probabilities of suffering any type of bullying. Consequently, this class was labeled *Bullied (All Types)*. Individuals assigned to Latent Class 3 had medium probabilities of suffering

bullying, specifically of being laughed at, with the peculiarity of having a very low probability of being excluded by peers.

This class was therefore labeled *Potential Friendly Teasing* (10.76%, $n = 1,248$), because it could capture the reality of suffering some common forms of aggression, such as being laughed at, within a group of friends. Finally, individuals belonging to Latent Class 4 (72.62%, $n = 8,423$) showed a very low probability of suffering any kind of bullying. This class was labeled *Not Bullied* (Figure 1).

Figure 1. Model of four latent classes based on types of bullying: item probabilities.

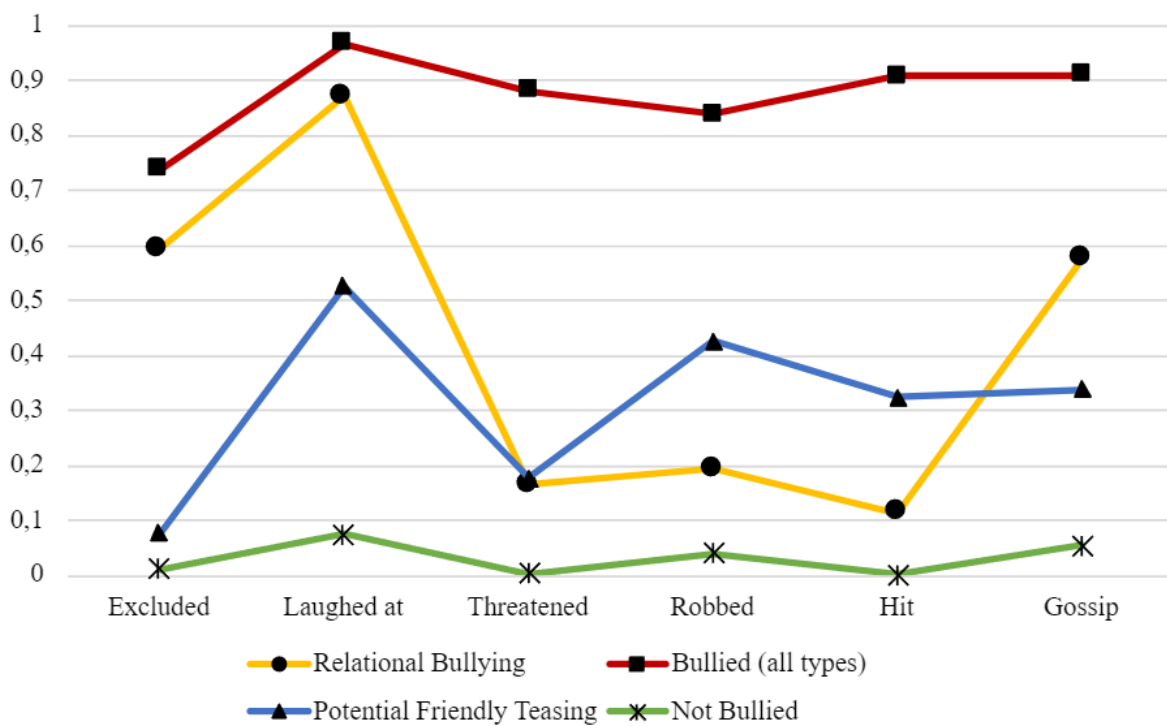


Table 1. Latent Class Models Fit to Bullying Behaviors.

Models	Model fit criteria					Diagnostic criteria		
	LL	AIC	BIC	SABIC	LMR (p)	BLRT (p)	Smallest class size (%)	Entropy
<i>Total sample</i>								
1 Class	-31,007.730	62,027.460	62,071.612	62,052.545	—	—	100	—
2 Classes	-25,146.702	50,319.404	50,415.067	50,373.755	11,545.813 ($<.001$)	11,722.056 ($<.001$)	23	0.848
3 Classes	-24,532.157	49,104.314	49,251.487	49,187.929	1,210.611 ($<.001$)	1,229.091 ($<.001$)	6	0.801
4 Classes	-24,412.339	48,878.677	49,077.362	48,991.559	236.033 ($<.001$)	239.636 ($<.001$)	6	0.757
5 Classes	-24,395.375	48,858.750	49,108.945	49,000.897	32.235 (.337)	32.727 ($<.001$)	4	0.747
<i>Girls</i>								
1 Class	-14,307.973	28,627.945	28,667.970	28,648.903	—	—	100	—
2 Classes	-11,953.636	23,933.272	24,019.992	23,978.682	4,632.352 ($<.001$)	4,708.673 ($<.001$)	23	0.831
3 Classes	-11,744.173	23,528.345	23,661.761	23,598.206	412.137 ($<.001$)	418.927 ($<.001$)	4	0.804
4 Classes	-11,676.039	23,406.077	23,586.188	23,500.390	134.059 ($<.001$)	136.268 ($<.001$)	4	0.766
5 Classes	-11,664.182	23,396.364	23,623.170	23,515.128	23.329 (.029)	23.714 (.013)	4	0.795
<i>Boys</i>								
1 Class	-16,488.713	32,989.425	33,029.387	33,010.321	—	—	100	—
2 Classes	-12,928.710	25,883.420	25,970.003	25,928.693	7,004.462 ($<.001$)	7,120.006 ($<.001$)	19	0.872
3 Classes	-12,569.110	25,178.220	25,331.425	25,247.871	707.528 ($<.001$)	719.200 ($<.001$)	7	0.801
4 Classes	-12,519.491	25,092.983	25,272.810	25,187.012	97.627 ($<.001$)	99.237 ($<.001$)	5	0.788
5 Classes	-12,506.293	25,080.586	25,307.035	25,198.993	25.968 (.019)	26.397 ($<.001$)	4	0.738

Note. LL = Log-likelihood; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SABIC = Adjusted BIC; LMR = Lo-Mendell-Rubin log-likelihood ratio test; BLRT = Bootstrapped likelihood ratio test.

Multigroup LCA

To test measurement invariance between boys and girls across latent classes, two models (the heterogeneous model and the homogeneous model) were tested and compared. Table 2 presents the fit criteria for those two models, and Table 3 presents a relative model fit.

Most absolute-fit statistics (log-likelihood, AIC, and SABIC) provide better values for the heterogeneous model. Moreover, regarding relative model fit, the heterogeneous model fits significantly better than the homogeneous model. According to these results, it can be said that conditional response probabilities are not the same for girls and boys, which means that latent classes do not have the same sizes across genders. Therefore, measurement invariance is not met between boys and girls across latent classes.

Table 2. Absolute Model Fit for Measurement Invariance.

Model	Log-likelihood	AIC	BIC	SABIC	df	L ²
Heterogeneous	-32235.184	64580.367	64985.095	64810.311	72	135.803
Homogeneous	-32324.333	64710.666	64938.785	64840.271	96	314.102

Table 3. Relative Model Fit for Measurement Invariance.

Model	L ²	df (deviance)	Deviance	p
Homogeneous	314.102	NaN	NaN	NaN
Heterogeneous	135.803	24	178.299	< .001

Class Prevalence by Group and Multinomial Logistic Regression of Latent Classes

To examine potential sources of the lack of measurement invariance, class prevalence by group was calculated, and a multinomial logistic regression was performed. Thus, the differences between girls and boys across classes were further examined, and gender's influence on class membership was tested. Table 4 displays class prevalence by group. Different class distributions are found for boys and girls.

It can be seen that the majority of girls (71.1%) tend to belong to the *Not Bullied* class, as well as the majority of boys, although with a smaller percentage (65.2%). Differences are more pronounced among the other three classes. On one hand, 17% of girls belong to the *Relational Bullying* class, as opposed to the 7.8% of boys who do so. On the other hand, a similar reverse pattern can be found in the *Potential Friendly Teasing* class: 17.5% of boys belong to it, as opposed to the 8% of girls. Last, 9.5% of boys belong to the *Bullied (All Types)* class, as opposed to only 3.8% of girls. The differences are clear: Whereas for boys, *Potential Friendly Teasing* is the most likely class after *Not Bullied*, it is the second least likely class for girls, after *Bullied (All types)*. Additionally, whereas for girls, *Relational Bullying* is the most likely class after *Not Bullied*, it is the least likely class for boys, even less likely than the *Not Bullied* class.

Table 4. Class Prevalence by Group.

	Bullied (All-Types)	Relational Bullying	Potential Friendly Teasing	Not Bullied
Female	.038 (3.8%)	.170 (17%)	.080 (8%)	.711 (71.1%)
Male	.095 (9.5%)	.078 (7.8%)	.175 (17.5%)	.652 (65.2%)

Finally, Table 5 displays the results of the multinomial logistic regression. Gender was the predictor variable, and the latent class membership was the outcome variable. The *Not Bullied* class was used as the reference class, and girls were set as the reference group for gender. Note that an odds ratio (OR) < 1 means that the reference group (girls) is more related to a specific latent class, and an OR > 1 means that the comparison group (boys) is more related to a specific latent class.

Compared to the *Not Bullied* class, (a) boys are 2.7 times more likely than girls to be in the *Bullied (All Types)* class, (b) boys are almost 2.4 times more likely than girls to be in the *Potential Friendly Teasing* class, and (c) girls are almost twice as likely as boys to be in the *Relational Bullying* class. These results are consistent with those obtained from the class prevalence by group.

Table 5. Results of the multinomial logistic regression.

	b (SE)	OR	OR 95% CI
<i>Bullied (all-types) vs. Not Bullied</i>			
Intercept	-3.933*** (0.188)		
Girls (ref.) vs. Boys	1.003*** (0.099)	2.726	[2.245-3.311]
<i>Relational Bullying vs. Not Bullied</i>			
Intercept	-0.738*** (0.147)		
Girls (ref.) vs. Boys	-0.690*** (0.118)	0.501	[0.398-0.632]
<i>Potential Friendly Teasing vs. Not Bullied</i>			
Intercept	-3.051*** (0.322)		
Girls (ref.) vs. Boys	0.869*** (0.148)	2.384	[1.784-3.185]

Note. ***p < .001.

Discussion

This study was intended to explore boys' and girls' response processes to bullying victimization items and address measurement invariance between the gender groups to reject the existence of measurement bias, which could threaten validity and fairness in bullying assessment. An LCA revealed four latent classes or profiles. Those classes represent different response patterns to the *Students' Experience of Being Bullied Scale* from the PISA 2018. The results of a multigroup LCA show that girls and boys display differences in those patterns.

The lack of measurement invariance may indicate that the genders live very different experiences of bullying victimization, which in turn may influence their understanding of bullying realities. Those differences in the understanding of the phenomenon can mediate the response processes girls and boys develop when answering bullying items, a situation that fosters the existence of bias in the measurement process. Comparative interpretations of bullying victimization scores between boys and girls should therefore be taken with caution, because fair comparisons cannot be made if the same construct has not been assessed for both gender groups. Therefore, gender differences in response processes should be further examined using qualitative methods, such as cognitive interviewing and/or web probing. Both qualitative methods could provide evidence of the response process, such as the respondents' interpretations of the items, experiences of being bullied that come to their minds when they respond to the items, decisions about the response options they chose, and reports of any difficulties encountered with the scale items (Beatty & Willis, 2007; Behr et al., 2017). Both cognitive interviewing and web probing could also serve as validity evidence of students' responses. To investigate girls' and boys' response processes more thoroughly, it is important to know how different their bullying experiences are, to what extent those differences can be a potential threat to the validity of comparative interpretations of bullying victimization scores,

and what the sources of such differences are. For example, gender stereotypes and roles can play a part in determining how different the bullying experiences for boys and girls are. Youths “do gender” when interacting with others, that is, “their specific type of gender construction articulates with particular social situations” (Messerschmidt, 2009, p. 86). Also, youths are influenced by the acceptance or rejection of their peers when “doing gender,” making them more likely to conform to traditional gender stereotypes and roles, because peers tend to accept people more when their gender expression is congruent with their “sex category” (Messerschmidt, 2009).

In this study, a first step in examining the extent to which differences in girls’ and boys’ response processes can compromise the validity of group comparisons in bullying victimization assessments was taken. The realities of being bullied and not being bullied are represented in two profiles, consistent with the *All-Types* and *Uninvolved* classes, which were identified in previous studies (e.g., Lee et al., 2022; Nylund et al., 2007; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022). It was also found that it is more likely for boys to be bullied, and the opposite for girls, that is, it is more likely for girls to not be bullied, consistent with the literature (e.g., Carrera-Fernández et al., 2013; Hanzi & Akça, 2020; Lee et al., 2022; OECD, 2019b; Wang et al., 2010; Zhao et al., 2022). Note that this does not imply that girls are not bullied, but it seems that girls are less likely to be bullied than boys. Moreover, in line with previous literature that specified the existing types of bullying (e.g., Björkqvist et al., 1992; Olweus, 1994; Rivers & Smith, 1994), the third profile found in our study corresponds to students who are suffering from relational bullying, an indirect type of bullying, which is more likely for girls to experience (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Rivers & Smith, 1994; Rosen & Nofziger, 2019; Shu et al., 2023). Surprisingly, the last profile did not match the other type of bullying—direct bullying—as we expected. This profile, which we called Potential Friendly Teasing, comprises students who have medium probabilities of being bullied

(specifically of being laughed at) but a low probability of being excluded by their peers, which can be considered a contradiction, based on previous findings that link bullying victimization with more loneliness and isolation (e.g., OECD, 2019b; UNESCO, 2019). Boys were also more likely to belong to this class. This finding could be related to the fact that the PISA 2018 revealed that boys reported having a stronger sense of belonging at school than girls despite experiencing more bullying than their female counterparts (OECD, 2019b).

Based on the literature on bullying, we came across three possible explanations for the *Potential Friendly Teasing* profile. The first one, which inspired the class label, would be that this profile comprises students who are bullied by their friends, which is a reality that has not been widely studied (Mishna et al., 2008; Wei & Jonson-Reid, 2011). Some bullied students may not even recognize aggressions perpetrated by their friends as bullying (Mishna et al., 2008), either because cognitive dissonance does not let them recognize their friend is hurting them or because they consider those behaviors fun and friendly teasing that is generally mutual and part of the way they bond (Mishna et al., 2008; Wei & Jonson-Reid, 2011). Boys appear to be more prone to finding activities that involve aggression fun and enjoyable (Wei & Jonson-Reid, 2011), which would be consistent with our finding that boys are more likely to be part of this *Potential Friendly Teasing* profile. The second explanation could be that this profile comprises students who are bullied but have friends who are on their side and act as allies or defenders (e.g., Chen et al., 2020). Lastly, the third would be that this profile comprises students who have “frenemies” (Wójcik & Flak, 2021), students who act as the bullied student’s friend when the bully is not around but stop doing so when the bully appears. In the latter situation, they can act as disengaged onlookers (students who do not get involved in the bullying situation) or even as active or passive supporters for the bully (Wójcik & Flak, 2021). These ambivalent behaviors could be enough for the bullied student to not feel excluded. These three possibilities also need to be further examined in future studies with cognitive interviews and/or

web probing to explore the realities of students of this profile and to develop a proper comprehensive explanation of the *Potential Friendly Teasing* profile.

The main objective of this study is entirely related to diversity and social inclusion issues. Ensuring fairness in evaluation when assessing demographic groups is crucial to understand relevant social phenomena, such as bullying experiences at school. In this research, we focused on fairness by examining measurement equivalence between girls' and boys' bullying experiences assessments. Obtaining validity evidence of response processes to ensure fair assessments and valid comparisons of test and scale scores is crucial to make fair inferences and valid interpretations without advantaging or disadvantaging any group of people in evaluations and interventions. With that in mind, following an ecological viewpoint (Zumbo et al., 2015), the vital experiences of all groups of comparison can be analyzed, as we have done in this study with boys' and girls' bullying experiences, to fully understand the phenomenon and develop effective interventions taking into account diversity (e.g., developing targeted interventions for girls and boys).

Limitations and Future Research

This is a pre-pandemic sample, given that these participants were recruited for the PISA 2018. We opted to use the PISA 2018 instead of the PISA 2022 because the latter cycle was conducted during the COVID-19 pandemic, when most students alternated between in-person and online classes, highlighting the possibility that the pandemic context greatly influenced the results. Nevertheless, studies with a post-pandemic sample could be more faithful to current social realities. Society has been struggling with many changes since the COVID-19 pandemic, and experiences in classrooms have changed. That is why researchers should examine the influence of changes in society due to the pandemic with a post-pandemic sample and possibly conduct a latent transition analysis (e.g., Abarda et al., 2020) to test changes across time.

Moreover, future studies with international or cross-cultural samples could provide evidence of the extent to which findings depend on cultural differences by carrying out a multilevel model (e.g., Peugh, 2010) controlling for country. A multilevel model can also be useful to control for schools in future studies to take into account, for instance, the influence of the different kinds of schools' respondents attend.

Another limitation is the low number of items the scale used in this study which made it difficult to obtain more detailed measures of bullying experiences. Future studies should try to replicate these results with longer scales that may add more information. Additionally, we collapsed some response categories in this study, leaving response options dichotomized. Future research could analyze the differences in frequency or intensity of bullying victimization by analyzing data sets with polytomous responses.

Implications for Practice

Relying on test scores to guide classroom intervention could represent a double-edged sword, because if interpretations of test scores are biased, interventions will be biased or not effective enough. It is important to carry out effective and fair interventions to address and prevent bullying cases in classrooms, and in order to do so, it is necessary to understand the different realities in place and reject the possibility of biases endangering the validity and fairness of interpretations of test scores. For example, a screener could help, as long as it is accompanied by a proper follow-up (e.g., via cognitive interviewing) in which the evaluator ensures the validity and fairness of comparative interpretations of scores. Is it true that girls suffer from less bullying victimization, or does it appear to be so because girls suffer from a kind of bullying victimization that is more difficult to detect and flag as a bullying behavior (relational bullying)? Are some cases of bullying not being detected in classrooms because they appear to be friendly teasing? These questions need to be asked to effectively care for and

protect students, and they can be answered through a validation of test scores that involves measurement invariance and examinations of the students' response processes. The understanding of differences in students' response processes when responding to bullying scales items could guide the development of targeted interventions for boys and girls.

Declarations

Author Contributions

Following the CRediT roles taxonomy (available in <https://credit.niso.org/>), contributions are: M-CN-G: conceptualization, data curation, formal analysis, visualization, writing – original draft, writing – review & editing; J-LP: conceptualization, supervision, writing – review & editing; AP: formal analysis, writing – review & editing. All authors read and approved the final manuscript.

Data Availability

The author(s) declare that all materials (e.g., scripts, databases) needed to replicate the results derived from this study can be found at: https://osf.io/uqje6/?view_only=2c78aa31295a4429a8033dccaace62c2.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the FPU Program of the Spanish Ministry of Education, Culture, and Sport under Grant (FPU19/02908), and the Andalusian Government R&D program (PY20_00898).

References

- Abarda, A., Dakkon, M., Azhari, M., Zaaloul, A., & Khabouze, M. (2020). Latent Transition Analysis (LTA): A method for identifying differences in longitudinal change among unobserved groups. *Procedia Computer Science*, 170, 1116–1121.
<https://doi.org/10.1016/j.procs.2020.03.059>
- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
<https://www.apa.org/science/programs/testing/standards>
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311.
<https://doi.org/10.1093/poq/nfm006>
- Behr, D., Meitinger, K., Braun, M., & Kaczmarek, L. (2017). *Web probing—implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions*. GESIS Survey Guidelines.
https://doi.org/10.15465/gesis-sg_en_023
- Björkqvist, K., Lagerspetz, K. M. J., & Kaukiainen, A. (1992). Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive*

Behavior, 18(2), 117–127. [https://doi.org/10.1002/1098-2337\(1992\)18:2%3C117::AID-AB2480180205%3E3.0.CO;2-3](https://doi.org/10.1002/1098-2337(1992)18:2%3C117::AID-AB2480180205%3E3.0.CO;2-3)

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. The Guilford Press.

Carrera-Fernández, M. V., Lameiras-Fernández, M., Rodríguez-Castro, Y., Failde-Garrido, J. M., & Calado-Otero, M. (2013). Bullying in Spanish secondary schools: Gender-based differences. *Spanish Journal of Psychology*, 16(21), 1–14. <https://doi.org/10.1017/sjp.2013.37>

Chen, G., Zhang, W., Zhang, W., & Deater-Deckard, K. (2020). A “defender protective effect” in multiple-role combinations of bullying among Chinese adolescents. *Journal of Interpersonal Violence*, 35(7–8), 1587–1609. <https://doi.org/10.1177/0886260517698278>

Chung, J. Y., & Lee, S. (2020). Are bully-victims homogeneous? Latent class analysis on school bullying. *Children and Youth Services Review*, 112, 104922. <https://doi.org/10.1016/j.childyouth.2020.104922>

Díaz-Aguado, M. J., Martínez-Arias, R., Falcón, L., & Alvariño, M. (2023). *Acoso escolar y ciberacoso en España en la infancia y en la adolescencia*. Fundación ColaCao y Universidad Complutense de Madrid. <https://fundacioncolacao.org/investigacion>

DiStefano, C., Shi, D., & Morgan, G. B. (2020). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(2), 237–249. <https://doi.org/10.1080/10705511.2020.1803073>

- Falla, D., Ortega-Ruiz, R., Runions, K., & Romera, E. M. (2020). Why do victims become perpetrators of peer bullying? Moral disengagement in the cycle of violence. *Youth & Society*, 54(3), 397–418. <https://doi.org/10.1177/0044118X20973702>
- Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decálogo para el Análisis Factorial de los ítems de un test. *Psicothema*, 34(1), 7–17. <https://doi.org/10.7334/psicothema2021.456>
- Fundación ANAR. (2022). *Evolución de la violencia a la infancia en España según las víctimas*. Fundación ANAR. https://www.anar.org/wp-content/uploads/2021/12/Estudio_Evolucio%CC%81n-de-la-Violencia-a-la-Infancia-en-Espan%CC%83a-segu%CC%81n-las-Vi%CC%81ctimas-2.pdf
- Fundación ANAR & Fundación Mutua Madrileña. (2023). *La opinión de los/as estudiantes. V Informe de Prevención del Acoso Escolar en Centros Educativos*. Fundación ANAR y Fundación Mutua Madrileña. <https://www.anar.org/anary-mutua-madrilena-presentan-el-v-informe-la-opinion-de-los-estudiantes-sobreacoso-escolar/>
- Gao, T., Mei, S., Zhou, X., Cao, H., Liang, L., Zhou, C., & Meng, X. (2023). Latent classes of bullying perpetration and victimization among adolescents: Associations with problem behaviors. *Children and Youth Services Review*, 150, 107013. <https://doi.org/10.1016/j.childyouth.2023.107013>
- Gladden, M. R., Vivolo-Kantor, A. M., Hamburger, M. E., & Lumpkin, C. D. (2014). *Bullying surveillance among youths: Uniform definitions for public health and recommended data elements. Version 1.0*. Centers for Disease Control and Prevention. <https://stacks.cdc.gov/view/cdc/21596>

- Goldweber, A., Waasdorp, T. E., & Bradshaw, C. P. (2013). Examining associations between race, urbanicity, and patterns of bullying involvement. *Journal of Youth and Adolescence*, 42, 206–219. <https://doi.org/10.1007/s10964-012-9843-y>
- Hanzi, I. H., & Akça, S. Ö. (2020). School bullying. *Journal of Education and Future*, 18, 113–121. <https://doi.org/10.30786/jef.691717>
- He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2), 1–19. <https://doi.org/10.9707/2307-0919.1111>
- Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., Wolfe, M., & Reid, G. (2015). Bullying and suicidal ideation and behaviors: A meta-analysis. *Pediatrics*, 135(2), 496–509. <https://doi.org/10.1542/peds.2014-1864>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Kankaraš, M., Moors, G., & Vermunt, J. K. (2018). Testing for measurement invariance with Latent Class Analysis. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis, methods and applications (2nd ed.)*. Routledge. <https://doi.org/10.4324/9781315537078>
- Kim, Y. S., & Leventhal, B. (2008). Bullying and suicide. A review. *International Journal of Adolescent Medical Health*, 20(2), 133–154. <https://doi.org/10.1515/ijamh.2008.20.2.133>
- Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In I. B. Weigner (Ed.), *Handbook of psychology (2nd ed.)*. Wiley.

- Lanza, S. T., & Cooper, B. R. (2016). Latent Class Analysis for developmental research. *Child Development Perspectives*, 10(1), 59–64. <https://doi.org/10.1111/cdep.12163>
- Lee, J. (2021). Pathways from childhood bullying victimization to young adult depressive and anxiety symptoms. *Child Psychiatry & Human Development*, 52, 129–140. <https://doi.org/10.1007/s10578-020-00997-4>
- Lee, J., Abel, N., & Holmes, J. L. (2017). Validation of measures of cyberbullying perpetration and victimization in emerging adulthood. *Research on Social Work Practice*, 27(4), 456–467. <https://doi.org/10.1177/1049731515578535>
- Lee, J., Roh, B. R., & Yang, K. E. (2022). Exploring the association between social support and patterns of bullying victimization among school-aged adolescents. *Children and Youth Services Review*, 136, 106418. <https://doi.org/10.1016/j.childyouth.2022.106418>
- Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), 1151–1169. <https://doi.org/10.6018/analesps.30.3.199361>
- Messerschmidt, J. W. (2009). “Doing gender”: The impact and future of a salient sociological concept. *Gender and Society*, 23(1), 85–88. <https://doi.org/10.1177/0891243208326253>
- Mishna, F., Wiener, J., & Pepler, D. (2008). Some of my best friends—experiences of bullying within friendships. *School Psychology International*, 29(5), 549–573. <https://doi.org/10.1177/0143034308099201>
- Modecki, K. L., Minchin, J., Harbaugh, A. G., Guerra, N. G., & Runions, K. C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, 55, 602–611. <https://doi.org/10.1016/j.jadohealth.2014.06.007>

- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide (8th ed.)*. Muthén & Muthén.
- Muthén, B., Muthén, L., & Asparouhov, T. (2015). *Estimator choices with categorical outcomes*. <https://www.statmodel.com/download/EstimatorChoices.pdf>
- Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among U.S. youth prevalence and association with psychological adjustment. *JAMA: The Journal of the American Medical Association*, 285(16), 2094–2100. <https://doi.org/10.1001/jama.285.16.2094>
- Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does Latent Class Analysis say? *Child Development*, 78(6), 1706–1722. <https://doi.org/10.1111/j.1467-8624.2007.01097.x>
- Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about Latent Class Analysis. *Translational Issues in Psychological Science*, 4(4), 440–461. <https://doi.org/10.1037/tps0000176>
- Observatorio Estatal de la Convivencia Escolar. (2010). *Estudio estatal sobre la convivencia escolar en la educación secundaria obligatoria*. <https://sede.educacion.gob.es/publiventa/detalle.action?cod=13567>
- OECD. (2017a). *PISA 2015. Informe Español. El bienestar de los estudiantes*. OECD Publishing. https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=18701
- OECD. (2017b). *PISA 2015 results (volume III): Students' well being*. OECD Publishing. <https://www.oecd.org/pisa/publications/pisa-2015-results-volumeiii-9789264273856-en.htm>

- OECD. (2018). *PISA 2018 technical report*. OECD Publishing.
<https://www.oecd.org/pisa/data/pisa2018technicalreport/>
- OECD. (2019a). *PISA 2018. Programa para la Evaluación Internacional de los Estudiantes*. Informe Español. OECD Publishing.
https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=20372
- OECD. (2019b). *PISA 2018 results (volume III): What school life means for students' lives*. OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-iii_acd78851-en
- Olweus, D. (1978). *Aggression in the schools: Bullies and whipping boys*. Hemisphere.
- Olweus, D. (1994). Bullying at school: Basic facts and effects of school based intervention program. *Journal of Child Psychology and Psychiatry*, 35(7), 1171–1190.
<https://doi.org/10.1111/j.1469-7610.1994.tb01229.x>
- Olweus, D. (2013). School bullying: Development and some important challenges. *Annual Review of Clinical Psychology*, 9, 751–780. <https://doi.org/10.1146/annurev-clinpsy-050212-185516>
- Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>
- Rivers, I., & Smith, P. K. (1994). Types of bullying behavior and their correlates. *Aggressive Behavior*, 20, 359–368. [https://doi.org/10.1002/1098-2337\(1994\)20:5%3C359::AID-AB2480200503%3E3.0.CO;2-J](https://doi.org/10.1002/1098-2337(1994)20:5%3C359::AID-AB2480200503%3E3.0.CO;2-J)

- Rosen, N. L., & Nofziger, S. (2019). Boys, bullying, and gender roles: How hegemonic masculinity shapes bullying behavior. *Gender Issues*, 36, 295–318. <https://doi.org/10.1007/s12147-018-9226-0>
- Rudnev, M. (2018). Testing for invariance of latent classes: Group-as-covariate approach. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Crosscultural analysis, methods and applications (2nd ed.)*. Routledge. <https://doi.org/10.4324/9781315537078>
- Shu, Y., Ota, M., & Kato, H. (2023). The diverse nature of school bullying and its relation to depression and the efficacy of coping strategies in Japan. *International Journal of Bullying Prevention*. Advance online publication. <https://doi.org/10.1007/s42380-023-00181-x>
- Swearer, S. M., & Hymel, S. (2015). Understanding the psychology of bullying. Moving toward a social-ecological diathesis-stress model. *American Psychologist*, 70(4), 344–353. <https://doi.org/10.1037/a0038929>
- Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2), 169–181. <https://doi.org/10.1108/QAE-06-2017-0034>
- UNESCO. (2019). *Behind the numbers: Ending school violence and bullying*. UNESCO. <https://doi.org/10.54675/TRVR4270>
- Wang, J., Iannotti, R. J., Luk, J. W., & Nansel, T. R. (2010). Co-occurrence of victimization from five subtypes of bullying: Physical, verbal, social exclusion, spreading rumors and cyber. *Journal of Pediatric Psychology*, 35(10), 1103–1112. <https://doi.org/10.1093/jpepsy/jsq048>

- Wei, H., & Jonson-Reid, M. (2011). Friends can hurt you: Examining the coexistence of friendship and bullying among early adolescents. *School Psychology International*, 32(3), 244–262. <https://doi.org/10.1177/0143034311402310>
- Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A guide to best practice. *Journal of Black Psychology*, 46(4), 287–311. <https://doi.org/10.1177/0095798420930932>
- Wójcik, M., & Flak, W. (2021). Frenemy: A new addition to the bullying circle. *Journal of Interpersonal Violence*, 36(19–20), 11131–11154. <https://doi.org/10.1177/0886260519880168>
- Zhao, D., Wang, D., He, Z., Yuan, S., Zhu, D., & Liu, H. (2022). Victim profiles and the protective role of school anti-bullying norms: A study of Chinese adolescents. *Current Psychology*, 42, 31835–31852. <https://doi.org/10.1007/s12144-022-04204-6>
- 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. <https://doi.org/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. <https://doi.org/10.1080/15434303.2014.972559>

CAPÍTULO VI: ESTUDIO III

**DO DIFFERENT EXPERIENCES OF BULLYING IMPACT
RESPONSE PROCESSES? A MIXED-METHODS APPROACH
TO OBTAIN VALIDITY EVIDENCE OF BULLYING SCALES'
SCORES**

**¿LAS DIFERENTES EXPERIENCIAS DE BULLYING
INFLUYEN EN LOS PROCESOS DE RESPUESTA? UNA
APROXIMACIÓN MIXTA PARA OBTENER EVIDENCIAS DE
VALIDEZ DE LAS PUNTUACIONES DE UNA ESCALA DE
BULLYING**

M. Carmen Navarro-González^{1,2} and José-Luis Padilla^{1,2}

¹Department of Methodology for Behavioral Sciences, University of Granada, Granada, Spain

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

This chapter is submitted to *Journal of Interpersonal Violence*.

Resumen

Conocer cómo los chicos y las chicas interpretan las experiencias de *bullying* es esencial para asegurar evaluaciones justas y válidas a partir de las que diseñar intervenciones eficaces y libres de sesgo. Las diferentes experiencias de *bullying* de chicos y chicas pueden influir en su comprensión del fenómeno, y como consecuencia, provocar diferencias en los procesos de respuesta a los ítems de los instrumentos de evaluación. Este estudio aporta evidencias de validez basadas en los procesos de respuesta para la *escala de experiencias de bullying* (ST038) del Cuestionario del Estudiante del PISA 2018. Utilizando un diseño mixto explicativo, exploramos las experiencias de *bullying* de chicas y chicos, y cómo esas diferentes experiencias influyeron en sus procesos de respuesta al contestar los ítems de la escala. Se obtuvieron datos cuantitativos y cualitativos mediante un cuestionario web y entrevistas cognitivas a 20 estudiantes españoles de secundaria (9 chicas, 11 chicos; $M = 15.25$ años; $DT = 0.44$). Los resultados revelaron diferencias de género notables en las experiencias de *bullying* y en los procesos de respuesta. Los chicos reportaron experimentar más agresiones directas, pero tendían a interpretarlas como bromas no problemáticas entre amigos. Las chicas reportaron experimentar más agresiones indirectas, y generalmente reconocieron verse afectadas psicológicamente por ellas. Discutimos cómo las entrevistas cognitivas aportan evidencias de validez sobre los procesos de respuesta a los ítems de victimización de *bullying*, así como las implicaciones para la validez y la justicia de las interpretaciones comparativas, y para el diseño de intervenciones no sesgadas.

Palabras clave: acoso escolar, diferencias de género, procesos de respuesta, entrevista cognitiva, métodos mixtos, validez.

Abstract

Understanding how boys and girls interpret bullying experiences is essential to ensure fair and valid assessments, which in turn serve as the basis for designing effective and unbiased interventions. The differing experiences of bullying among boys and girls may shape their understanding of the phenomenon and, consequently, lead to variations in how they respond to items in assessment instruments. This study presents validity evidence based on response processes for the *Students' Experience of Being Bullied Scale* (ST038) included in the PISA 2018 Student Questionnaire. Employing an explanatory mixed-methods design, we examined the bullying experiences of girls and boys, and how these differing experiences influenced their response processes and patterns when responding to the scale's items. Quantitative and qualitative data were collected through a web-based questionnaire and cognitive interviews conducted with 20 Spanish high-school students (9 girls, 11 boys; $M = 15.25$ years; $SD = 0.44$). The results revealed notable gender-based differences in both bullying experiences and response processes. Boys were more likely to experience direct forms of aggression, yet tended to interpret them as "friendly teasing" rather than problematic behavior. In contrast, girls were more likely to experience indirect forms of aggression, and generally reported being psychologically affected by them. The implications for validity, fairness, and the design of bias-free interventions, as well as how cognitive interviews can provide validity evidence regarding response processes for items on bullying victimization scales, are discussed.

Keywords: bullying, gender differences, response processes, cognitive interviewing, mixed-methods, validity.

Introduction

Bullying is an unwanted, aggressive behavior intended to harm a person seen as less powerful (Gladden et al., 2014; Lee, 2021; Olweus, 1978, 2013). Consequences of bullying can be devastating for children and adolescents, eventually deriving in psychological (e.g., depression, anxiety, trauma, suicidal ideation), relational (e.g., increased loneliness, risky sexual behaviors), and educational problems (e.g., poor academic performance, increased school absenteeism, and low sense of belonging to school) (e.g., Cheng et al., 2020; Holt et al., 2015; OECD, 2019; Snodgrass et al., 2024; Swearer & Hymel, 2015; UNESCO, 2019). All these negative consequences make bullying a significant social and educational problem that needs to be prevented in schools and high schools.

Prevalences of bullying vary widely depending on the study (from less than 10% to over 65%; Chung & Lee, 2020). For example, the Preventive Psychology Unit of the Complutense University of Madrid and ColaCao Foundation (Díaz-Aguado et al., 2023) found that, whereas only 6.2% of students identified themselves as bullying victims, 19.2% of students reported having received some kind of abusive behavior from their peers (e.g., name calling, teasing, etc.). Differences in prevalence like these suggest that, at least partially, the wording of questions and item stems in bullying victimization scales matters for accurately capturing the real nature of bullying in schools, as differences in wording can lead to differences in how respondents interpret and answer those items.

Differences in the response processes can affect the validity of scores interpretations from bullying victimization measurement instruments, especially when such differences are associated with distinct demographic groups. If a biased interpretation of scores —stemming from group differences in response processes— is made, it may result in the underrepresentation of bullying, which in turn can lead to ineffective interventions. Moreover, when comparisons between different student groups are based on biased interpretations of scores, not only is validity compromised, but fairness is also at risk, as some comparison groups may be “advantaged or disadvantaged because of characteristics irrelevant to the intended construct” (American Educational Research Association [AERA] et al., 2014, p.50). This situation could lead to biased and unfair interventions, highlighting issues related to validity evidence based on the consequences of testing (AERA et al., 2014). Thus, valid and fair interpretations of bullying victimization scores are crucial to comprehensively understand the phenomenon and, in turn, to develop effective and fair interventions. To achieve this, it is necessary to verify whether the phenomenon being measured aligns with the intended construct —validity assessment—, and to examine if the same phenomenon is being measured across all comparison groups —measurement invariance assessment.

In this study, we focus on providing validity evidence of response processes when comparing Spanish girls’ and boys’ bullying victimization scores. Prior research has shown that these groups have different experiences with bullying (e.g., Rosen & Nofziger, 2021; Shu et al., 2023), which may result in a situation where the same phenomenon is not being measured across both groups, endangering validity and fairness of the assessment. To address this, we explored response processes of boys and girls to the *Students’ Experience of Being Bullied Scale* used in the Programme for International Student Assessment [PISA] 2018 Student Questionnaire. This scale was selected due to the widespread influence the PISA program has all around the world, as well as the scale consisting of items that intend to capture both direct

and relational bullying. Although a more recent PISA cycle was published in 2022, we opted for the PISA 2018 because the PISA 2022 was conducted during the COVID-19 pandemic, a period when most schools operated remotely, potentially affecting bullying prevalence and experiences. In fact, prevalence of bullying in PISA 2022 compared to PISA 2018 has decreased (OECD, 2023) in most countries, and the influence of the pandemic on those results cannot be ruled out.

Gender Differences in Bullying Victimization

It has been shown that girls' and boys' experiences of bullying victimization differ significantly. On the one hand, boys tend to both perpetrate and experience more direct forms of aggression intended to cause immediate harm (physical and verbal aggressions, such as hitting, kicking, pushing, insulting, or threatening), commonly labeled as “direct bullying”. On the other hand, girls are more likely to engage and be targeted by indirect forms of aggression aimed at damaging reputation and/or relationships (relational aggressions, such as gossiping or social exclusion, commonly referred to as “indirect” or “relational bullying”) (e.g., Rosen & Nofziger, 2021; Shu et al., 2023). In addition, boys tend to report having suffered more bullying than girls (e.g., Hanci & Akça, 2020; Lee et al., 2022; OECD, 2019; Zhao et al., 2022).

We found a similar pattern in a previous study (Navarro-González et al., 2025), which used Latent Class Analysis (LCA) to explore response patterns to identify different profiles of bullying victims, as well as to examine differences in these profiles between girls and boys through a multigroup LCA assessing measurement invariance. Measurement invariance between boys and girls was not supported, reinforcing the idea of girls and boys having very different experiences of bullying victimization: girls were more likely than boys to experience relational bullying, whereas boys tended to experience more bullying overall. Interestingly, instead of identifying a profile corresponding to “direct bullying”, we found a profile

characterized by moderate probabilities of experiencing various forms of bullying, except for social exclusion. Boys were more likely than girls to belong to this profile. These gender differences may compromise the validity of comparative interpretations of bullying victimization scores between girls and boys, as fairness requires measurement invariance, that is, the same phenomenon must be measured across groups (AERA et al., 2014).

Therefore, it becomes necessary to further examine how differences in bullying victimization experiences between boys and girls may lead to distinct understandings or cognitive representations of the phenomenon. These differences, in turn, could result in variations in their response processes to bullying victimization items, thereby increasing the risk of biased assessment (He & van de Vijver, 2012).

Response Processes to Bullying Victimization Items

One way to examine the different experiences of bullying and how they affect the interpretations of item stems is by exploring response processes when answering bullying victimization items. This helps determine whether the items “reflect the same construct for all test takers” and “the same meaning or interpretation for all individuals” (AERA et al., 2014, p.50). In other words, it provides validity evidence based on response processes with the aim of addressing fairness in testing.

When examining response processes, researchers explore the different cognitive “routes” respondents take to answer an item. According to the “question-and-answer” model developed by Tourangeau (2018), this cognitive “route” involves four stages, which do not necessarily occur in a linear sequence: 1) comprehension of item stem; 2) retrieval of information requested; 3) judgment and estimation; and 4) reporting the answer.

If the different experiences and understandings that boys and girls have regarding bullying victimization influence how they comprehend the items, retrieve relevant information from memory, or estimate and report their answers, then the assessment may not reflect the same phenomenon across gender groups. This would compromise the fairness of interpretations of test scores.

Therefore, it is essential to examine whether these cognitive “routes” are the same for girls and boys when responding to bullying victimization items, and if not, to understand in what way they differ. Furthermore, according to the Ecological Model of Item Response (Zumbo et al., 2015), it is also necessary to investigate potential contextual explanations for any lack of measurement invariance, in order to understand why such differences arise.

A Mixed-Methods Approach to Validation of Response Processes

Mixed-methods research enables a more comprehensive understanding of the phenomenon under study, by considering “multiple viewpoints, perspectives, positions, and standpoints” (Johnson et al., 2007, p.113), thereby overcoming the limitations of quantitative and qualitative methods alone (Creswell & Plano Clark, 2011). Several studies (e.g., Benítez et al., 2019; Castillo-Díaz & Padilla, 2013) have adopted a mixed-methods approach to explore respondents’ response processes by integrating both quantitative and qualitative data and analyses.

In this context, mixed-methods research can be instrumental in gathering information about differences in response processes of girls and boys when answering bullying victimization items, as well as in identifying the sources of any observed lack of measurement invariance. This approach helps to better understand the nature of potential biases present in the assessment process. Such insights could be achieved through a mixed-methods study that

integrates quantitative and qualitative methods at the design, method, and reporting levels (Fetters et al., 2013).

One of the most commonly used qualitative methods for exploring respondents' response processes —while also allowing integration with quantitative data and analyses— is cognitive interviewing. This method enables the collection of information regarding all stages of the “question-and-answer” process described by Tourangeau (2018): respondents' interpretations of items, the memories that come to their minds, the judgment strategies used when estimating answers, and the decisions made when selecting their final responses (Beatty & Willis, 2007).

Cognitive interviewing is defined as “the administration of draft survey questions while collecting additional verbal information about the survey responses” (Beatty & Willis, 2007, p.287). The cognitive interview protocol consists of several probes designed to explore how respondents constructed their answers to the target questions, collecting information about anything that “sheds light on the broader circumstances that their answers were based upon” (Beatty & Willis, 2007, p.288). Examples of such probes include: “*What were you thinking when you read this question?*” or “*What came to mind when you answered this question?*” (Beatty & Willis, 2007).

The qualitative results from cognitive interviewing can be integrated with the interviewees' quantitative responses to the items under examination to better understand the meaning of those responses from a multiperspective approach that provides a comprehensive understanding of the phenomenon. This integration can be achieved narratively, following a weaving approach (Fetters et al., 2013), in which both qualitative and quantitative results are conjointly presented, using qualitative findings to explain quantitative results, and vice versa.

The Current Study

This study aims to gather validity evidence based on response processes for the *Students' Experience of Being Bullied Scale* from the PISA 2018 Student Questionnaire scores, using a mixed-methods approach. The goal is to identify and rule out potential measurement “artifacts” or biases that could compromise the validity and fairness of comparative interpretations of bullying scores between Spanish girls and boys. The study addresses the following research questions:

1. What are bullying experiences of Spanish adolescents?
2. Do Spanish boys and girls differ in their response process to bullying items?
3. Do gender differences in bullying experiences compromise the validity and fairness of comparative scores interpretations?

To address these questions, the study follows an explanatory mixed-method design integrating quantitative and qualitative data to gain a deeper understanding of response processes to bullying item scales.

Methods

Participants

We interviewed a total of 20 students (9 girls and 11 boys), all of whom were 15 or 16 years old ($M = 15.25$ years; $SD = 0.44$), and attended compulsory secondary education (known in Spain as *Educación Secundaria Obligatoria, E.S.O.*). Participants were selected among the 15 and 16 year-old students of a public high-school in Andalusia (Spain), with the assistance of the school counselor. The final number of interviews was determined using the saturation criterion. All participants and their parents were Spanish natives, except one girl whose mother

was born in another country. Participation was voluntary and included an economic compensation (15 euros). Both students and their parents provided informed consent to participate in the study, which was approved by the Ethics in Human Research Committee of the University of Granada (3796/CEIH/2023).

Measures

Web questionnaire. Prior to the cognitive interview, participants responded to a web-based questionnaire that included background and sociodemographic questions (age, gender, and immigration status), as well as the scales ST038 and ST034 from the PISA 2018 Student Questionnaire.

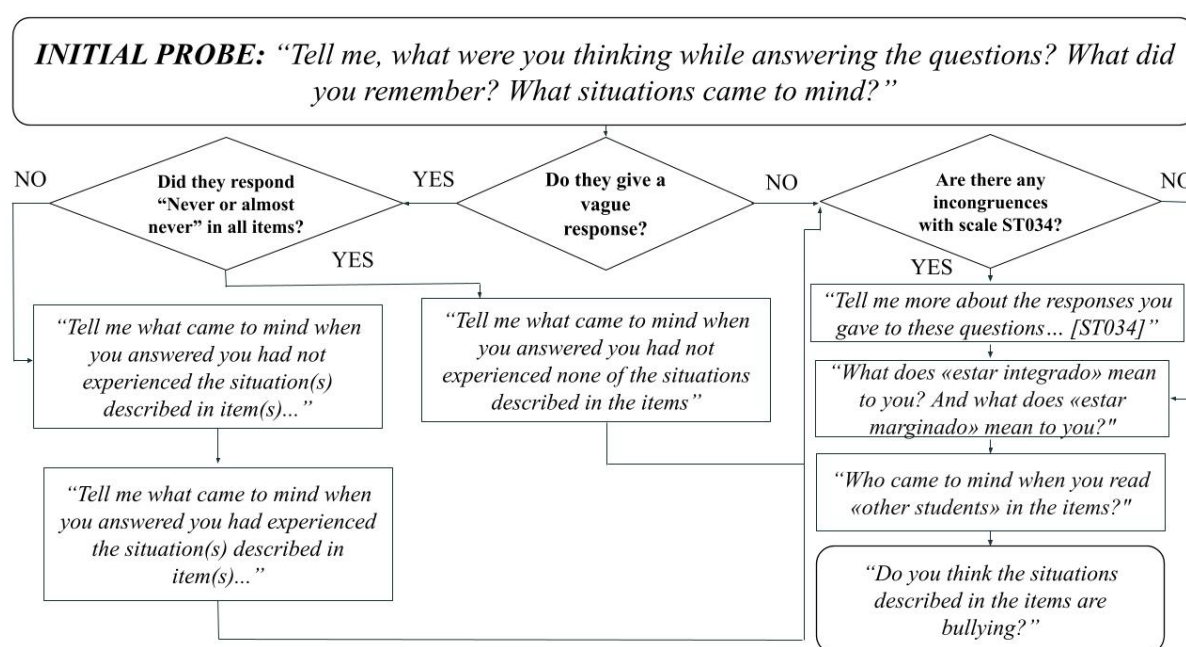
The scale ST038 is intended to measure the *students' experience of being bullied at school*. It consists of six items, each rated on a four-point Likert scale (1 – Never or almost never, 2 – A few times a year, 3 – A few times a month, 4 – Once a week or more). The items are “*Other students left me out of things on purpose*” (In Spanish: “*Otros alumnos me han excluído a propósito*”), “*Other students made fun of me*” (In Spanish: “*Otros alumnos se han reído de mí*”), “*I was threatened by other students*” (In Spanish: “*Otros alumnos me han amenazado*”), “*Other students took away or destroyed things that belonged to me*” (In Spanish: “*Otros alumnos me han quitado o han roto mis cosas*”), “*I got hit or pushed around by other students*” (In Spanish: “*Otros alumnos me han golpeado o empujado*”), and “*Other students spread nasty rumors about me*” (In Spanish: “*Otros alumnos han difundido rumores horribles sobre mí*”).

The scale ST034 was included to complement the scale ST038 by providing additional information. It is designed to measure the *students' sense of belonging to school*. The scale consists of six items, each rated on a four-point Likert scale (1 – Strongly disagree, 2 – Disagree, 3 – Agree, 4 – Strongly agree). Items are “*I feel like an outsider (or left out of things)*

at school” (In Spanish: “*Me siento marginado en el centro*”), “I make friends easily at school” (In Spanish: “*Hago amigos fácilmente en el centro*”), “I feel like I belong at school” (In Spanish: “*Me siento integrado en el centro*”), “I feel awkward and out of place in my school” (In Spanish: “*Me siento incómodo y fuera de lugar en mi centro*”), “Other students seem to like me” (In Spanish: “*Caigo bien a otros alumnos*”), and “I feel lonely at school” (In Spanish: “*Me siento solo en el centro*”).

Cognitive interview protocol. A cognitive interview protocol was developed and pre-registered on April 10th 2024. It is publicly available in both Spanish and English (https://osf.io/pb65y/?view_only=d4ab0de3c5d1476ba1cb14a56885d8dd). The protocol included probes designed to explore participants’ cognitive response processes when answering the items, such as “Tell me, what were you thinking while answering the questions? What did you remember? What situations came to mind?”. A flow diagram of the full interview process is depicted in Figure 1.

Figure 1. Flow diagram of the cognitive interview process.



Procedure

All cognitive interviews were conducted in-person at the high-school during the last weeks of April 2024. The interviews were carried out individually. Each participant first completed the scales on a computer in the presence of the interviewer. Then, following a retrospective design, they answered several probes about their response processes while answering the previous items. Each interview lasted approximately 30-45 minutes. Audio recordings were made of all interviews.

Data Analysis

First, audio recordings of the cognitive interviews were transcribed. The qualitative data was analyzed through inductive reasoning, following the procedure described by Miller et al. (2014), which is based on the principles of Grounded Theory (Glaser & Strauss, 1967). The procedure consists of five steps: 1) conducting interviews to generate interview text; 2) synthesizing the interview text to produce detailed summaries; 3) comparing summaries across participants to develop a thematic schema; 4) comparing identified themes across subgroups to construct an advanced schema; and 5) drawing conclusions to produce final study findings. This approach allows for a successive and systematic reduction of information. All qualitative analyses were conducted using Nvivo software (QSR International, 2022). Two independent researchers analyzed the data.

Results

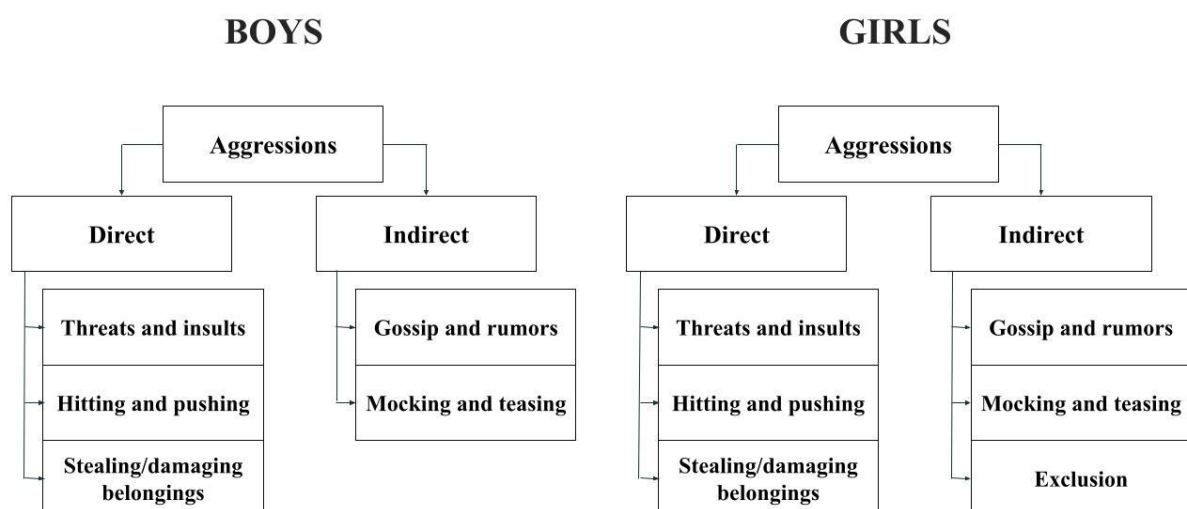
Situations and Experiences of Bullying

During the interviews, students mentioned six types of bullying-related situations: 1) experienced situations (aggressions they experienced first-hand), 2) witnessed situations

(aggressions they witnessed, without being victims), 3) not experienced situations (aggressions they explicitly stated not having received), 4) hypothetical situations (aggressions they have not received, but they consider likely to occur elsewhere), 5) situations in which the interviewee might have been the perpetrator, and 6) other situations, such as arguments, misunderstandings, etc.

Regarding experienced situations, the types of aggressions mentioned during the interviews align with those described in the ST038 items, distinguishing the two typologies: direct and indirect aggressions. When comparing girls and boys (Figure 2), both groups reported experiencing the same types of aggressions, with one exception: while girls mentioned having been excluded by their peers, boys did not report this type of aggression —neither as experienced, nor witnessed. In fact, 9 out of 11 boys explicitly declared that they had never experienced any situation of exclusion.

Figure 2. Typologies of experienced aggressions compared across gender: advanced schema.



In total, students reported 60 situations of experienced aggressions (23 reported by girls and 37 by boys). The most frequently mentioned situations among boys were threats and insults (11 situations reported by 4 boys), mocking and teasing (9 situations reported by 7 boys), and hitting and pushing (9 situations reported by 8 boys). On the other hand, the most frequently mentioned situations among girls were mocking and teasing (9 situations reported by 7 girls), gossip and rumors (5 situations reported by 4 girls), and exclusion (3 situations reported by 3 girls).

Students also described several characteristics associated with each situation they experienced in their speeches. The most frequently mentioned characteristics by both boys and girls are listed in Table 1, these represent the characteristics most commonly mentioned by each group.

It can be observed that boys mentioned more characteristics of the situations than girls. Most of these characteristics referred to the same situations of mocking and teasing and hitting and pushing (e.g., *“It’s... messing around. Sometimes, with friends, there’s some pushing. Sometimes pushing is a little bit harder, but no, never too much. We even fight sometimes, but it’s always as a joke. We’re friends, nothing serious ever happens. Everything we do is a joke and we don’t mind.”* [original version: *“Son... tonterías. A veces, con amigos, empujones, que a veces van un poquito más serios. Pero no, nunca salen más lejos de eso. A veces nos peleamos y de todo, pero siempre es de broma. Pero somos amigos, nunca pasa nada. Todo eso lo hacemos de broma y a nosotros no nos importa.”*]). Although both groups identified their friends as perpetrators of some aggressions, boys tend to experience these situations through characteristics such as “low gravity”, “reciprocal” or “as a joke”, as illustrated in the previous example. In contrast, girls interpret these experiences as acts of betrayal and tend to end the friendship afterwards (e.g., *“No, she’s not my friend anymore. Because I don’t want her to make up more stories about me, and I don’t trust her anymore.”* [original version: *“No, ya no*

es mi amiga. Porque no quiero que se invente más cosas más y ya no me fío de ella.”]). It is also worth noting that, while girls express their psychological distress associated with the aggressions they experienced, boys are more likely to normalize these situations and explicitly state that they do not feel psychological distress.

Table 1. Most frequent characteristics of experienced situations across gender.

Characteristics of situations	<i>f</i>
Boys	
Friends as perpetrators	10/11
Low gravity	10/11
Reciprocal	9/11
Lack of psychological distress	9/11
“As a joke”	9/11
High frequency	7/11
Normalization	7/11
Lack of repercussion	7/11
Girls	
Psychological distress	7/9
Friends as perpetrators	5/9
Not the only victim	5/9

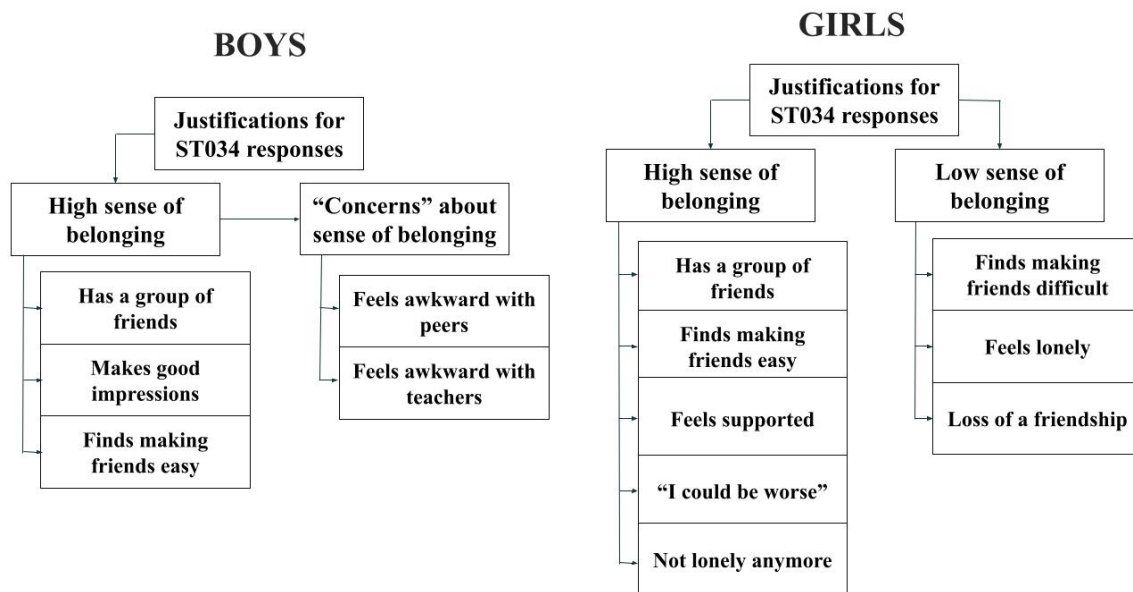
Regarding witnessed situations, similar patterns can be found across genders. Boys tend to witness more situations of threats and insults (8 situations witnessed by 6 boys) and hitting and pushing (7 situations witnessed by 3 boys), while girls are more likely to witness situations involving gossip and rumors (4 situations witnessed by 3 girls) and exclusion (4 situations witnessed by 4 girls).

Sense of Belonging to School and Bullying

Almost all participants reported a strong sense of belonging to their school, regardless of the number of aggressions they had experienced. Only one reported a low sense of belonging, and she was the only participant who was suffering from exclusion at the time of the interview. Two boys, despite reporting a high sense of belonging, expressed some concerns about feeling uncomfortable in certain situations —specifically with some peers (not their friends), and with teachers.

The reasons students gave for having a high sense of belonging to school vary by gender. While both boys and girls mentioned having a group of friends and finding it easy to make friends as the main reasons, girls also highlighted that their current sense of belonging is linked to the end of previous experience of exclusion (e.g., *“Right now, I feel that I can talk about myself, that they respect me, and that they include me in their activities.”* [original version: *“Siento que ahora mismo sí puedo contar mis cosas y que sí me respetan y me integran en las actividades.”*]). All justifications students provided for their answers to the ST034 scale are depicted in Figure 3.

Figure 3. Justifications for ST034 responses compared across gender: advanced schema.



Cognitive Representation of Certain Terms and Response Processes

Throughout the interview, several terms from the scales used were explored to understand how students interpreted them. The first term examined was “integrado”, which appeared in the item “*I feel like I belong at school*” (In Spanish: “*Me siento integrado en el centro*”), from the ST034 scale. Boys and girls seem to have understood the term in a similar way. Both boys and girls believed the term means inclusion, having positive interactions and relationships with peers, feeling well at school, having friends, and maintaining good relationships with teachers. Only girls also mentioned being popular, knowing the school rules, and not having any problems. In contrast, only boys mentioned being involved in extracurricular activities, feeling well with friends, performing well academically, feeling well outside of school, and being treated well by others.

The second term under examination was “marginado”, which appeared in the item “*I feel like an outsider (or left out of things) at school*” (In Spanish: “*Me siento marginado en el*

centro”), from the ST034 scale. Most students (both boys and girls) interpreted the term as referring to isolation, being a victim of aggressions by peers, not being liked by others, feeling unwell at school, and lacking interaction with peers. Only boys additionally mentioned feeling inferior and frustrated. However, two girls and two boys had problems understanding the term, even admitting they had never heard it before.

The cognitive representation of the expression “other students” was also examined, as all items from the ST038 scale start with this expression. When reading it, both girls and boys primarily thought of the perpetrators of the aggressions they recalled. However, some students (3 boys and 2 girls) also thought of the victims of those aggressions. The perpetrators they thought of were diverse. Both boys and girls mentioned friends as perpetrators, as well as non-friends. Among these, both girls and boys referred to classmates, students from other classes, and older students. Only girls also mentioned students from other schools and ex-friends.

Regarding students’ response processes, some girls (6 out of 9) and boys (8 out of 11) did not report certain aggressions that they later communicated during the interview. The majority (5 out of 6 girls and 6 out of 8 boys) explained they had not reported these situations because they had not considered them severe enough. There are gender differences in the reasons behind this perception. Girls gave a variety of explanations (e.g., thinking only physical aggressions are severe, thinking rumors and gossip are just “silly things adolescents do”). In contrast, most boys (5 out of 6) thought the situations were not severe because they involved a reciprocal aggression between them and their friends, which they perceived as joking. On the other hand, three boys reported on the scale situations that they did not consider to be aggressions —again, referring to reciprocal aggression among friends. This phenomenon did not occur among girls.

Validity Evidence Based on Consequences: Conceptualization of Bullying

In order to test whether participants thought of “bullying” or not when reading ST038 items, the term “bullying” was deliberately avoided throughout the interview until the final question, in which we asked if they believed the ST038 items represented bullying behaviors. The possibility that students may have thought of bullying when answering ST038 items could have influenced their responses, particularly in terms of the “underreporting” of situations not considered severe enough—an issue discussed in the previous section. Five students (2 girls and 3 boys) spontaneously mentioned the term “bullying” before the final question, possibly indicating that ST038 items prompted them to think about bullying. One girl described a situation in which another girl accused her of being a bully, two boys mentioned situations of witnessing bullying, and one girl and one boy clarified that their experiences with aggression were not bullying.

Regarding students’ responses to the final interview question, all of them thought that the behaviors represented in the ST038 items were bullying. Only a girl made an exception in the case of rumors and gossip, explaining “*rumors are often lies. Whatever they say about you, you know who you are, you know if you did or not what they say. I think you should have a clear conscience and don’t mind what they say.*” [original version: “*Porque los rumores suelen ser mentira. Entonces, que lo que digan de ti, tú sabes realmente cómo eres o si lo has hecho o no lo has hecho. Entonces yo creo que tú tienes que tener la conciencia tranquila de que te dé igual lo que digan.*”].

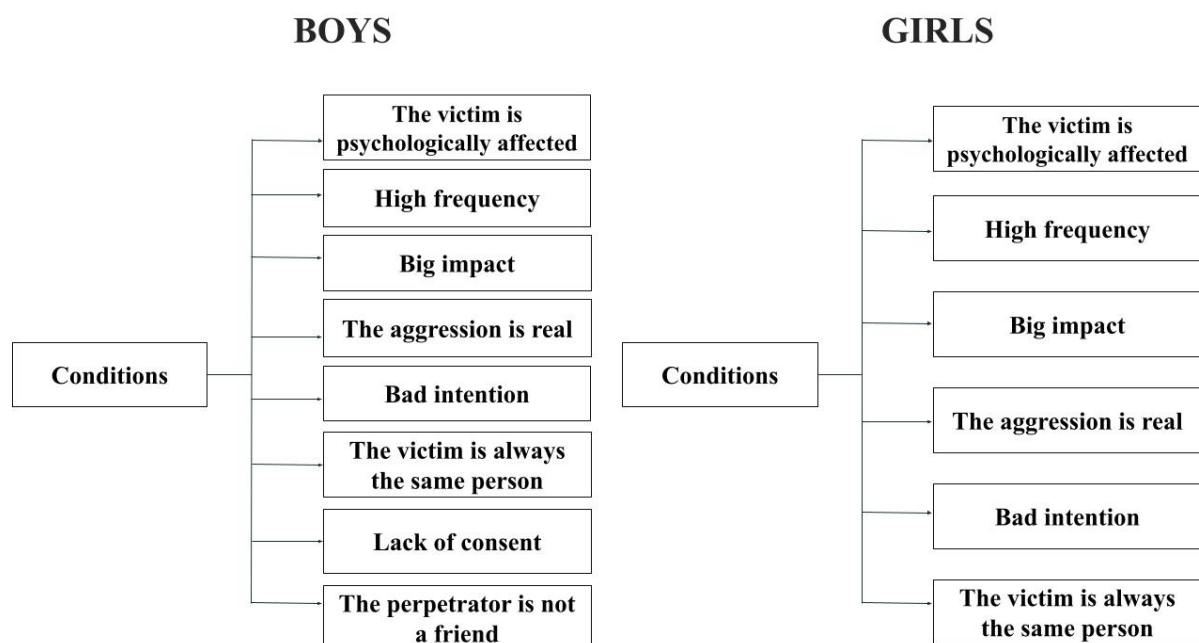
The main reason both boys and girls gave for considering the ST038 as examples of bullying was that victims of such behaviors are psychologically affected by such behaviors. Additionally, girls also mentioned that those were unjustified, unethical, and carried out with

bad intentions by people they described as “bad”, and that such behaviors are not perpetrated by people who love you.

Most girls (6 out of 9) and one boy thought that all behaviors described in the ST038 items are bullying under any circumstance. In contrast, the remaining students (10 out of 11 boys and 3 out of 9 girls) believed that those behaviors are bullying as long as certain conditions are met. Those conditions are presented in Figure 4.

It is worth noting that only boys mentioned the condition that the perpetrator is not a friend, implying that if friends perpetrate any of the ST038 behaviors, they do not consider them to be bullying (e.g., *“If it’s between friends, no. You control the levels of how they say things to you or what is happening at the moment.”*; [original version: *“Si es entre amigos, no. Tú ya controlas los niveles de como te lo dicen o lo que está pasando en ese momento.”*])).

Figure 4. Conditions to consider ST038 behaviors as bullying compared across gender: advanced schema.



Lastly, some students felt personally addressed by the final question of the interview. Two boys acknowledged having been bullied, while two girls and three boys felt the need to clarify that the aggressions they experienced were not bullying.

Discussion

In this study, cognitive interviewing enabled us to delve into the differences in response processes between boys and girls when answering bullying victimization items. While the understanding of potentially complex or ambiguous terms was found to be similar across genders, the fact that girls and boys have markedly different experiences of bullying, as previous literature suggested (e.g., Hanci & Akça, 2020; Navarro-González et al., 2025; Rosen & Nofziger, 2021; Shu et al., 2023; Lee et al., 2022; OECD, 2019; Zhao et al., 2022), was reinforced. Boys are more likely to experience direct forms of aggression, whereas girls tend to be subjected to indirect forms. Even when girls frequently report direct experiences of mocking and teasing, these are qualitatively different from those reported by boys: for girls, such experiences are often psychologically distressing, while boys tend to perceive them as minor incidents occurring in a “joking” context. A similar pattern was found in the case of witnessed aggressions.

Differential gender socialization may underline these differences (Eagly, 1987; Eagly & Wood, 2012), as adolescents tend to conform to traditional gender roles and stereotypes to avoid peer rejection—something more likely to occur when gender expression does not align with the “sex category” (Messerschmidt, 2009).

On the one hand, girls are often socialized to be kind and friendly toward others (e.g., Diekmann & Eagly, 2008), which aligns with their tendency to engage in indirect aggression

that avoids direct confrontation or physical threats, behaviors that would contradict a friendly and kind demeanor. On the other hand, boys are often socialized under the roles of hegemonic masculinity, which normalize instrumental violence and aggression (e.g., Dieckman & Eagly, 2008). This is consistent with the pattern of direct bullying often observed among boys, as well as with their tendency to downplay the seriousness, impact, and psychological distress associated with such experiences. Boys also report that these situations are often reciprocal within their peers group, possibly serving as a form of bonding —friendly teasing—, as some authors suggest (Mishna et al. 2008; Wei & Jonson-Reid, 2011). This may help explain the “paradox” in which boys experience more aggression and bullying than girls, yet report a stronger sense of belonging to school (OECD, 2019).

These different experiences of bullying victimization appear to influence how girls and boys understand what bullying means, with their “cognitive representations” of the phenomenon differing significantly. Girls are more inclined to identify an act of aggression as bullying, and tend to interpret almost any aggressive behavior as bullying. In contrast, boys are more likely to minimize and normalize aggressions, only labeling it as bullying when certain criteria are met, such as high frequency or when the perpetrator and victim are not friends. Otherwise, the aggression is often perceived as friendly teasing.

Both boys and girls believe that for an act to be considered bullying, the victim must be psychologically affected. This belief may help explain why girls are more likely than boys to classify aggressive situations as bullying. Due to gender socialization (e.g., Dieckman & Eagly, 2008), girls are generally more comfortable expressing their emotions and acknowledging the psychological impact of the aggressions they experience. This openness may lead them to more readily identify such experiences as bullying. Boys, on the other hand, are often socialized to hide and mask their emotions (e.g., Dieckman & Eagly, 2008), and may claim not to be

psychologically affected by the aggressions, making it more difficult for them to recognize bullying in these situations.

Therefore, there are two distinct “realities” and “cognitive representations” that shape how boys and girls respond to bullying victimization items, highlighting the significant differences in their response processes:

1. Girls tend to experience indirect aggressions that have a psychological impact on them. Generally, these types of experiences make it easier for girls to identify bullying behaviors, being more accurate than boys in doing so. However, for some girls, experiencing indirect aggressions may increase the risk of underreporting such aggressions, as they may believe that only physical and direct aggressions are “real” bullying.

2. Boys tend to experience direct aggressions that, on the surface, do not appear to psychologically affect them and are often perceived as normal, as a joke, or falling under the category of friendly teasing. Depending on the evaluators’ perspective and the intended purpose of the assessment, different reporting issues may arise: 1) some evaluators may assume that all forms of aggression are problematic, regardless of how boys perceive or report them. Their premise is that boys are indeed psychologically affected by these aggressions, even if they do not openly acknowledge it. These evaluators would therefore aim to detect every instance of aggressive behavior that boys experience or perpetrate. In this case, the main challenge these evaluators would be faced is underreporting, as boys who do not see themselves as bullying victims may fail to report certain aggressions, considering them not “serious enough” or simply “friendly teasing”; 2) other evaluators may believe that only aggressions accompanied by psychological distress are problematic. Their premise is that if boys report not being

psychologically affected, then they truly are not. These evaluators would focus on identifying only severe aggressions experienced with psychological affectation, distinguishing them from those perceived as friendly teasing. In this case, the issue would be overreporting of “false positives”, as some boys may report aggressions that they do not consider bullying, but rather friendly teasing.

This study has highlighted the importance of obtaining validity evidence from response processes through a mixed-methods approach, which is necessary in order to fully understand boys’ and girls’ responses and scores on tests and questionnaires, and to develop fair and inclusive assessments, in terms of diversity. It has also shown that there are several idiosyncrasies influencing girls’ and boys’ responses that evaluators must consider if they aim to make valid interpretations and fair comparisons between the two groups. The traditional finding that “*boys experience more bullying than girls*” is not just as simple as it seems. Understanding the issues of underreporting or overreporting —depending on the gender socialization— is key to unpacking this finding and truly grasping its meaning.

Limitations and Future Research

After discussing the impact of gender socialization on boys’ and girls’ responses, a natural question arises: are boys so accustomed to violence that they no longer perceive it as problematic, or are they simply raised to be “tough” and therefore act *as if* such aggressions do not psychologically affect them? Unfortunately, a limitation of this study is that the cognitive interviews we conducted did not yield enough information to answer this question. Detecting “lies” or “repressed emotions” is an inherently difficult task, but addressing this question could be a valuable challenge for future researchers seeking to further unravel the phenomenon of bullying.

Other limitations in this study are related to the sample. We only interviewed 20 students from the same country and the same public high school. It would be interesting for future researchers to broaden the sample to enable comparisons between public and private schools, and to replicate these findings in other countries and cultures. Web probing—the implementation of cognitive interviewing techniques in online questionnaires— (Behr et al., 2017), could be the perfect tool for this purpose, as it allows researchers to reach larger sample sizes and more geographically diverse populations with ease.

Implications for Practice

The design and development of interventions to prevent and address bullying by school counselors, principals, and policymakers is often based on data from bullying scales. To give meaning to the results of these scales, it is essential to understand how boys and girls interpret key concepts in bullying item stems and how they construct their responses.

Our study illustrates how cognitive interviews can help in understanding what bullying scale items mean to boys and girls, the differences in interpretation between gender groups, and what they consider to be bullying experiences. Practitioners and counselors in schools can complement quantitative data from bullying scales with qualitative evidence from cognitive interviews to gain a more comprehensive understanding of how students at their schools experience and perceive bullying. Armed with a more comprehensive understanding, they can design more effective interventions that are tailored to the real context of their schools. For example, cognitive interviews can provide evidence on whether the phenomenon of “friendly teasing” is present and whether it is experienced differently by boys and girls. Moreover, understanding the impact that value judgments associated with the term “bullying” may have—and whether such judgments are present in their schools— is essential for preparing and interpreting responses to bullying scales.

Declarations

Conflicts of Interest

None.

Author Contributions

Following the CRediT roles taxonomy (available in <https://credit.niso.org/>), contributions are: M-CN-G: conceptualization, investigation, formal analysis, visualization, writing – original draft, writing – review & editing; J-LP: conceptualization, supervision, writing – review & editing. All authors read and approved the final manuscript.

Human Rights

This research was approved by the Ethics in Human Research Committee of the University of Granada (3796/CEIH/2023). Both students and their parents provided informed consent to participate in the study. Privacy and confidentiality was preserved through anonymization and restricted data storage, according to current legislation (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos y Garantías de Derechos Digitales).

Funding

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

References

- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association [AERA]. Retrieved from <https://www.apa.org/science/programs/testing/standards>
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287-311. <https://doi.org/10.1093/poq/nfm006>
- Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). *Web probing—implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions*. GESIS Survey Guidelines. https://doi.org/10.15465/gesis-sg_en_023
- 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*, 51(1), 1-34. <https://doi.org/10.1177/0049124119852390>
- Castillo-Díaz, M., & Padilla, J. L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, 114, 963-975. <https://doi.org/10.1007/s11205-012-0184-8>

- Cheng, W., Guan, Y., & Luo, Y. (2020). Campus bullying and sense of belonging to school and campus security among junior high school students: a cross-sectional survey. *Chinese Journal of Public Health*, 36(6), 889-894.
<https://doi.org/10.11847/zgggws1124168>
- Chung, J. Y., & Lee, S. (2020). Are bully-victims homogeneous? Latent Class Analysis on school bullying. *Children and Youth Services Review*, 112, 104922.
<https://doi.org/10.1016/j.childyouth.2020.104922>
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. 2nd edition. Sage Publications. Retrieved from
<https://psycnet.apa.org/record/2006-11884-000>
- Díaz-Aguado, M. J., Martínez-Arias, R., Falcón, L., & Alvariño, M. (2023). *Acoso escolar y ciberacoso en España en la infancia y en la adolescencia*. Fundación ColaCao y Universidad Complutense de Madrid. Retrieved from
<https://fundacioncolacao.org/investigacion>
- Diekmann, A. B., & Eagly, A. H. (2008). Of men, women, and motivation. In J. Y. Shah, & W. L. Gardner (Eds.), *Handbook of motivation science*. Guilford Press. Retrieved from
<https://psycnet.apa.org/record/2008-00543-000>
- Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Psychology Press. <https://doi.org/10.4324/9780203781906>
- Eagly, A. H., & Wood, W. (2012). Social role theory. In P. van Lange, A. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories in social psychology*. Sage.
<https://doi.org/10.4135/9781446249222.n49>

- Fetters, M. D., Curry, L.A., & Cresswell, J. W. (2013). Achieving integration in mixed methods designs – Principles and practices. *Health Services Research*, 48(6), 2134-2156. <https://doi.org/10.1111/1475-6773.12117>
- Gladden, M. R., Vivolo-Kantor, A. M., Hamburger, M. E., & Lumpkin, C. D. (2014). *Bullying surveillance among youths: uniform definitions for public health and recommended data elements. Version 1.0*. Centers for Disease Control and Prevention. Retrieved from <https://stacks.cdc.gov/view/cdc/21596>
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. Aldine. <https://doi.org/10.4324/9780203793206>
- Hanzi, I. H., & Akça, S. Ö. (2020). School bullying. *Journal of Education and Future*, 18, 113-121. <https://doi.org/10.30786/jef.691717>
- He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2). <https://doi.org/10.9707/2307-0919.1111>
- Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., Wolfe, M., & Reid, G. (2015). Bullying and suicidal ideation and behaviors: a meta-analysis. *Pediatrics*, 135(2), 496-509. <https://doi.org/10.1542/peds.2014-1864>
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112-133. <https://doi.org/10.1177/1558689806298224>
- Lee, J. (2021). Pathways from childhood bullying victimization to young adult depressive and anxiety symptoms. *Child Psychiatry & Human Development*, 52, 129-140. <https://doi.org/10.1007/s10578-020-00997-4>

- Lee, J., Roh, B. R., & Yang K. E. (2022). Exploring the association between social support and patterns of bullying victimization among school-aged adolescents. *Children and Youth Services Review*, 136, 106418. <https://doi.org/10.1016/j.childyouth.2022.106418>
- Miller, K., Willson, S., Chepp, V., & Ryan, J. M. (2014). Analysis. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive Interviewing Methodology*. Wiley. <https://doi.org/10.1002/9781118838860>
- Messerschmidt, J. W. (2009). “Doing gender”: The impact and future of a salient sociological concept. *Gender and Society*, 23(1), 85-88. <https://www.doi.org/10.1177/0891243208326253>
- Mishna, F., Wiener, J., & Pepler, D. (2008). Some of my best friends—experiences of bullying within friendships. *School Psychology International*, 29(5), 549-573. <https://doi.org/10.1177/0143034308099201>
- Navarro-González, M. C., Padilla, J. L., & Postigo, A. (2025). Testing comparability of gender bullying victimization assessments through a latent class approach. *Journal of Interpersonal Violence*, 00(0) [online first], 1-25. <https://doi.org/10.1177/08862605251325926>
- OECD (2019). *PISA 2018 results (volume III): what school life means for students’ lives*. OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/education/pisa-2018-results-volume-iii_acd78851-en
- OECD (2023). *PISA 2022 results (volume II): learning during - and from - disruption*. OECD Publishing. Retrieved from https://www.oecd.org/en/publications/pisa-2022-results-volume-ii_a97db61c-en/full-report.html

- Olweus, D. (1978). *Aggression in the schools: bullies and whipping boys*. Hemisphere.
Retrieved from <https://psycnet.apa.org/record/1979-32242-000>
- Olweus, D. (2013). School bullying: development and some important challenges. *Annual Review of Clinical Psychology*, 9, 751-780. <https://doi.org/10.1146/annurev-clinpsy-050212-185516>
- QSR International (2022). NVivo (Version 13) [Software]. Retrieved from <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>
- Rosen, N. L., & Nofziger, S. (2019). Boys, bullying, and gender roles: how hegemonic masculinity shapes bullying behavior. *Gender Issues*, 36, 295-318. <https://doi.org/10.1007/s12147-018-9226-0>
- Shu, Y., Ota, M., & Kato, H. (2023). The diverse nature of school bullying and its relation to depression and the efficacy of coping strategies in Japan. *International Journal of Bullying Prevention*, 1-15. <https://doi.org/10.1007/s42380-023-00181-x>
- Snodgrass, M. A., Smith, S. L., & Gregus, S. (2024). Understanding children and adolescents' experiences being bullied: a mixed-methods study. *Trends in Psychology*, 1-22. <https://doi.org/10.1007/s43076-024-00385-0>
- Swearer, S. M., & Hymel, S. (2015). Understanding the psychology of bullying. Moving toward a social-ecological diathesis-stress model. *American Psychologist*, 70(4), 344-353. <https://doi.org/10.1037/a0038929>
- Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2), 169-181. <https://doi.org/10.1108/QAE-06-2017-0034>

UNESCO (2019). *Behind the numbers: ending school violence and bullying*. UNESCO.

<https://doi.org/10.54675/TRVR4270>

Wei, H., & Jonson-Reid, M. (2011). Friends can hurt you: examining the coexistence of friendship and bullying among early adolescents. *School Psychology International*, 32(3), 244-262. <https://doi.org/10.1177/0143034311402310>

Zhao, D., Wang, D., He, Z., Yuan, S., Zhu, D., & Liu, H. (2022). Victim profiles and the protective role of school anti-bullying norms: a study of Chinese adolescents. *Current Psychology*. <https://doi.org/10.1007/s12144-022-04204-6>

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. <https://doi.org/10.1080/15434303.2014.972559>

CAPÍTULO VII: DISCUSIÓN GENERAL

1. DISCUSIÓN GENERAL E INTEGRACIÓN DE HALLAZGOS

Asegurar la validez y la justicia de las interpretaciones que se basan en la comparación de las mediciones obtenidas por diferentes grupos demográficos, lingüísticos y/o culturales es imprescindible para tomar decisiones importantes que no beneficien o perjudiquen a un grupo de personas en particular (American Educational Research Association [AERA] et al., 2014). Especialmente crucial es esta cuestión cuando de las interpretaciones comparativas de las puntuaciones de un test dependen el desarrollo de estrategias para detectar y/o intervenir en problemáticas con importantes y devastadoras consecuencias, como la victimización de *bullying*. Para cumplir dicho cometido, en la presente tesis doctoral se propone una estrategia “mixta” de validación que incluye el estudio tanto de la invarianza métrica como de los procesos de respuesta a los ítems.

Entender la falta de invarianza métrica como un problema de validez (Zumbo y Padilla, 2019), abre el camino a utilizar el estudio de la invarianza como forma de aportar evidencias de validez, no solo basadas en la estructura interna —como tradicionalmente se propone (AERA et al., 2014)—, sino también basadas en los procesos de respuesta. Esto es así debido a que, si se indagan sus causas, la falta de invarianza puede indicar la presencia de fuentes de sesgo que han influido diferencialmente en los procesos de respuesta de personas pertenecientes a diferentes grupos sociodemográficos, lingüísticos y/o culturales (Zumbo, 2007; Zumbo et al., 2015). Así, indagar en las causas de la falta de invarianza —o de la presencia de funcionamiento diferencial de los ítems (DIF)—, se convierte en la tarea de

indagar en los procesos de respuesta de las personas que contestan a los ítems (e.g., Padilla y Benítez, 2014; Sireci y Benítez, 2023). De esta manera, se harían “visibles” diversos sesgos que, de existir, pondrían en riesgo la justicia de las mediciones al beneficiar o perjudicar a un grupo de personas concreto, tales como el sesgo androcéntrico (e.g., Bonilla-Campos, 2010; García-Calvente et al., 2010). Además, explorar las causas de la falta de invarianza o de la presencia de DIF significa también conocer mejor el fenómeno de estudio y el contexto en el que viven las personas que contestan a los instrumentos de evaluación, visibilizando las diferentes realidades que existen y las “diversas formas de ser humano”, que moldean la manera que tienen las personas de interactuar con los instrumentos de evaluación (Zumbo, 2021; Zumbo, 2023; Zumbo, en prensa).

A lo largo de la tesis doctoral se ha evidenciado que el estudio integrado de la invarianza métrica y de los procesos de respuesta en una estrategia mixta de validación es útil para asegurar la validez y la justicia de las mediciones, como se ha mostrado al desarrollar y aplicar dicha estrategia en el contexto de la evaluación de victimización de *bullying* en chicas y chicos adolescentes, cumpliéndose así el objetivo general de la tesis.

El primer estudio (Estudio I) evaluó el estado de conocimientos sobre los resultados, las condiciones de aplicación, y las ventajas e inconvenientes de los métodos cuantitativos y cualitativos para el análisis de la invarianza o equivalencia métrica, a través de una revisión sistemática mixta de la literatura científica. Así, con esta revisión sistemática mixta, se consiguió realizar un compendio-resumen de las formas en las que se ha abordado el estudio de la invarianza métrica y el DIF relativos al género en el contexto de las medidas educativas, cómo se han aplicado y cuáles son sus pros y sus contras. El hecho de haber realizado una revisión sistemática mixta posibilitó la inclusión de estudios cuantitativos, cualitativos y mixtos, así como una extracción de datos de los mismos tanto cuantitativa como cualitativa. Esto permitió obtener un corpus de conocimiento más amplio y completo sobre las técnicas,

con datos sobre su frecuencia de aplicación, las características principales de las pruebas sobre las que se aplican normalmente (e.g., formato de respuesta, número de ítems, modo de administración...), el tamaño de la muestra más común, los pros y contras de su aplicación, en qué medida es frecuente que tras utilizar cada una de las técnicas se exploren las causas de la falta de invarianza o de la presencia de DIF, y qué procedimientos se utilizan en caso de que se exploren.

De esta manera, se encontró que los métodos más utilizados para estudiar la invarianza métrica o el DIF fueron el análisis factorial confirmatorio (CFA) multigrupo y el Mantel-Haenszel (MH), manteniéndose en el “podio” de ser los métodos tradicionales para abordar esta cuestión. Sin embargo, métodos como los modelos de clases latentes (LCA) o SIBTEST son bastante prometedores, a pesar de que no fueron muchos los estudios que los utilizaron, ya que permiten fácilmente el posterior estudio de las causas de la falta de invarianza o de DIF, así como el abordaje de la influencia de diversas variables, permitiendo su adaptación a datos multinivel. Por otro lado, el método más utilizado para indagar sobre las causas de la falta de invarianza o de la presencia de DIF fue el análisis del contenido de los ítems, aunque en la mayoría de los casos dicho análisis fue meramente descriptivo y superficial, sin ahondar en explicaciones complejas y profundas para las causas. También se encontraron otros métodos como la entrevista, los grupos focales, el juicio de expertos o el método Delphi, aunque usados en mucha menor frecuencia, probablemente debido a que la implementación de estos métodos es más costosa que el mero análisis de contenido.

En conclusión, se vio que la tendencia general era realizar los análisis más populares para estudiar la invarianza métrica y el DIF, probablemente sin dedicar mucho tiempo a estudiar de forma rigurosa qué técnica hubiera sido mejor teniendo en cuenta las características y objetivos de los estudios. También, la mayoría de los estudios no buscaron las causas de la falta de invarianza o de la presencia de DIF, y se dedicaron únicamente a eliminar los ítems

que resultaron problemáticos en los análisis, sin preocuparse por indagar más sobre los factores responsables, como se propone hacer en la “tercera generación del DIF” (Zumbo, 2007), y mucho menos conceptualizando el DIF de acuerdo a la “cuarta generación” (Zumbo, en prensa). Además, la mayoría de los estudios que sí lo hicieron, aplicaron técnicas que generaron explicaciones superficiales sobre estas causas. Por un lado, esta aparente “falta de interés” por la rigurosidad de elegir la técnica más apropiada de detección de DIF o falta de invarianza, y para explorar sus causas podría ser fruto de la desconexión que existe a menudo entre la teoría y la práctica (e.g., Padilla et al., 2006), llegando a ser un problema la limitada “transferencia” que parece haber a la hora de implementar los avances del campo de la Metodología en otras áreas de la Psicología más aplicadas. Otro problema que es posible que esté influyendo en esta cuestión es lo costoso de la implementación de algunas técnicas, ya que muchas de ellas requieren de formación específica y pueden dilatar significativamente el tiempo que requeriría llevar a cabo el estudio, como por ejemplo, la entrevista cognitiva —que, de hecho, ningún estudio de la revisión sistemática utilizó—. Solucionar estos problemas resulta un reto para la investigación futura en el campo, que deberá tratar, de manera activa, de ser más accesible para otras áreas. Estos hallazgos ponen de manifiesto cómo a menudo el abordaje del DIF se centra únicamente en la parte más estadística o más relacionada con las técnicas de detección, olvidando el propósito de la detección del DIF en sí, que es asegurar la validez y la justicia de las mediciones (Zumbo, en prensa).

Los hallazgos de esta revisión sistemática sirvieron para guiar el diseño de los estudios posteriores (Estudios II y III), para los que fue necesario elegir un método cuantitativo para estudiar la invarianza métrica y un método cualitativo para explorar las causas de la falta de dicha invarianza. Por un lado, se tomó la decisión de utilizar como método estadístico el LCA para el Estudio II, puesto que, como se ha comentado anteriormente, es un método que permite fácilmente indagar en las causas de la falta de invarianza con el estudio de los patrones de

respuesta, e incluso a través de capturar diferencias de carácter cualitativo entre las diferentes clases, al permitir “dar sentido” a las mismas en función de la probabilidad de contestar de una manera determinada a cada uno de los ítems, teniendo en cuenta el propio contenido de los mismos. Además, permite traer a la ecuación variables que a priori pueden no ser manifiestas, cuestión importante a la hora de abordar las causas de la invarianza de manera exploratoria, o al estudiar un problema complejo en el que podrían no estar “viéndose” todas las variables que influyen en las diferencias entre ciertos grupos de personas. Es, además, una de las técnicas recomendadas por Bruno D. Zumbo et al. (2015). Por otro lado, se optó por utilizar entrevistas cognitivas para el Estudio III, ya que, a pesar de que ningún estudio de la revisión sistemática las utilizara, sí se encontró que la entrevista fue el único método para explorar las causas de la invarianza sobre el que se resaltó como pro su utilidad para indagar en los “procesos cognitivos” de los y las participantes. Además, numerosos estudios avalan la entrevista cognitiva como estrategia idónea para obtener evidencias de validez sobre los procesos de respuesta (e.g., Beatty y Willis, 2007; Benítez et al., 2019; Benítez y Padilla, 2014; Padilla y Benítez, 2014; Sireci y Benítez, 2023).

De esta manera, en los Estudios II y III se llevó a cabo la estrategia mixta de validación, mediante la cual se examinaron empíricamente las condiciones de aplicación tanto de los métodos cuantitativos como de los cualitativos usados para estudiar la invarianza métrica y los procesos de respuesta. Esta estrategia mixta de validación se desarrolló a través de un diseño secuencial explicativo (Fetters et al., 2013), cuya integración se ha llevado a cabo en distintos niveles, tal y como se detalló en el Capítulo III (Metodología).

Así, en primer lugar (Estudio II) se analizó la información cuantitativa relativa a la invarianza métrica de las respuestas de chicos y chicas a la escala de experiencias de *bullying* (ST038) del Cuestionario del Estudiante del Programa para la Evaluación Internacional de

Estudiantes (*Programme for International Student Assessment* [PISA]) 2018, a través de un LCA. Se encontraron cuatro perfiles o clases latentes:

- ❖ **Bullied (All Types):** personas con alta probabilidad de responder afirmativamente en todos los ítems.
- ❖ **Not Bullied:** personas con baja probabilidad de responder afirmativamente en todos los ítems.
- ❖ **Relational Bullying:** personas con alta probabilidad de responder afirmativamente los ítems representativos de *bullying* indirecto o relacional, pero baja probabilidad de responder afirmativamente a los demás ítems.
- ❖ **Potential Friendly Teasing:** personas con probabilidad media de responder afirmativamente a casi todos los ítems, excepto a los ítems referidos a las amenazas y a la exclusión social, para los cuales presentaron una probabilidad muy baja de responder afirmativamente.

Al realizar un LCA multigrupo, se constató que no se mantenía la invarianza métrica entre chicas y chicos a través de las clases. Mientras que en ambos grupos sí parecía mantenerse la estructura de cuatro clases, las probabilidades de pertenecer a cada una no eran iguales para chicos que para chicas. Para las chicas, era 2 veces más probable que para los chicos pertenecer a la clase *Relational Bullying*, mientras que para éstos era 2.4 veces más probable que para ellas pertenecer a la clase *Potential Friendly Teasing* y 2.7 veces más probable pertenecer a la clase *Bullied (All Types)*. Estos resultados pusieron de manifiesto las diferencias que existen entre chicos y chicas en sus experiencias de *bullying*, y en su mayoría fueron consistentes con la literatura previa (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Hanzi y Akça, 2020; Lee et al., 2022; Nylund et al., 2007; Olweus, 1994; Rivers y Smith, 1994; Rosen y

Nofziger, 2019; Shu et al., 2023; Wang et al., 2010; Zhao et al., 2022). A pesar de eso, la clase *Potential Friendly Teasing* suscitó más preguntas que respuestas: ¿por qué hay personas — especialmente chicos— que, a pesar de ser víctimas de ciertos comportamientos agresivos no se sienten excluidas?, ¿qué realidad está representando esta clase?

Con estos hallazgos en mente, se desarrolló el protocolo de entrevista cognitiva para el siguiente estudio (Estudio III), en el que se llevaría a cabo la aplicación del método cualitativo que complementaría y ayudaría a explicar los resultados cuantitativos del Estudio II. Por ejemplo, para explorar mejor la naturaleza de la clase *Potential Friendly Teasing*, se decidió incluir, además de la escala de experiencias de *bullying* (ST038) del Cuestionario del Estudiante del PISA 2018, la escala sobre sentido de pertenencia al centro escolar (ST034), presente también en el mismo cuestionario. Así, se exploraron de una manera más completa las experiencias de exclusión de las y los adolescentes. Por tanto, se alcanzó una integración a nivel del diseño (Fetters et al., 2013), puesto que, a través del diseño mixto secuencial explicativo, los hallazgos de la fase cuantitativa informaron la conceptualización de la fase cualitativa y esta última complementó y amplió los resultados de la primera; y también se alcanzó una integración a nivel del método (Fetters et al., 2013), puesto que el protocolo de la entrevista cognitiva se desarrolló a partir de los hallazgos de la fase cuantitativa, influyendo los resultados de la primera recolección de datos en determinar la información a incluir en la segunda recolección de datos —*building*— (Fetters et al., 2013).

De esta manera, en el Estudio III se indagaron los procesos de respuesta de chicos y chicas a las escalas de experiencias de *bullying* (ST038) y de sentido de pertenencia al centro escolar (ST034) del Cuestionario del Estudiante del PISA 2018, a través de entrevistas cognitivas. Se encontró que, en efecto, las experiencias de *bullying* de chicas y chicos eran diferentes, tal y como se había empezado a atisbar en el Estudio II cuando no se encontró invarianza métrica entre chicos y chicas a través de las clases.

Los chicos, por un lado, mencionaron más agresiones de tipo directo (e.g., amenazas, insultos, golpes, empujones), y ninguno mencionó haber sido víctima de exclusión, mientras que las chicas mencionaron más agresiones de tipo indirecto (e.g., difusión de rumores, exclusión), siguiendo el patrón encontrado en la literatura previa (e.g., Björkqvist et al., 1992; Carrera-Fernández et al., 2013; Rivers y Smith, 1994; Rosen y Nofziger, 2019; Shu et al., 2023). Ambos grupos también mencionaron haber sufrido en numerosas ocasiones situaciones de risas y burlas. No obstante, se comprobó a través de las entrevistas que chicos y chicas no interpretan ni viven de la misma manera las situaciones de agresión, aunque sean las mismas. Por su parte, las chicas admitieron que las agresiones sí les molestaban, y que incluso habían roto amistades en el pasado si esas amistades les habían “hecho algo”. Por otra parte, los chicos calificaron casi todas las agresiones que mencionaron como de baja gravedad y como “bromas” que se hacían habitualmente entre amigos, de forma recíproca, y que no molestaban a nadie. Hemos calificado esta realidad que viven los chicos como *friendly teasing*.

Además, en general, los chicos mencionaron más situaciones de agresión que las chicas, siendo esto consistente con los tradicionales hallazgos que afirman que los chicos reciben más *bullying* que las chicas (e.g., Carrera-Fernández et al., 2013; Hanzi y Akça, 2020; Lee et al., 2022; Wang et al., 2010; Zhao et al., 2022). Sin embargo, a pesar de sufrir más comportamientos agresivos, la percepción que tienen los chicos de estas situaciones es, como se ha mencionado anteriormente, “normalizadora” y no negativa. Esto es, a pesar de que los chicos estén más familiarizados con las situaciones violentas en su día a día y las sufran con más frecuencia, en su mayoría, no las perciben como situaciones que les afecten negativamente y las categorizan como “bromas o juegos” que tienen con su grupo de amigos —*friendly teasing*—. Sin embargo, las chicas, a pesar de estar envueltas en un menor número de situaciones agresivas, cuando lo están, sí las perciben como negativas y reconocen sentirse psicológicamente afectadas por ellas, generando problemas serios en su sentido de pertenencia

al centro y en las relaciones con sus iguales. Esto no quiere decir que no haya chicos que sufren psicológicamente por las agresiones que se les acometen —los hay—, pero, según lo que los propios chicos manifestaron en la entrevista sobre las condiciones para considerar una agresión como *bullying* o no, sí que parece que para que un chico llegue a admitir estar afectado y envuelto en un problema, las agresiones tienen que llegar a cotas muy altas (e.g., ser muy frecuentes, tener un gran impacto), y no ser perpetradas por un amigo de la víctima. Cabe también hacerse la pregunta de si realmente los chicos no sufren por estas situaciones de *friendly teasing* o de si enmascaran su sufrimiento debido a los estereotipos y roles de género que los llevan a “hacerse los fuertes” y a esconder su debilidad (e.g., Dieckman y Eagly, 2008). Ambas opciones son plausibles y, con los datos que tenemos hasta el momento, no somos capaces de dar una respuesta contundente a la pregunta, puesto que detectar ese tipo de incongruencias en el discurso sigue siendo un reto para la investigación.

Se encontró también que estas diferentes experiencias pueden tener influencias diferenciales en los procesos de respuesta de chicas y chicos cuando contestan a ítems sobre victimización de *bullying*. Por un lado, con muchas chicas se corre el riesgo de que no reporten en el cuestionario todas las agresiones que han sufrido porque piensan que el *bullying* son “palabras mayores”, y que se refiere únicamente a agresiones directas o incluso físicas, y en consecuencia tienden a minimizar la importancia de las agresiones indirectas que ellas sufren. No obstante, en general las chicas tienen la “ventaja” de admitir más fácilmente cuándo una agresión las está afectando psicológicamente, lo que las predispone a detectar con más facilidad situaciones de *bullying*. Por otro lado, los chicos, como se ha mencionado anteriormente, necesitan más condicionantes para considerar una situación como *bullying*, y también les cuesta más admitir que una agresión les afecta, por lo que en general será más difícil que identifiquen una situación como *bullying*. En su lugar, tienden a vivir las agresiones como *friendly teasing*, y eso presenta una serie de consecuencias en su forma de contestar a los cuestionarios. Algunos

chicos, por el hecho de considerar las agresiones como *friendly teasing*, podrían no reportar en el cuestionario agresiones que hayan sufrido porque para ellos son “juegos”. Otros chicos, sin embargo, a pesar de considerarlas “juegos”, las seguirán reportando en el cuestionario. Dependiendo de la postura que tomen o de los objetivos que tengan las personas evaluadoras, una de estas cuestiones resultará problemática porque representará un problema de *underreporting* u *overreporting*, respectivamente.

Estos hallazgos del Estudio III, obtenidos a través de entrevistas cognitivas, sirven también para explicar de manera más exhaustiva los hallazgos obtenidos en el Estudio II y así generar un corpus de conocimiento más amplio y comprehensivo. Así, se procede a continuación a culminar la integración de esta estrategia mixta de validación, llevando a cabo la integración al nivel de la interpretación y la comunicación de los resultados (Fetters et al., 2013). Es, como se comentó en el Capítulo III (Metodología), una integración narrativa que sigue una *weaving approach* (Fetters et al., 2013), en la que se explican y detallan juntos los hallazgos cuantitativos y cualitativos, “tema por tema” o “concepto por concepto”. A continuación, por tanto, se explica con más detalle, utilizando la información obtenida en el Estudio III, la naturaleza de cada clase encontrada en el Estudio II, a través del LCA:

- ❖ **Bullied (All Types):** esta clase está conformada por personas con alta probabilidad de responder afirmativamente en todos los ítems. Esto es, son los chicos y las chicas que reciben todos los tipos de agresiones. Principalmente, la mayoría de las personas pertenecientes a esta clase son chicos. Esto podría ser así porque, en la mayoría de los casos, cuando una chica sufre *bullying*, recibe únicamente agresiones de tipo relacional y es infrecuente que una chica se vea envuelta en agresiones más directas.
- ❖ **Not Bullied:** esta clase la conforman personas con baja probabilidad de responder afirmativamente en todos los ítems, es decir, son las chicas y los chicos que no reciben

ninguna de las agresiones. Esta es, afortunadamente, la clase más numerosa, puesto que tanto la mayoría de los chicos como la mayoría de las chicas pertenecen a ella.

❖ **Relational Bullying:** en esta clase se encuentran las personas con alta probabilidad de responder afirmativamente los ítems representativos de *bullying* indirecto o relacional, y con baja probabilidad de responder afirmativamente a los demás ítems. La gran mayoría de personas pertenecientes a esta clase son chicas, cuestión que se corrobora con los testimonios de las entrevistas cognitivas. El hecho de que las chicas manifestaran sufrir más agresiones indirectas y el hecho de que los chicos prácticamente no mencionaran sufrir situaciones de ese tipo —dándose el caso de que ninguno de los chicos mencionó haber sido víctima de exclusión—, confirman que esta clase está conformada, principalmente, por chicas que reciben *bullying* de tipo relacional.

❖ **Potential Friendly Teasing:** esta clase la conforman personas con probabilidad media de responder afirmativamente a casi todos los ítems, excepto al relativo a las amenazas y al relativo a la exclusión social. La gran mayoría de las personas pertenecientes a esta clase son chicos. Inicialmente, esta clase resultó ser la más enigmática de las que se encontraron en el Estudio II, puesto que no conformó el patrón esperado de *bullying* directo, sino que presentaba personas de las que, sobre todo, se reían y se burlaban, pero a las que nunca excluían. Además, a algunas de ellas les quitaban sus pertenencias, difundían rumores sobre ellas, e incluso les golpeaban, aunque a muy pocas las amenazaban. En la discusión del Estudio II se barajaron varias posibilidades para explicar qué tipo de personas eran las pertenecientes a esta clase, pero no se supo dar una respuesta definitiva a cuál de esas posibilidades era la correcta. Una de las finalidades del Estudio III fue darle respuesta a esta pregunta. Con la información de

las entrevistas cognitivas se puede inferir que esta clase está conformada, en efecto, por chicos que experimentan *friendly teasing*, el cual ha demostrado tener un patrón compatible con el que se encontró en el LCA: la agresión que con más frecuencia repiten dentro del *friendly teasing* son las risas y las burlas, y ocasionalmente también se quitan sus pertenencias. A veces, incluso se golpean entre ellos —manifiestan que, “jugando” —, pero no es habitual que pasen a amenazarse, puesto que eso lo viven como algo “que va más en serio”. Como es una actividad que realizan siempre entre amigos, nunca se excluyen unos a otros. Por último, fueron los chicos los que mayoritariamente describieron este tipo de experiencias, concordando todo con los hallazgos del Estudio II.

Por último, a la luz de estos resultados, cabría preguntarse por qué no se obtuvo en el Estudio II una clase que representara a chicos que sólo sufren *bullying* directo, como sí se obtuvo una que representa a chicas que sólo sufren *bullying* relacional. A pesar de que, con los datos que se han obtenido, no se puede dar una respuesta definitiva a esta pregunta, es posible que la respuesta tenga que ver con los roles y estereotipos de género. Mientras que a las chicas sí se las socializa para que no sean directamente agresivas (e.g., Diekman y Eagly, 2008), haciendo menos probable que lleven a cabo conductas agresivas directas como golpear, amenazar o insultar, los roles y estereotipos de género de los chicos no les cohiben a la hora de realizar agresiones indirectas. En otras palabras, que una chica sea ruda y golpee o amenace a alguien es contrario a los roles y estereotipos de género que se les imponen a las chicas, pero que un chico aisle socialmente a otro o difunda rumores sobre él no es contrario a los roles y estereotipos de género que se les imponen a los chicos. Esta situación es la que puede generar que las chicas, por un lado, sí que tiendan a perpetrar y sufrir solamente agresiones indirectas, mientras que por otro lado, los chicos perpetren y sufran todo tipo de agresiones —y no solamente las directas—.

2. LIMITACIONES E INVESTIGACIONES FUTURAS

A lo largo del trabajo se han identificado una serie de limitaciones que es preciso señalar para que se tengan en cuenta a la hora de interpretar los resultados de la tesis doctoral, y de guiar la dirección de investigaciones futuras. En primer lugar, el análisis de datos cuantitativo se realizó sobre una base de datos ya existente de la edición del año 2018 del estudio PISA. A pesar de que esta base de datos es suficientemente extensa y representativa, puesto que está recogida bajo los estándares de calidad del estudio PISA, se trata de una muestra “pre-pandémica”. A pesar de eso, se prefirió trabajar con los datos del PISA 2018 en lugar de con los datos del PISA 2022 porque esta última edición se realizó en tiempos en los que todavía no se había alcanzado “la nueva normalidad” después de la pandemia del COVID-19, y muchos centros impartían la docencia *online* o de forma híbrida. Esto pudo hacer que los resultados de variables tan dependientes de la interacción social como son las relacionadas con el clima escolar o *bullying* se pudieran ver de alguna manera alteradas por la situación de excepcionalidad, y, por tanto, no representaran fidedignamente la realidad imperante en los centros educativos. Aun así, los datos del PISA 2018 no reflejan los cambios que podrían haberse producido a raíz de la pandemia y otros recientes acontecimientos a nivel mundial y social, y, por tanto, no serían completamente representativos de la sociedad actual, teniendo en cuenta la posibilidad de que existan diferencias generacionales entre los y las adolescentes actuales y los de hace siete u ocho años, atribuibles a los cambios sociales provocados por la pandemia. Así, en investigaciones futuras sería conveniente analizar la consistencia de los resultados con datos de ediciones más recientes del estudio PISA, como la próxima que está prevista para este mismo año 2025.

Otra indicación para la investigación futura podría ser la de ampliar las variables de estudio, tanto para estudios cuantitativos como cualitativos. Sería interesante comprobar si los

resultados y evidencias obtenidas se replican en diferentes tipos de centros (público vs. privado), diferentes entornos (urbano vs. rural), e incluso en diferentes países o culturas. Así, sería conveniente, sobre todo para el análisis cualitativo, contar con una muestra más amplia y variada de participantes. Un método muy adecuado para conseguir muestras más grandes y más extendidas geográficamente es el *web probing*, que consiste en la aplicación de las técnicas propias de la entrevista cognitiva en el entorno web (Behr et al., 2017). Al tratarse de un método que se desarrolla a través del entorno web, facilita enormemente el acceso a números mayores de personas, así como a personas que se encuentran lejos geográficamente y con las que sería muy costoso contactar a través de vías tradicionales basadas en el contacto en persona, como la entrevista cognitiva (Behr et al., 2017). Investigaciones futuras podrían utilizar el *web probing* para ampliar la muestra de la recogida de datos cualitativa y replicar, e incluso ampliar los resultados obtenidos en la tesis doctoral.

Un reto para la investigación futura es el de, a nivel más sustantivo, tratar de responder a la pregunta de si los chicos realmente no se ven afectados psicológicamente por las agresiones que sufren e interpretan en el contexto del *friendly teasing*, o de si están ocultando su debilidad a causa de los “mandatos de género” (e.g., Dieckman y Eagly, 2008). Como es bien sabido, detectar mentiras o incongruencias de este tipo es extremadamente difícil, y hacerlo con un nivel de seguridad notorio, aún más. Por eso, presentamos esta posible línea futura de investigación como un reto a superar, animando a las investigaciones futuras a que exploren métodos —probablemente basados en la entrevista cognitiva— para detectar este tipo de incongruencias.

Por último, es preciso tener en cuenta que esta estrategia mixta de validación se ha probado en un contexto determinado —la evaluación de la victimización de *bullying*— con un instrumento de medida determinado —la escala de experiencias de *bullying* (ST038) del Cuestionario del Estudiante del PISA 2018—, por lo que resulta necesario seguir probando esta

estrategia en investigaciones futuras con otros instrumentos de medida y en otros contextos de evaluación, examinando otros fenómenos o constructos y utilizando diferentes grupos de comparación a la hora de examinar la invarianza métrica. Solo con la aplicación repetida y diversa de esta estrategia se conseguirá, a su vez, aportar evidencias de validez para la misma, que consigan apoyar la idea de que puede ser generalizable y utilizada de forma efectiva, independientemente del contexto de evaluación que se quiera examinar.

3. IMPLICACIONES PRÁCTICAS

La presente tesis doctoral presenta importantes implicaciones prácticas, tanto a nivel sustantivo como a nivel metodológico. Por un lado, a nivel sustantivo, los hallazgos de esta tesis doctoral sobre cómo se manifiesta el fenómeno del *bullying* en chicas y chicos alientan a que en el futuro, a la hora de hacer evaluaciones sobre dicho fenómeno, se tengan en cuenta estas diferencias de género en las experiencias de *bullying*, así como los problemas que generan diferencialmente en ambos grupos (e.g., chicas que no consideran que estén recibiendo *bullying* porque no son agresiones directas, o chicos que no consideran que estén recibiendo *bullying* porque lo interpretan como *friendly teasing*). Además, la tesis abre un camino para indagar cómo esas diferencias de género se trasladan a los procesos de respuesta a los ítems de los instrumentos de evaluación. Al tener en cuenta estas cuestiones, se podrán interpretar debidamente las puntuaciones de ambos grupos en las evaluaciones y se podrán sacar conclusiones más acertadas y fieles a la realidad que chicos y chicas viven, detectando de manera más precisa casos que estén sufriendo *bullying* o que estén en riesgo de sufrirlo, incluso detectando algunos que de otra manera no se hubieran detectado. A su vez, esta mejor comprensión del fenómeno permitirá diseñar programas de prevención y de intervención más ajustados a la realidad y personalizados según el género, en lugar de implementar programas

genéricos y estandarizados que no contemplan las idiosincrasias y singularidades que tienen las experiencias de *bullying* de chicas y chicos. Así, en las chicas podría ponerse más atención a las formas de agresión indirectas, a la minimización de la importancia de las mismas, a las relaciones de amistad rotas de repente, a desmitificar la idea de que el *bullying* es solamente directo, y a alentar a que comuniquen a sus centros si están sufriendo alguna situación que las hace sentirse mal o incómodas emocionalmente. En los chicos, podría ponerse atención a cómo interactúan los grupos de amigos entre ellos, a desmitificar la idea de que un “amigo” no puede hacerte daño o perpetrar *bullying*, a fomentar el desarrollo de su inteligencia emocional para que aprendan a identificar cuándo se sienten mal o les duele algo que les han hecho, e incluso a trabajar cuestiones más ligadas a la masculinidad tradicional como la legitimación de la violencia o el miedo a mostrarse como “débiles”. En definitiva, podrían dejarse atrás las interpretaciones simplistas, tales como: “los chicos reciben más *bullying* que las chicas”, que son interpretaciones que dirigen injustamente la mirada y la atención hacia uno de los dos grupos y que no reflejan la complejidad de una realidad en la que ambos grupos sufren a su manera, y se sustituirían por interpretaciones del fenómeno que realmente nos dirijan a desarrollar intervenciones que suplan las necesidades concretas de cada grupo, de manera individualizada y justa.

Además, a nivel metodológico, la tesis doctoral desarrolla una estrategia de validación y muestra la utilidad de los métodos mixtos para aportar evidencias de validez, en especial, las basadas en los procesos de respuesta. Así, se demuestra la eficacia de métodos cuantitativos como los modelos de clases latentes para abordar el problema de la invarianza métrica y comenzar a explorar las causas de la falta de invarianza, tal y como proponen Bruno D. Zumbo et al. (2015), y también la eficacia de métodos cualitativos como la entrevista cognitiva para indagar en los procesos de respuesta de las personas que contestan un test. Esta estrategia mixta de validación está además centrada en abordar el problema de la justicia en las mediciones,

puesto que se centra en aportar evidencias de validez para apoyar las interpretaciones comparativas de las puntuaciones de dos o más grupos de personas que se diferencian a nivel sociodemográfico, lingüístico y/o cultural, con el objetivo de nunca beneficiar o perjudicar a uno de los grupos por cuestiones irrelevantes al fenómeno o constructo objeto de medición, de acuerdo con las indicaciones de los *Standards for Educational and Psychological Testing* (AERA et al., 2014). De esta manera, esta estrategia mixta de validación que se desarrolla y propone en la presente tesis doctoral sirve, por un lado, para que se aborden los procesos de evaluación e interpretaciones de las puntuaciones de manera más justa y ética, y por otro lado, para que futuros y futuras profesionales que deseen realizar interpretaciones de puntuaciones de tests válidas y justas se aproximen con un abordaje de la validez actualizado y metodológicamente preciso, basado en la definición actual de validez de los *Standards for Educational and Psychological Testing* (AERA et al., 2014), en conceptualizaciones recientes mediante las cuales la invarianza métrica se considera un problema de validez (Zumbo y Padilla, 2019), y en el Modelo Ecológico de Respuesta a los Ítems (Zumbo et al., 2015), que tiene en cuenta que diferentes variables pueden influir en los procesos de respuesta al contestar un conjunto de ítems. Además, esta estrategia mixta de validación está inmersa en la “cuarta generación del DIF” (Zumbo, en prensa), ya que tiene como objetivo explorar concienzudamente las realidades y experiencias de las personas evaluadas, para tenerlas en cuenta a la hora de entender cómo han interactuado con el instrumento de evaluación e interpretar sus puntuaciones en él de manera válida y justa, visibilizando todas las realidades y “formas de ser humano”. Esta estrategia mixta de validación podría suponer un elemento y un punto de partida que trace un camino marcado por la evaluación y la ciencia inclusivas por el que guiar a futuras y futuros profesionales cuando administren tests o cuestionarios, y también a futuras investigaciones en el campo de la Metodología.

4. CONCLUSIONES

A lo largo de esta tesis doctoral, se ha desarrollado y probado una estrategia mixta de validación centrada en aportar evidencias de validez basadas en los procesos de respuesta y en asegurar la justicia de las interpretaciones comparativas de las mediciones, a la par que se han aportado explicaciones sobre las diferencias entre chicas y chicos en su percepción del clima escolar y sus experiencias de *bullying*, permitiendo una comprensión más exhaustiva y comprehensiva del fenómeno.

Así, en primer lugar, se ha demostrado que con la combinación de los métodos cuantitativo y cualitativo se ha accedido a datos y evidencias a las que probablemente no se habría accedido utilizando únicamente uno de los métodos. Por ejemplo, gracias a la información que se obtuvo de las entrevistas cognitivas en la parte cualitativa, pudo otorgarse una explicación completa y precisa de la naturaleza de la clase *Potential Friendly Teasing* que se encontró con el LCA en la parte cuantitativa. Esta información no se habría obtenido y no se podría haber dado una explicación clara y fidedigna para dicha clase de haber solamente contado con los métodos cuantitativos. Este tipo de información resulta esencial a la hora de interpretar de manera válida y justa las mediciones y a la hora de comparar las puntuaciones entre diferentes grupos de personas (en este caso, entre chicos y chicas), sin beneficiar o perjudicar a uno de ellos por razones ajenas a los objetivos de la medición. Este corpus de conocimiento más comprehensivo que se consigue con la integración de ambos métodos posibilita un entendimiento del fenómeno de estudio más fiel a la realidad, y por tanto, más ecológico y más válido, y en consecuencia, más justo.

Además, con los Estudios II y III de la tesis doctoral se ha avanzado en la aportación de conocimiento sustantivo del fenómeno del *bullying*, apoyando la idea que ya se discutía en

literatura previa de que las experiencias de *bullying* de chicas y chicos son diferentes, y afinando y aportando, además, información más precisa sobre esas diferencias en el contexto español, con el beneficio que supone el uso de una metodología mixta. Así, se ha llegado a la conclusión de que se puede afirmar que existen dos realidades muy diferentes en cuanto a cómo viven las experiencias de victimización de *bullying* chicos y chicas, que afectan diferencialmente a sus procesos de respuesta al contestar ítems sobre victimización de *bullying*:

- ❖ **Las chicas** tienden a perpetrar y recibir únicamente agresiones indirectas o relacionales, debido posiblemente a la socialización de género que reciben (e.g., Diekman y Eagly, 2008). También, son más propensas a admitir que las agresiones les afectan —posiblemente a causa, de nuevo, de su socialización de género (e.g., Diekman y Eagly, 2008)—, haciendo esto que sea más probable que detecten situaciones de *bullying*. A pesar de que están envueltas en menos experiencias violentas que los chicos, no hay que desdeñar el impacto psicológico y social que generan en ellas aquellas que sufren. Además, algunas chicas podrían no reportar algunas de sus experiencias debido a que consideran que el *bullying* “real” es de tipo físico.
- ❖ **Los chicos**, en general, cuando sufren *bullying* reciben agresiones de todo tipo, aunque las más preponderantes son las agresiones de tipo directo. Estas experiencias violentas son muy habituales en su día a día, y la mayoría las viven como “bromas” o “juegos” que realizan con sus amigos, y no como experiencias negativas que les afecten psicológicamente. Esto puede ser debido a su socialización de género, que ha legitimado la violencia y las luchas de poder (e.g., Diekman y Eagly, 2008). Esta forma de experimentar las agresiones como *friendly teasing* puede hacer que parte de estos chicos no las reporten en los cuestionarios sobre victimización de *bullying*. Si las personas evaluadoras consideran que, independientemente de que los chicos no admitan que las agresiones les afectan, esas agresiones pueden tener un impacto negativo en

ellos y es necesario detectarlas, deben tener en cuenta que se enfrentan a un problema de *underreporting*. Otros chicos, por su parte, a pesar de no considerar estas agresiones como problemáticas, es posible que sí las reporten en los cuestionarios. Si las personas evaluadoras consideran que no son iguales las agresiones que la víctima percibe como negativas que aquellas que forman parte del contexto de *friendly teasing*, y que deberían distinguirse unas de las otras en la evaluación, deben tener en cuenta que se enfrentan a un problema de *overreporting* de “falsos positivos”.

En definitiva, en esta tesis doctoral se ha hecho patente que una estrategia mixta de validación basada en el LCA y las entrevistas cognitivas es capaz de generar un corpus de conocimiento amplio que, a su vez, permite aportar evidencias de validez que de otra manera no se hubieran obtenido, y que contribuyen a abordar el problema de la equivalencia métrica y de la justicia en las mediciones, cuestiones imprescindibles para llevar a cabo una ciencia y una práctica profesional rigurosa, ética e inclusiva.

REFERENCIAS

REFERENCIAS

American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME] (1966). *Standards for Educational and Psychological Tests and Manuals*. American Psychological Association.

American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME]. (1974). *Standards for Educational and Psychological Tests*. American Psychological Association.

American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME]. (1985). *Standards for Educational and Psychological Testing*. American Psychological Association.

American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME]. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association [AERA]. Recuperado de <https://www.aera.net/Publications/Books/Standards-1999-Ed>

American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME]. (2014). *Standards for Educational and Psychological Testing* [Estándares para Pruebas Educativas y

- Psicológicas]. American Educational Research Association [AERA]. Recuperado de <https://www.aera.net/Standards14-Spanish>
- American Psychological Association [APA]. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51(2, Pt. 2), 1–38. <https://doi.org/10.1037/h0053479>
- American Psychological Association [APA]. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, 47, 1597–1611. Recuperado de <https://www.apa.org/ethics/code>
- Angoff, W. (1993). Perspective on differential item functioning methodology. En P. W. Holland y H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203357811>
- Ayuso, N., Murillo, A. C. y 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. <https://doi.org/10.1109/TE.2020.3004075>
- Bazeley, P. (2024). Conceptualizing integration in mixed methods research. *Journal of Mixed Method Research*, 18(3), 225–234. <https://doi.org/10.1177/15586898241253636>
- Beatty, P. C. y Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311. <https://doi.org/10.1093/poq/nfm006>
- Behr, D., Meitinger, K., Braun, M. y Kaczmirek, L. (2017). *Web probing—implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions*. GESIS Survey Guidelines. https://doi.org/10.15465/gesis-sg_en_023

- Benítez, I. y Padilla J. L. (2014). Cognitive interviewing in mixed research. En K. Miller, S. Willson, V. Chepp y J. L. Padilla (Eds.), *Cognitive Interviewing Methodology*. Wiley.
<https://doi.org/10.1002/9781118838860>
- Benítez, I., van de Vijver, F. y Padilla, J. L. (2019). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research*, 51(1), 1–34.
<https://doi.org/10.1177/0049124119852390>
- Bertocchi, G. y Bozzano, M. (2020). Gender gaps in education. En K. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Springer.
https://doi.org/10.1007/978-3-319-57365-6_26-1
- Björkqvist, K., Lagerspetz, K. M. J. y Kaukiainen, A. (1992). Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive Behavior*, 18(2), 117–127. [https://doi.org/10.1002/1098-2337\(1992\)18:2%3C117::AID-AB2480180205%3E3.0.CO;2-3](https://doi.org/10.1002/1098-2337(1992)18:2%3C117::AID-AB2480180205%3E3.0.CO;2-3)
- Bonilla-Campos, A. (2010). Psicología, diferencias y desigualdades: límites y posibilidades de la perspectiva de género feminista. *Quaderns de Psicologia*, 12(2), 65–80.
<https://doi.org/10.5565/rev/qpsicologia.806>
- Booth, A. (2006). Clear and present questions: formulating questions for evidence based practice. *Library Hi Tech*, 24(3), 355–368. <https://doi.org/10.1108/07378830610692127>
- Camerini, A. L. y Schulz, P. J. (2018). Social desirability bias in child-report social well-being: Evaluation of the children’s social desirability short scale using item response theory and examination of its impact on self-report family and peer relationships. *Child Indicators Research*, 11, 1159–1174. <https://doi.org/10.1007/s12187-017-9472-9>

- Carrasco, C., Mayordomo, M., Domínguez, M. y Alabart, A. (2004). *El trabajo con mirada de mujer. Propuesta de una encuesta de población activa no androcéntrica*. Consejo Económico y Social.
- Carrera-Fernández, M. V., Lameiras-Fernández, M., Rodríguez-Castro, Y., Failde-Garrido, J. M. y Calado-Otero, M. (2013). Bullying in Spanish secondary schools: Gender-based differences. *Spanish Journal of Psychology*, 16(21), 1–14.
<https://doi.org/10.1017/sjp.2013.37>
- Carroll, S. J. (2018). The gender gap as a tool for women's political empowerment: the formative years 1980–1984. En A. Maxwell y T. Shields (Eds.), *The Legacy of Second-Wave Feminism in American Politics*. Springer. <https://doi.org/10.1007/978-3-319-62117-3>
- Castillo, M. y Padilla, J. L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, 114, 963–975.
<https://doi.org/10.1007/s11205-012-0184-8>
- Chen, M. Y. y Zumbo, B. D. (2017). Ecological framework on item responding as validity evidence: an application of multilevel DIF modeling using PISA data. En B. D. Zumbo y A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research*. Springer. <https://doi.org/10.1007/978-3-319-56129-5>
- Chung, J. Y. y Lee, S. (2020). Are bully-victims homogeneous? Latent class analysis on school bullying. *Children and Youth Services Review*, 112, 104922.
<https://doi.org/10.1016/j.childyouth.2020.104922>

- Cizek, G. J., Rosenberg, S. L. y Koons, H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397–412. <https://doi.org/10.1177/0013164407310130>
- Cobrerros, L., Galindo, J. y Raigada, T. (2024). *Mujeres en STEM. Desde la educación básica hasta la carrera laboral*. Esade. Center for Economic Policy. Recuperado de <https://www.esade.edu/ecpol/wp-content/uploads/2024/03/Mujeres-en-STEM-2024-1.pdf>
- Colegio Oficial de Psicólogos [COP]. (1987). *Código deontológico del psicólogo*. Colegio Oficial de Psicólogos. Recuperado de <https://www.cop.es/index.php?page=codigodeontologico>
- Comisión Europea (2009). *Manual. El género en la investigación*. European Communities y Ministerio de Ciencia e Innovación. <https://doi.org/10.2777/23655>
- Cooke, A., Smith, D. y Booth, A. (2012). Beyond PICO: The SPIDER tool for qualitative evidence synthesis. *Qualitative Health Research*, 22(10), 1435–1443. <https://doi.org/10.1177/1049732312452938>
- Crabtree, B. F., Miller, W. L., Tallia, A. F., Cohen, D. J., DiCicco-Bloom, B., McIlvain, H. E., Aita, H. A., Scott, J. G., Gregory, P. B., Stange, K. C. y McDaniel, R. R. (2005). Delivery of clinical preventive services in family medicine offices. *Annals of Family Medicine*, 3(5), 430–435. <https://doi.org/10.1370/afm.345>
- Creswell, J. W. y Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. 2nd edition. Sage Publications. Recuperado de <https://psycnet.apa.org/record/2006-11884-000>

- Crowne, D. P. y Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349-354.
<https://doi.org/10.1037/h0047358>
- Díaz-Aguado, M. J., Martínez-Arias, R., Falcón, L. y Alvariño, M. (2023). *Acoso escolar y ciberacoso en España en la infancia y en la adolescencia*. Fundación ColaCao y Universidad Complutense de Madrid. Recuperado de <https://fundacioncolacao.org/investigacion>
- Diekmann, A. B. y Eagly, A. H. (2008). Of men, women, and motivation. En J. Y. Shah, y W. L. Gardner (Eds.), *Handbook of motivation science*. Guilford Press. Recuperado de <https://psycnet.apa.org/record/2008-00543-000>
- Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Psychology Press. <https://doi.org/10.4324/9780203781906>
- Eagly, A. H. y Wood, W. (2012). Social role theory. En P. van Lange, A. Kruglanski, y E. T. Higgins (Eds.), *Handbook of theories in social psychology*. Sage.
<https://doi.org/10.4135/9781446249222.n49>
- Ferrando, P. J. y Chico, E. (2000). Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne. *Psicothema*, 12(3), 383–389. Recuperado de <https://www.psicothema.com/pii?pii=346>
- Ferrer-Pérez, V. (2019). Una mirada feminista a la psicología social. *International Journal of Social Psychology: Revista de Psicología Social*, 34(2), 1–25.
<https://doi.org/10.1080/02134748.2019.1589784>

- Fetters, M. D., Curry, L. A. y Creswell, J. W. (2013). Achieving integration in mixed methods designs – Principles and practices. *Health Services Research*, 48(6), 2134–2156. <https://doi.org/10.1111/1475-6773.12117>
- Flake, J. K., Davidson, I. J., Wong, O. y Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, 77(4), 576–588. <https://doi.org/10.1037/amp0001006>
- Garb, H. N. (2021). Race bias and gender bias in the diagnosis of psychological disorders. *Clinical Psychology Review*, 90, 102087. <https://doi.org/10.1016/j.cpr.2021.102087>
- García-Calvente, M. M., Jiménez-Rodrigo, M. L. y Martínez-Morante, E. (2010). *Guía para incorporar la perspectiva de género a la investigación en salud*. Escuela Andaluza de Salud Pública. Recuperado de <https://www.easp.es/project/guia-para-incorporar-la-perspectiva-de-genero-a-la-investigacion-en-salud/>
- Gladden, M. R., Vivolo-Kantor, A. M., Hamburger, M. E. y Lumpkin, C. D. (2014). *Bullying surveillance among youths: Uniform definitions for public health and recommended data elements. Version 1.0*. Centers for Disease Control and Prevention. Recuperado de <https://stacks.cdc.gov/view/cdc/21596>
- Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D. y Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema*, 30(1), 104–109. <https://doi.org/10.7334/psicothema2017.183>
- Gutiérrez, S., Sanz, J., Espinosa, R., Gesteira, C. y García-Vera, M. P. (2016). La Escala de Deseabilidad Social de Marlowe-Crowne: baremos para la población general española y desarrollo de una versión breve. *Anales de Psicología*, 32(1), 206–217. <https://doi.org/10.6018/analesps.32.1.185471>

- Hanzi, I. H. y Akça, S. Ö. (2020). School bullying. *Journal of Education and Future*, 18, 113–121. <https://doi.org/10.30786/jef.691717>
- He, J. y van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2), 1–19. <https://doi.org/10.9707/2307-0919.1111>
- Hervás, A. (2022). Género femenino y autismo: infra detección y mis diagnósticos. *Medicina (Buenos Aires)*, 82(1), 37–42. Recuperado de <https://pubmed.ncbi.nlm.nih.gov/35171806/>
- Hogan, T. P. y Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, 64(5), 802–812. <https://doi.org/10.1177/0013164404264120>
- Holland, P. W. y Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203056905>
- Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., Wolfe, M. y Reid, G. (2015). Bullying and suicidal ideation and behaviors: A meta-analysis. *Pediatrics*, 135(2), 496–509. <https://doi.org/10.1542/peds.2014-1864>
- Hong, Q. N., Pluye, P., Bujold, M. y 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. <https://doi.org/10.1186/s13643-017-0454-2>

- Hurtado-García, I. (2019). Gender bias in ADHD: the pathologization of gender roles. En E. Pérez-Sedeño, L. S. Almendros, S. García-Dauder y E. Ortega-Arjonilla (Eds.), *Knowledges, Practices and Activism from Feminist Epistemologies*. Vernon Press.
- Johnson, R. B., Onwuegbuzie, A. J. y Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2), 112–133. <https://doi.org/10.1177/1558689806298224>
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kim, Y. S. y Leventhal, B. (2008). Bullying and suicide. A review. *International Journal of Adolescent Medical Health*, 20(2), 13–154. <https://doi.org/10.1515/ijamh.2008.20.2.133>
- Koskey, K. L. K. y Stewart, V. C. (2014). A concurrent mixed methods approach to examining the quantitative and qualitative meaningfulness of absolute magnitude estimation scales in survey research. *Journal of Mixed Methods Research* 8(2), 180–202. <https://doi.org/10.1177/1558689813496905>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Lee, J. (2021). Pathways from childhood bullying victimization to young adult depressive and anxiety symptoms. *Child Psychiatry & Human Development*, 52, 129–140. <https://doi.org/10.1007/s10578-020-00997-4>

- Lee, J., Roh, B. R. y Yang, K. E. (2022). Exploring the association between social support and patterns of bullying victimization among school-aged adolescents. *Children and Youth Services Review*, 136, 106418. <https://doi.org/10.1016/j.childyouth.2022.106418>
- Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gülmezoglu, M., Noyes, J., Booth, A., Garside, R. y 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. <https://doi.org/10.1371/journal.pmed.1001895>
- Lizarondo, L., Stern, C., Carrier, J., Godfrey, C., Rieger, K., Salmond, S., Apóstolo, J., Kirkpatrick, P. y Loveday, H. (2020). Mixed methods systematic reviews. En E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, Z. Jordan (Eds.), *JBIManual for Evidence Synthesis*. <https://doi.org/10.46658/JBIMES-24-07>
- Menéndez-Monte, B., Alcedo-Rodríguez, M. A. y Fontanil-Gómez, Y. (2022). La discriminación de las mujeres en el “Manual diagnóstico de los trastornos mentales-DSM-5”. En J. García-Fernández, G. A. Giacobbe y R. Riestra-Camacho (Eds.), *Misoginia y Filoginia. Fuerzas discursivas simbólicas en la narrativa internacional*. Dykinson. Recuperado de <https://www.dykinson.com/libros/misoginia-y-filoginia/9788411226400/>
- Merone, L., Tsey, K., Russell, D. y Nagle, C. (2021). Sex and gender gaps in medicine and the androcentric history of medical research. *Australian and New Zealand Journal of Public Health*, 45(5), 424–426. <https://doi.org/10.1111/1753-6405.13139>
- Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3º ed., pp. 13–104). Macmillan.

- Miller, K., Willson, S., Chepp, V. y Ryan, J.M. (2014). Analysis. En K. Miller, S. Willson, V. Chepp y J.L. Padilla (Eds.), *Cognitive Interviewing Methodology*. Wiley.
<https://doi.org/10.1002/9781118838860>
- Moseholm, E. y Fетters, M. D. (2017). Conceptual models to guide integration during analysis in convergent mixed methods studies. *Methodological Innovations*, 10(2), 1–11.
<https://doi.org/10.1177/2059799117703118>
- Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10, 121–132.
<https://doi.org/10.3102/10769986010002121>
- Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. En H. Wainer y H. Braun (Eds.), *Test validity* (pp. 213–238). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203056905>
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 551–585. <https://doi.org/10.1007/BF02296397>
- Muthén, B. O., Kao, C. y Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, 28, 1–22. <https://doi.org/10.1111/j.1745-3984.1991.tb00340.x>
- Muthén, B. O. y Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, 10, 133–142.
<https://doi.org/10.3102/10769986010002133>

- Nylund, K., Bellmore, A., Nishina, A. y Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does Latent Class Analysis say? *Child Development*, 78(6), 1706–1722. <https://doi.org/10.1111/j.1467-8624.2007.01097.x>
- Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2018). *PISA 2018 technical report*. OECD Publishing. Recuperado de <https://www.oecd.org/pisa/data/pisa2018technicalreport/>
- Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2019a). *PISA 2018. Programa para la Evaluación Internacional de los Estudiantes. Informe Español*. OECD Publishing. Recuperado de https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=20372
- Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2019b). *PISA 2018 results (volume III): What school life for students' lives*. OECD Publishing. <https://doi.org/10.1787/acd78851-en>
- Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2023). *PISA 2022. Programa para la Evaluación Internacional de los Estudiantes. Informe Español*. OECD Publishing. Recuperado de <https://www.educacionfpydeportes.gob.es/inee/evaluaciones-internacionales/pisa/pisa-2022.html>
- Olweus, D. (1978). *Aggression in the schools: Bullies and whipping boys*. Hemisphere. Recuperado de <https://psycnet.apa.org/record/1979-32242-000>
- Olweus, D. (1994). Bullying at school: Basic facts and effects of school based intervention program. *Journal of Child Psychology and Psychiatry*, 35(7), 1171–1190. <https://doi.org/10.1111/j.1469-7610.1994.tb01229.x>

- Olweus, D. (2013). School bullying: Development and some important challenges. *Annual Review of Clinical Psychology*, 9, 751–780. <https://doi.org/10.1146/annurev-clinpsy-050212-185516>
- Padilla, J. L. y Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Padilla, J. L., Gómez, J., Hidalgo, M. D. y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema*, 18(2), 307–312. Recuperado de <https://www.psicothema.com/pi?pii=3215>
- 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. *British Medical Journal*, 372(71), 1–9. <https://doi.org/10.1136/bmj.n71>
- Pearson, A., White, H., Bath-Hextall, F., Salmond, S., Apóstolo, J. y Kirkpatrick, P. (2015). A mixed-methods approach to systematic reviews. *International Journal of Evidence-Based Healthcare*, 13(3), 121–131. <https://doi.org/10.1097/XEB.0000000000000052>
- Perrin, F. (2021). Can the historical gender gap index deepen our understanding of economic development? *Journal of Demographic Economics*, 88(3), 379–417. <https://doi.org/10.1017/dem.2020.34>
- Pluye, P. y Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health*, 35(1), 29–45. <https://doi.org/10.1146/annurev-publhealth-032013-182440>

- Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. y Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme*. <https://database.inahta.org/article/2684>
- Rivers, I. y Smith, P. K. (1994). Types of bullying behavior and their correlates. *Aggressive Behavior*, 20, 359–368. [https://doi.org/10.1002/1098-2337\(1994\)20:5%3C359::AID-AB2480200503%3E3.0.CO;2-J](https://doi.org/10.1002/1098-2337(1994)20:5%3C359::AID-AB2480200503%3E3.0.CO;2-J)
- Rosen, N. L. y Nofziger, S. (2019). Boys, bullying, and gender roles: How hegemonic masculinity shapes bullying behavior. *Gender Issues*, 36, 295–318. <https://doi.org/10.1007/s12147-018-9226-0>
- Saint-Arnault, D. y Fethers, M. D. (2011). RO1 funding for mixed methods research: lessons learned from the mixed-method analysis of japanese depression project. *Journal of Mixed Methods Research*, 5(4), 309–329. <https://doi.org/10.1177/1558689811416481>
- Sakellariou, C. y 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. <https://doi-org/10.1016/j.ijer.2021.101821>
- Serrano-Montilla, C., Lozano, L. M. y Padilla, J. L. (2025). Police officers' perspectives when handling intimate partner violence against women cases: A mixed-method approach using focus group and semantic network analysis. *Journal of Family Violence*, 1–15. <https://doi.org/10.1007/s10896-025-00913-1>
- Shealy, R. y Stout, W. F. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159–194. <https://doi.org/10.1007/BF02294572>

- Shear, B. R. y Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement. En B. D. Zumbo y E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. Springer. https://doi.org/10.1007/978-3-319-07794-9_6
- Shu, Y., Ota, M. y Kato, H. (2023). The diverse nature of school bullying and its relation to depression and the efficacy of coping strategies in Japan. *International Journal of Bullying Prevention*. Advance online publication. <https://doi.org/10.1007/s42380-023-00181-x>
- Shünemann, H., Brožek, J., Guyatt, G. y Oxman, A. (2013). GRADE Handbook. GRADE Working Group. Recuperado de <https://gdt.grade.org/app/handbook/handbook.html>
- Sireci, S. y Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, 35(3), 217–226. <https://doi.org/10.7334/psicothema2022.447>
- Sirriyeh, R., Lawton, R., Gardner, P. y Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice*, 18(4), 746–752. <https://doi.org/10.1111/j.1365-2753.2011.01662.x>
- Stern, C., Lizarondo, L., Carrier, J., Godfrey, C., Rieger, K., Salmond, S., Apóstolo, J., Kirkpatrick, P. y Loveday, H. (2020). Methodological guidance for the conduct of mixed methods systematic reviews. *JBIS Evidence Synthesis*, 18(10), 2108–2118. <https://doi.org/10.11124/JBISIR-D-19-00169>
- Swearer, S. M. y Hymel, S. (2015). Understanding the psychology of bullying. Moving toward a social-ecological diathesis-stress model. *American Psychologist*, 70(4), 344–353. <https://doi.org/10.1037/a0038929>

- Tashakkori, A. y Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches*. Sage Publications.
- Thomas, J., Kneale, D., McKenzie, J. E., Brennan, S. E. y Bhaumik, S. (2024). Determining the scope of the review and the questions it will address. En J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page y V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Collaboration. Recuperado de <https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current>
- Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, 26(2), 169–181. <https://doi.org/10.1108/QAE-06-2017-0034>
- United Nations Educational, Scientific and Cultural Organization [UNESCO]. (2019). *Behind the numbers: Ending school violence and bullying*. UNESCO. <https://doi.org/10.54675/TRVR4270>
- Vescio, T. K., Schlenker, K. A. y Lenes, J. G. (2010). Power and sexism. En A. Guinote y T. K. Vescio (Eds.), *The social psychology of power*. Guilford Press. Recuperado de <https://psycnet.apa.org/record/2010-13145-013>
- Wang, J., Iannotti, R. J., Luk, J. W. y Nansel, T. R. (2010). Co-occurrence of victimization from five subtypes of bullying: Physical, verbal, social exclusion, spreading rumors and cyber. *Journal of Pediatric Psychology*, 35(10), 1103–1112. <https://doi.org/10.1093/jpepsy/jsq048>
- Zhao, D., Wang, D., He, Z., Yuan, S., Zhu, D. y Liu, H. (2022). Victim profiles and the protective role of school anti-bullying norms: A study of Chinese adolescents. *Current Psychology*, 42, 31835–31852. <https://doi.org/10.1007/s12144-022-04204-6>

- 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. <https://doi.org/10.1080/15434300701375832>
- Zumbo, B. D. (2021). *Equity and fairness at the nexus of data science, psychometrics, digital innovation, and social justice* [Featured Invited Address]. Quantitative Methods Section of the Canadian Psychological Association (CPA) [virtual conference]. https://brunozumbo.com/?page_id=31
- Zumbo, B. D. (2023). A dialectic on validity: Explanation-focused and the many ways of being human. *International Journal of Assessment Tools in Education*, 10(Special Issue), 1–96. <https://doi.org/10.21449/ijate.1406304>
- Zumbo, B. D. (en prensa). Looking through the lens of DIF: Embracing the many ways of being human. *The Score*.
- Zumbo, B. D. y Chan, E. K. H. (2014). Setting the stage for validity and validation in social, behavioral, and health sciences: trends in validation practices. En B. D. Zumbo y E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences*. Springer. https://doi.org/10.1007/978-3-319-07794-9_1
- Zumbo, B. D. y Gelin, M. N. (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. *Journal of Educational Research and Policy Studies*, 5, 1–23.
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O. y 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.

<https://doi.org/10.1080/15434303.2014.972559>

Zumbo, B. D. y Padilla, J. L. (2019). The Interplay between survey research and psychometrics, with a focus on validity theory. En P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis y A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing*. Wiley. <https://doi.org/10.1002/9781119263685>

MATERIAL SUPLEMENTARIO

MATERIAL SUPLEMENTARIO DEL CAPÍTULO IV

Supplementary Material I. Search Strategy Adapted for Web of Science (English and Spanish Terms).

1. TS=(student* OR adolescent* OR teenager* OR young* OR estudiante* OR alumn* OR j?ven*)
2. AND TS=("gender gap" OR "gender inequalities" OR (diferencia* AND género))
3. AND TS=("measurement invariance" OR "measurement equivalence" OR "differential item functioning" OR DIF OR "item bias" OR "cognitive interview*" OR "in-depth interview*" OR "web probing" OR "focus group*" OR (invarianza AND medida*) OR (equivalencia AND medida*) OR "funcionamiento diferencial de los ítems" OR (sesgo* AND ítem*) OR "entrevista* cognitiva*" OR "entrevista* en profundidad" OR "grupo* focal*" OR "grupo* de discusión")
4. AND TS=((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) OR (evaluaci* AND educa*) OR (evaluaci* AND escolar) OR (evaluaci* AND académic*) OR "programa* educativo*" OR (estudiant* AND evaluaci*) OR (estudiant* AND program*) OR "rendimiento educativo" OR "rendimiento escolar" OR "rendimiento académico" OR (rendimiento AND estudiant*))

Supplementary Material II. Data Items.

All categories with (*) will be flexible to have emergent subcategories added to them.

Bibliometric indicators — bibliometric characteristics of the articles.

*First author's last name**

Article title

*Year of publication**

*First author's country**

Citations — number of citations each article has at the moment of data extraction.

Journal — name of the journal in which each article is published*

JCR — Is the journal indexed in Journal Citation Reports?

Yes

Q1

Q2

Q3

Q4

No

Subject area — JCR subject area of the journal (if several, highest percentile)*

Methodological aspects of the studies — methodological characteristics of each study (when an article includes more than one study, the methodological characteristics are extracted separately for each one).

Type of study — states if the study uses only quantitative methods, only qualitative methods, or both quantitative and qualitative methods.

Only quantitative

Only qualitative

Quantitative + qualitative

Data collection — states if study collects data or resorts to databases from previous studies.

Study collects data

Study uses available databases

Participants — main participants' characteristics.

Type of participants — states if participants are students, their teachers, their parents, or a combination (students + teachers + parents; students + teachers; students + parents; teachers + parents).

Only students

Only teachers

Only parents

Students + teachers + parents

Students + teachers

Students + parents

Teachers + parents

Country — country of origin of participants*

Sample size

Mean age of students

Participant recruitment — states how participant recruitment was carried out in each study*

Differential Item Functioning (DIF)/measurement invariance analysis — main characteristics of DIF/measurement invariance analyses.

Target variable(s) — states if variable(s) assessed belong to cognitive or non-cognitive domains.

Cognitive domain*

Non-cognitive domain*

Instrument(s) — measurement instrument(s) used to assess target variable(s).

Name of the instrument*

Number of items

Response format

Open-ended answers

Closed-ended answers

Both

Not specified

Item score format

Dichotomous

Polytomous

Both

Not specified

Test administration mode

Paper-based

Computer-based

Mixed

Not specified

Approach to DIF/measurement invariance analysis — states if the authors did not consider the source of DIF/lack of measurement invariance at all in their research, if the authors mentioned the source of DIF/lack of measurement invariance in their work, but did not collect systematic empirical evidence or conduct specific statistical analyses to examine the source of DIF/lack of measurement invariance, or if the authors systematically investigated sources of DIF/lack of measurement invariance with empirical evidence and/or statistical analyses.

No source(s) of DIF/lack of measurement invariance

No DIF found/measurement invariance met

Mention of source(s) of DIF/lack of measurement invariance

Investigation about source(s) of DIF/lack of measurement invariance

DIF detection/invariance testing method(s) used — based on Hidalgo & Gómez-Benito's (2010) classification. Some strategies are susceptible to deletion in case no studies used them.

Parametric*

Logistic regression / Ordinal logistic regression (OLR)

Log-linear models

Latent class models

IRT log-likelihood ratio test (IRTLR)

Lord chi-square statistic

Area measures

IRT Differential Item Functioning Test (DIFT)

Confirmatory Factor Analysis (CFA)

Multiple-Indicator Multiple-Cause (MIMIC)

Multilevel Analysis

Non-parametric*

Mantel-Haenszel (MH)

Standardization

SIBTEST / Poly-SIBTEST / CATSIB

Characteristics of DIF detection/invariance testing method(s) used

Justifications for using each method (pros)*

Limitations of each method (cons)*

Explanations for gender DIF/lack of measurement invariance — main characteristics of analyses used to obtain explanations for gender DIF/lack of measurement invariance.

*Qualitative method(s) used**

*Other methods**

Characteristics of method(s) used

Justifications for using each method (pros)*

Limitations of each method (cons)*

*Causes of gender DIF/lack of measurement invariance**

Supplementary Material III. Outcomes.

Primary Outcomes

Relating to the first objective: the primary outcome for this objective is the identification of the DIF detection/invariance testing techniques used, as reported by the author(s). If one study used more than one technique, all of them were extracted.

Relating to the second objective: the primary outcome for this objective is the identification of the qualitative methods used to obtain the possible explanations for gender DIF/lack of measurement invariance, as reported by the author(s). The non-qualitative methods used to obtain explanations for gender DIF/lack of measurement invariance were also extracted, in order to have more information.

Relating to the third objective: the primary outcomes for this objective are the description of the methodological aspects of the studies, the target variables for which DIF/measurement invariance is analyzed, and the characteristics of the instruments for which DIF/measurement invariance is analyzed. If one study assessed DIF or measurement invariance in more than one instrument, all of them were extracted.

Relating to the fourth objective: the primary outcomes for this objective are the justifications given by the author(s) for the use of each DIF detection/invariance testing technique and of each qualitative method, and the limitations given by the author(s) for each DIF detection/invariance testing technique and for each qualitative method.

Secondary Outcomes

The secondary outcomes are the bibliometric indicators, and causes of gender DIF/lack of measurement invariance reported by authors that examine such causes.

Supplementary Material IV-VII, XII. Data extraction and Critical Appraisal.

Available for download here (Nvivo, SPSS and Excel files): <https://osf.io/rvqk7/>

Supplementary Material VIII. References of Included Articles.

1. Abu-Hilal, M.M., Al-Maamari, S., Al-Harthy, I., Stocket, J., & Al-Nabhani, H.Z. (2019). Achievement and Self-Concept Relation: An Evidence of the Adequacy of the I/E Model with a Short Version of the Self-Description Questionnaire-II among School Students in Oman. *Italian Journal of Sociology of Education*, 11(3), 194-211. DOI: [10.14658/pupj-ijse-2019-3-11](https://doi.org/10.14658/pupj-ijse-2019-3-11)
2. Alivernini, F., Manganelli, S., Cavicchiolo, E., Chirico, A., & Lucidi, F. (2018). The Use of Self-Regulated Cognitive Strategies Across Students With Different Immigrant Backgrounds and Gender. *Journal of Psychoeducational Assessment*, 37(5), 652-664. DOI: [10.1177/0734282918785072](https://doi.org/10.1177/0734282918785072)
3. Ardic, E.O. & Gelbal, S. (2017). Cross—Group Equivalence of Interest and Motivation Items in PISA 2012 Turkey Sample. *Eurasian Journal of Educational Research*, 17(68), 223-240. DOI: [10.14689/ejer.2017.68.12](https://doi.org/10.14689/ejer.2017.68.12)
4. Bedyńska, S., Rycielski, P., & Jabłońska, M. (2021). Measuring Stereotype Threat at Math and Language Arts in Secondary School: Validation of a Questionnaire. *Frontiers in Psychology*, 12, 553964. DOI: [10.3389/fpsyg.2021.553964/full](https://doi.org/10.3389/fpsyg.2021.553964/full)
5. Bielinski, J. & Davison, M.L. (2001). A Sex Difference by Item Difficulty Interaction in Multiple-Choice Mathematics Items Administered to National Probability Samples. *Journal of Educational Measurement*, 38(1), 51-77. DOI: [10.1111/j.1745-3984.2001.tb01116.x](https://doi.org/10.1111/j.1745-3984.2001.tb01116.x)

6. Bofah, E.A. & Hannula, M.S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-Scale Assessments in Education*, 3(4), 1-36. DOI: [10.1186/s40536-015-0014-y](https://doi.org/10.1186/s40536-015-0014-y)
7. Bru, E., Virtanen, T., Kjetilstad, V., & Niemiec, C.P. (2019). Gender Differences in the Strength of Association between Perceived Support from Teachers and Student Engagement. *Scandinavian Journal of Educational Research*, 65(1), 153-168. DOI: [10.1080/00313831.2019.1659404](https://doi.org/10.1080/00313831.2019.1659404)
8. Brunner, M., Gogol, K.M., Sonnleitner, P., Keller, U., Krauss, S., & Preckel, F. (2013). Gender differences in the mean level, variability, and profile shape of student achievement: Results from 41 countries. *Intelligence*, 41(5), 378-395. DOI: [10.1016/j.intell.2013.05.009](https://doi.org/10.1016/j.intell.2013.05.009)
9. Calderón, F. & González, J. (2021). Analyses of the Factor Structure and Item Measurement Bias of a School Climate Scale in Chilean Students. *Frontiers in Education*, 6, 659398. DOI: [10.3389/feduc.2021.659398](https://doi.org/10.3389/feduc.2021.659398)
10. 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](https://doi.org/10.1016/j.stueduc.2019.100819)
11. 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](https://doi.org/10.3390/educsci11050201)
12. Çelik, M. & Özkan, Y.Ö. (2020). An Investigation of Differential Item Functioning in PISA 2015 Math Subtest in Terms of The Educational Zones and Gender. *Journal of*

- Measurement and Evaluation in Education and Psychology*, 11(3), 283-301. DOI: [10.21031/epod.715020](https://doi.org/10.21031/epod.715020)
13. Chaffee, K.E. & Plante, I. (2022). How Parents' Stereotypical Beliefs Relate to Students' Motivation and Career Aspirations in Mathematics and Language Arts. *Frontiers in Psychology*, 12, 796073. DOI: [10.3389/fpsyg.2021.796073](https://doi.org/10.3389/fpsyg.2021.796073)
14. 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](https://doi.org/10.1177/2158244019844085)
15. Cheema, J.R. (2017). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, 16(2), 152-166. DOI: [10.1080/17405629.2017.1358607](https://doi.org/10.1080/17405629.2017.1358607)
16. 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](https://doi.org/10.1080/10627197.2014.903650)
17. 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](https://doi.org/10.1177/0013164410387387)
18. 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](https://doi.org/10.3102/1076998609353111)
19. Demirbağ, C., Arikan, S., & Muğaloğlu, E.Z. (2020). Adaptation of the Self-efficacy Beliefs in STEM Education Scale and Testing Measurement Invariance across Groups.

- Journal of Measurement and Evaluation and Education and Psychology*, 11(2), 163-179. DOI: [10.21031/epod.675240](https://doi.org/10.21031/epod.675240)
20. Doudeen, H.M. & Annabi, H.A. (2008). Sex-Related Differential Item Functioning (DIF) Analysis of TIMSS. *Educational Sciences*, 35, 697-705. Retrieved from <https://journals.ju.edu.jo/DirasatHum/article/view/1807>
21. 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](https://doi.org/10.1007/s10763-020-10130-5)
22. Grover, R.K. & Ercikan, K. (2017). For Which Boys and Which Girls Are Reading Assessment Items Biased Against? Detection of Differential Item Functioning in Heterogeneous Gender Populations. *Applied Measurement in Education*, 30(3), 178-195. DOI: [10.1080/08957347.2017.1316276](https://doi.org/10.1080/08957347.2017.1316276)
23. Harris, A.M. & Carlton, S.T. (1993). Patterns of Gender Differences on Mathematics Items on the Scholastic Aptitude Test. *Applied Measurement in Education*, 6(2), 137-151. DOI: [10.1207/s15324818ame0602_3](https://doi.org/10.1207/s15324818ame0602_3)
24. 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](https://doi.org/10.1016/j.compedu.2017.06.003)
25. Heafner, T.L. & Fitchett, P.G. (2017). US history content knowledge and associated effects of race, gender, wealth, and urbanity: Item Response Theory (IRT) modeling of NAEP-USH achievement. *The Journal of Social Studies Research*, 42(1), 11-25. DOI: [10.1016/j.jssr.2017.01.001](https://doi.org/10.1016/j.jssr.2017.01.001)

26. 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(1), 177-188. DOI: [10.1007/s10964-020-01345-4](https://doi.org/10.1007/s10964-020-01345-4)
27. Hofer, S.I., Schumacher, R., & Rubin, H. (2017). The test of basic Mechanics Conceptual Understanding (bMCU): Using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics. *International Journal of STEM Education*, 4(18), 1-20. DOI: [10.1186/s40594-017-0080-5](https://doi.org/10.1186/s40594-017-0080-5)
28. Huang, T.W., Wu, P.C., & Mok, M.M.C. (2022). Examination of Gender-Related Differential Item Functioning Through Poly-BW Indices. *Frontiers in Psychology*, 13, 821459. DOI: [10.3389/fpsyg.2022.821459](https://doi.org/10.3389/fpsyg.2022.821459)
29. 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](https://doi.org/10.1111/ssm.12269)
30. 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](https://doi.org/10.1016/j.lindif.2013.12.003)
31. Jin, K.Y., Reichert, F., Cagasan, L.P., de la Torre, J., & Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring test dimensionality and performance differences. *Computers & Education*, 157, 103968. DOI: [10.1016/j.compedu.2020.103968](https://doi.org/10.1016/j.compedu.2020.103968)
32. Kaiser, T., Oberrauch, L., & Seeber, G. (2020). Measuring economic competence of secondary school students in Germany. *The Journal of Economic Education*, 51(3-4), 227-242. DOI: [10.1080/00220485.2020.1804504](https://doi.org/10.1080/00220485.2020.1804504)

33. 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](https://doi.org/10.1177/0734282910391623)
34. 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](https://doi.org/10.1080/15305058.2013.877911)
35. 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](https://doi.org/10.1016/j.childyouth.2017.06.058)
36. 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](https://doi.org/10.1080/02671522.2019.1633557)
37. Lane, S., Wang, N., & Magone, M. (1996). Gender-Related Differential Item Functioning on a Middle-School Mathematics Performance Assessment. *Educational Measurement: Issues and Practice*, 15(4), 21-27. DOI: [10.1111/j.1745-3992.1996.tb00575.x](https://doi.org/10.1111/j.1745-3992.1996.tb00575.x)
38. Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S. (2016). Differential Performance on National Exams: Evaluating Item and Bundle Functioning Methods using English, Mathematics, and Science Assessments. *SAGE Open*, 6(2), 215824401665379. DOI: [10.1177/2158244016653791](https://doi.org/10.1177/2158244016653791)
39. Lau, W.W.F. & Yuen, A.H.K. (2015). Factorial invariance across gender of a perceived ICT literacy scale. *Learning and Individual Differences*, 41, 79-85. DOI: [10.1016/j.lindif.2015.06.001](https://doi.org/10.1016/j.lindif.2015.06.001)

40. Le, L.T. (2009). Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items. *International Journal of Testing*, 9(2), 122-133. DOI: [10.1080/15305050902880769](https://doi.org/10.1080/15305050902880769)
41. 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](https://doi.org/10.1080/15305058.2014.922567)
42. Liou, P.Y. & Lin, J.J.H. (2021). Comparisons of Science Motivational Beliefs of Adolescents in Taiwan, Australia, and the United States: Assessing the Measurement Invariance Across Countries and Genders. *Frontiers in Psychology*, 12, 674902. DOI: [10.3389/fpsyg.2021.674902](https://doi.org/10.3389/fpsyg.2021.674902)
43. Lohbeck, A., Schultheiß, J., Petermann, U., & Petermann, F. (2016). Factor structure and measurement invariance of the Students' Self-report Checklist of Social and Learning Behaviour (SSL). *Learning and Individual Differences*, 51, 314-321. DOI: [10.1016/j.lindif.2016.08.038](https://doi.org/10.1016/j.lindif.2016.08.038)
44. 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>
45. 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](https://doi.org/10.26858/ijole.v2i1.4346)
46. 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](https://doi.org/10.1207/s15324818ame1904_4)

47. Mireles-Ríos, R., Simon, O., & Nylund-Gibson, K. (2020). The Critical Role of Teacher Emotional Support for Latinx Students. *Teachers College Record: The Voice of Scholarship in Education*, 122(12), 1-32. DOI: [10.1177/016146812012201209](https://doi.org/10.1177/016146812012201209)
48. Mucherah, W., Finch, W.H., & Keaikitse, S. (2012). Differential Bundle Functioning Analysis of the Self-Description Questionnaire Self-Concept Scale for Kenyan Female and Male Students Using the MIMIC Model. *International Journal of Testing*, 12(1), 78-99. DOI: [10.1080/15305058.2011.620724](https://doi.org/10.1080/15305058.2011.620724)
49. 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](https://doi.org/10.1002/pits.22330)
50. Oberrauch, L. & Kaiser, T. (2019). Economic competence in early secondary school: Evidence from a large-scale assessment in Germany. *International Review of Economics Education*, 35, 100172. DOI: [10.1016/j.iree.2019.100172](https://doi.org/10.1016/j.iree.2019.100172)
51. Ozdemir, B. & Alshamrani, A.H. (2020). Examining the Fairness of Language Test Across Gender with IRT-based Differential Item and Test Functioning Methods. *International Journal of Learning, Teaching and Educational Research*, 19(6), 27-45. DOI: [10.26803/ijlter.19.6.2](https://doi.org/10.26803/ijlter.19.6.2)
52. Pabian, S. & Vandebosch, H. (2015). Short-term longitudinal relationships between adolescents' (cyber)bullying perpetration and bonding to school and teachers. *International Journal of Behavioral Development*, 40(2), 162-172. DOI: [10.1177/0165025415573639](https://doi.org/10.1177/0165025415573639)

53. Pae, T.I. (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](https://doi.org/10.1177/0265532211434027)
54. Penner, A.M. (2003). International gender $\times$ item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, 95(3), 650-655. DOI: [10.1037/0022-0663.95.3.650](https://doi.org/10.1037/0022-0663.95.3.650)
55. Pomplun, M. & Omar, M.H. (2001). Do Reading Passages About War Provide Factorially Invariant Scores for Men and Women? *Applied Measurement in Education*, 14(2), 171-189. DOI: [10.1207/S15324818AME1402_4](https://doi.org/10.1207/S15324818AME1402_4)
56. Primi, C., Donati, M.A., Chiesi, F., & Morsanyi, K. (2017). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*, 24(2), 258-279. DOI: [10.1080/13546783.2017.1387606](https://doi.org/10.1080/13546783.2017.1387606)
57. 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](https://doi.org/10.1007/s10826-015-0182-y)
58. Rohatgi, A., Scherer, R., & Hatlevik, O.E. (2016). The role of ICT self-efficacy for students' ICT use and their achievement in a computer and information literacy test. *Computers & Education*, 102, 103-116. DOI: [10.1016/j.compedu.2016.08.001](https://doi.org/10.1016/j.compedu.2016.08.001)
59. Romine, W.L., Sadler, T.D., & Wulff, E.P. (2017). Conceptualizing Student Affect for Science and Technology at the Middle School Level: Development and Implementation of a Measure of Affect in Science and Technology (MAST). *Journal of Science Education and Technology*, 26(5), 534-545. DOI: [10.1007/s10956-017-9697-x](https://doi.org/10.1007/s10956-017-9697-x)

60. Ryan, K.E. & Fan, M. (1996). Examining Gender DIF on a Multiple-Choice Test of Mathematics: A Confirmatory Approach. *Educational Measurement: Issues and Practice*, 15(4), 15-20. DOI: [10.1111/j.1745-3992.1996.tb00574.x](https://doi.org/10.1111/j.1745-3992.1996.tb00574.x)
61. Schroeders, U. & Jansen, M. (2020). Science Self-Concept – More than the Sum of Its Parts? *The Journal of Experimental Education*, 90(2), 435-451. DOI: [10.1080/00220973.2020.1740967](https://doi.org/10.1080/00220973.2020.1740967)
62. Schwabe, F., McElvany, N., & Trendtel, M. (2014). The School Age Gender Gap in Reading Achievement: Examining the Influences of Item Format and Intrinsic Reading Motivation. *Reading Research Quarterly*, 50(2), 219-232. DOI: [10.1002/rrq.92](https://doi.org/10.1002/rrq.92)
63. Schweder, S. & Raufelder, D. (2019). Positive emotions, learning behavior and teacher support in self-directed learning during adolescence: Do age and gender matter? *Journal of Adolescence*, 73(1), 73-84. DOI: [10.1016/j.adolescence.2019.04.004](https://doi.org/10.1016/j.adolescence.2019.04.004)
64. 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](https://doi.org/10.1080/10904018.2015.1065747)
65. Shamsuddin, H., Razak, N.A., Thien, L.M., & Khairani, A.Z. (2020). Do Boys and Girls Interpret Mathematics Test Items Similarly? Insights from Rasch Model Analysis. *Asia Pacific Journal of Educators and Education*, 35(1), 17-36. DOI: [10.21315/apjee2020.35.1.2](https://doi.org/10.21315/apjee2020.35.1.2)
66. Shanmugan, S.K.S. (2018). Determining Gender Differential Item Functioning for Mathematics in Coeducational School Culture. *Malaysian Journal of Learning and Instruction*, 15(2), 83-109. DOI: [10.32890/mjli2018.15.2.4](https://doi.org/10.32890/mjli2018.15.2.4)

67. Shin, J., Lee, H., McCarthy-Donovan, A., Hwang, H., Yim, S., & Seo, E. (2015). Home and Motivational Factors Related to Science-Career Pursuit: Gender differences and gender similarities. *International Journal of Science Education*, 37(9), 1478-1503. DOI: [10.1080/09500693.2015.1042941](https://doi.org/10.1080/09500693.2015.1042941)
68. Shin, S., Rachmatullah, A., Roshayanti, F., Ha, M., & Lee, J.K. (2018). Career motivation of secondary students in STEM: A cross-cultural study between Korea and Indonesia. *International Journal for Educational and Vocational Guidance*, 18(2), 203-231. DOI: [10.1007/s10775-017-9355-0](https://doi.org/10.1007/s10775-017-9355-0)
69. Steegh, A., Höffler, T., Höft, L., & Parchmann, I. (2020). First steps toward gender equity in the chemistry Olympiad: Understanding the role of implicit GENDER-SCIENCE stereotypes. *Journal of Research in Science Teaching*, 58(1), 40-68. DOI: [10.1002/tea.21645](https://doi.org/10.1002/tea.21645)
70. Susongko, P., Arfiani, Y., & Kusuma, M. (2021). Determination of Gender Differential Item Functioning in Tegal Students' Scientific Literacy Skills with Integrated Science (SLiSIS) Test Using Rasch Model. *Jurnal Pendidikan IPA Indonesia*, 10(2), 270-281. DOI: [10.15294/jpii.v10i2.26775](https://doi.org/10.15294/jpii.v10i2.26775)
71. Syafrial, Ashadi, Saputro, S., & Sarwanto (2022). Trend Creative Thinking Perception of Students in Learning Natural Science: Gender and Domicile Perspective. *International Journal of Instruction*, 15(1), 701-716. DOI: [10.29333/iji.2022.15140a](https://doi.org/10.29333/iji.2022.15140a)
72. Taylor, C.S. & Lee, Y. (2012). Gender DIF in Reading and Mathematics Tests With Mixed Item Formats. *Applied Measurement in Education*, 25(3), 246-280. DOI: [10.1080/08957347.2012.687650](https://doi.org/10.1080/08957347.2012.687650)

73. Tayyaba, S. (2010). Mathematics achievement in middle school level in Pakistan: Findings from the First National Assessment. *International Journal of Educational Management*, 24(3), 221-249. DOI: [10.1108/09513541011031583](https://doi.org/10.1108/09513541011031583)
74. Tsai, L.T. & Chang, C.C. (2018). Measuring ocean literacy of high school students: Psychometric properties of a Chinese version of the ocean literacy scale. *Environmental Education Research*, 25(2), 264-279. DOI: [10.1080/13504622.2018.1542487](https://doi.org/10.1080/13504622.2018.1542487)
75. 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](https://doi.org/10.3389/fpsyg.2020.00622)
76. Tvedt, M.S., Bru, E., & Idsoe, T. (2019). Perceived Teacher Support and Intentions to Quit Upper Secondary School: Direct, and Indirect Associations via Emotional Engagement and Boredom. *Scandinavian Journal of Educational Research*, 65(1), 101-122. DOI: [10.1080/00313831.2019.1659401](https://doi.org/10.1080/00313831.2019.1659401)
77. Walstad, W.B. & Robson, D. (1997). Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. *The Journal of Economic Education*, 28(2), 155-171. DOI: [10.1080/00220489709595917](https://doi.org/10.1080/00220489709595917)
78. Wang, F., Leung, S.O., & Jiang, C. (2021). Psychometric Properties of Cognitive and Metacognitive Learning Strategies Scale Among Chinese Senior Secondary School Students. *Journal of Psychoeducational Assessment*, 39(6), 761-771. DOI: [10.1177/07342829211011806](https://doi.org/10.1177/07342829211011806)
- 79; 80. Wirthwein, L., Sparfeldt, J. R., Heyder, A., Buch, S. R., Rost, D.H., & Steinmayr, R. (2019). Sex differences in achievement goals: Do school subjects matter? *European*

Journal of Psychology of Education, 35(2), 403-427. DOI: [10.1007/s10212-019-00427-](https://doi.org/10.1007/s10212-019-00427-7)

[7](https://doi.org/10.1007/s10212-019-00427-7)

81. 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](https://doi.org/10.7334/psicothema2018.303)
82. Xia, M., Fosco, G.M., & Feinberg, M.E. (2015). Examining reciprocal influences among family climate, school attachment, and academic self-regulation: Implications for school success. *Journal of Family Psychology*, 30(4), 442-452. DOI: [10.1037/fam0000141](https://doi.org/10.1037/fam0000141)
83. Xu, J. (2017). Homework Expectancy Value Scale for high school students: Measurement invariance and latent mean differences across gender and grade level. *Learning and Individual Differences*, 60, 10-17. DOI: [10.1016/j.lindif.2017.10.003](https://doi.org/10.1016/j.lindif.2017.10.003)
84. 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](https://doi.org/10.12738/estp.2018.2.0317)
85. Young, E.L. & Sudweeks, R.R. (2005). Gender Differential Item Functioning in the Multidimensional Self Concept Scale With a Sample of Early Adolescent Students. *Measurement and Evaluation in Counseling and Development*, 38(1), 29-44. DOI: [10.1080/07481756.2005.11909767](https://doi.org/10.1080/07481756.2005.11909767)
86. Yu, J. & McLellan, R. (2019). Beyond academic achievement goals: The importance of social achievement goals in explaining gender differences in self-handicapping. *Learning and Individual Differences*, 69, 33-44. DOI: [10.1016/j.lindif.2018.11.010](https://doi.org/10.1016/j.lindif.2018.11.010)

87. 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](https://doi.org/10.1080/10627197.2004.9652959)

Supplementary Material IX. Study Characteristics.

Table IX.1.

Bibliometric Indicators of Articles.

	Citations	First author's country	Journal	Rank (JCR)
Abu-Hilal et al. (2019)	1	Oman	Italian Journal of Sociology of Education	Not indexed
Alivernini et al. (2018)	12	Italy	Journal of Psychoeducational Assessment	Q3
Ardic & Gelbal (2017)	2	Turkey	Eurasian Journal of Educational Research	Q3
Bedynska et al. (2021)	3	Poland	Frontiers in Psychology	Q1
Bielinski & Davison (2001)	35	USA	Journal of Educational Measurement	Q3
Bofah & Hannula (2015)	13	Finland	Large-scale Assessments in Education	Q1
Bru et al. (2019)	4	Norway	Scandinavian Journal of Educational Research	Q2
Brunner et al. (2013)	20	Germany	Intelligence	Q2
Calderón & González (2021)	0	Chile	Frontiers in Education	Q2
Cascella et al. (2020)	7	UK	Studies in Educational Evaluation	Q2
Cascella et al. (2021)	1	UK	Education Sciences	Q1
Çelik & Özkan (2020)	1	Turkey	Journal of Measurement and Evaluation in Education and Psychology	Q4
Chaffee & Plante (2022)	0	Canada	Frontiers in Psychology	Q1
Chang (2019)	7	Taiwan	SAGE Open	Q2
Cheema (2017)	11	USA	European Journal of Developmental Psychology	Q4
Chen & Jiao (2014)	20	USA	Educational Assessment	Q2

	Citations	First author's country	Journal	Rank (JCR)
Cheng et al. (2011)	39	Taiwan	Educational and Psychological	Q1
Cho & Cohen (2010)	75	USA	Journal of Educational and Behavioral Statistics	Q2
Demirbağ et al. (2020)	1	Turkey	Journal of Measurement and Evaluation in Education and Psychology	Q4
Doudeen & Annabi (2008)	0	United Arab Emirates	Dirasat, Educational Sciences	Not indexed
Ferretti & Giberti (2020)	5	Italy	International Journal of Science and Mathematics Education	Q1
Grover & Ercikan (2017)	7	Canada	Applied Measurement in Education	Q3
Harris & Carlton (1993)	60	USA	Applied Measurement in Education	Q3
Hatlevik et al. (2017)	18	Norway	Computers & Education	Q1
Heafner & Fitchett (2017)	1	USA	The Journal of Social Studies Research	Not indexed
Heyder et al. (2020)	7	Germany	Journal of Youth and Adolescence	Q1
Hofer et al. (2017)	8	Germany	International Journal of STEM Education	Q1
Huang et al. (2022)	0	Taiwan	Frontiers in Psychology	Q1
Innabi & Dodeen (2018)	4	Sweden	School Science and Mathematics	Q3
Jansen et al. (2014)	100	Germany	Learning and Individual Differences	Q2
Jin et al. (2020)	23	China	Computers & Education	Q1
Kaiser et al. (2020)	4	Germany	The Journal of Economic Education	Q3

	Citations	First author's country	Journal	Rank (JCR)
Kalaycioğlu & Berberoğlu (2011)	11	Turkey	Journal of Psychoeducational Assessment	Q3
Kan & Bulut (2014)	9	Turkey	International Journal of Testing	Q1
Kaye-Tzadok et al. (2017)	41	Israel	Children and Youth Services Review	Q1
Korpershoek et al. (2019)	5	The Netherlands	Research Papers in Education	Q1
Lane et al. (1996)	40	USA	Educational Measurement Issues and Practice	Q4
Latifi al. (2016)	1	Canada	SAGE Open	Q2
Lau & Yuen (2015)	15	China	Learning and Individual Differences	Q2
Le (2009)	38	Australia	International Journal of Testing	Q1
Lee & Geisinger (2014)	7	USA	International Journal of Testing	Q1
Liou & Lin (2021)	1	South Korea	Frontiers in Psychology	Q1
Lohbeck et al. (2016)	1	Germany	Learning and Individual Differences	Q2
Lyons-Thomas et al. (2014)	14	Canada	Education and Science	Q3
Mahmud & Nur (2018)	6	Indonesia	International Journal of Language Education	Q4
Mendes-Barnett & Ercikan (2006)	59	Canada	Applied Measurement in Education	Q3
Mireles-Ríos et al. (2020)	0	USA	Teachers College Record	Q3
Mucherah et al. (2012)	1	USA	International Journal of Testing	Q1
Nalipay et al. (2019)	9	China	Psychology in the Schools	Q3
Oberrauch & Kaiser (2019)	7	Germany	International Review of Economics Education	Q2

	Citations	First author's country	Journal	Rank (JCR)
Ozdemir & Alshamrani (2020)	1	Saudi Arabia	International Journal of Learning, Teaching and Educational Research	Not indexed
Pabian & Vandebosch (2015)	27	Belgium	International Journal of Behavioral Development	Q3
Pae (2012)	21	South Korea	Language Testing	Q1
Penner (2003)	69	USA	Journal of Educational Psychology	Q1
Pomplun & Omar (2001)	4	USA	Applied Measurement in Education	Q3
Primi et al. (2017)	25	Italy	Thinking & Reasoning	Q2
Raufelder et al. (2015)	34	Germany	Journal of Child and Family Studies	Q2
Rohatgi et al. (2016)	124	Norway	Computers & Education	Q1
Romine et al. (2017)	1	USA	Journal of Science Education and Technology	Q1
Ryan & Fan (1996)	0	USA	Educational Measurement Issues and Practice	Q4
Schroeders & Jansen (2020)	3	Germany	The Journal of Experimental Education	Q2
Schwabe et al. (2014)	71	Germany	Reading Research Quarterly	Q1
Schweder & Raufelder (2019)	22	Germany	Journal of Adolescence	Q2
Seo et al. (2016)	6	USA	International Journal of Listening	Not indexed
Shamsuddin et al. (2020)	0	Malaysia	Asia Pacific Journal of Educators and Education	Not indexed
Shanmugam (2018)	1	Malaysia	Malaysian Journal of Learning and Instruction	Q3

	Citations	First author's country	Journal	Rank (JCR)
Shin et al. (2015)	16	South Korea	International Journal of Science Education	Q1
Shin et al. (2018)	26	South Korea	International Journal of Educational and Vocational Guidance	Q2
Steegeh et al. (2020)	8	Germany	Journal of Research in Science Teaching	Q1
Susongko et al. (2021)	1	Indonesia	Jurnal Pendidikan IPA Indonesia	Not indexed
Syafrial et al. (2022)	1	Indonesia	International Journal of Instruction	Q2
Taylor & Lee (2012)	29	USA	Applied Measurement in Education	Q3
Tayyaba (2010)	9	UK	International Journal of Educational Management	Q3
Tsai & Chang (2018)	8	Taiwan	Environmental Education Research	Q1
Tsaousis et al. (2020)	0	USA	Frontiers in Psychology	Q1
Tvedt et al. (2019)	11	Norway	Scandinavian Journal of Educational Research	Q2
Walstad & Robson (1997)	97	USA	The Journal of Economic Education	Q3
Wang et al. (2021)	1	China	Journal of Psychoeducational Assessment	Q3
Wirthwein et al. (2019)	8	Germany	European Journal of Psychology and Education	Q2
Woitschach et al. (2019)	6	Canada	Psicothema	Q1
Xia et al. (2015)	27	USA	Journal of Family Psychology	Q1
Xu (2017)	17	USA	Learning and Individual Differences	Q1

	Citations	First author's country	Journal	Rank (JCR)
Yildirim & Büyüköztürk (2018)	4	Turkey	Educational Sciences Theory & Practice	Q4
Young & Sudweeks (2005)	14	USA	Measurement and Evaluation in Counseling and Development	Q3
Yu & McLellan (2019)	12	UK	Learning and Individual Differences	Q2
Zenisky et al. (2004)	36	USA	Educational Assessment	Q2

Table IX.2.

Methodological Aspects of Studies.

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Abu-Hilal et al. (2019)	700	-	Oman	Yes	QUAN	Self-concept
Alivernini et al. (2018)	263,683	15.60	Italy	No	QUAN	Self-regulation
Ardic & Gelbal (2017)	3,124	-	Turkey	No	QUAN	Motivation in Mathematics
Bedynska et al. (2021)	1,241	14.56	Poland	Yes	QUAN	Stereotype threat at Mathematics and Language Art lessons
Bielinski & Davison (2001)	86,437	-	USA	No	QUAN	Mathematics achievement
Bofah & Hannula (2015)	38,806	15.58	Multi-country	No	QUAN	Value, confidence and like learning Mathematics; teacher and parental support

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Bru et al. (2019)	1,265	-	Norway	Yes	QUAN	Teacher support and school engagement
Brunner et al. (2013)	275,369	-	Multi-country	No	QUAN	Mathematics achievement
Calderón & González (2021)	176,126	-	Chile	No	QUAN	School climate
Cascella et al. (2020)	56,061	-	Italy	No	MIXED	Mathematics achievement
Cascella et al. (2021)	1,647	-	Italy	Yes	QUAN	Mathematics achievement
Çelik & Özkan (2020)	2,409	-	Turkey	No	QUAN	Mathematics achievement
Chaffee & Plante (2022)	170	16.15	Canada	Yes	QUAN	Motivation in Mathematics and Language; parents' masculinity and femininity beliefs; parents' ability stereotypes
Chang (2019)	1,027	16.79	Taiwan	Yes	QUAN	Ocean literacy
Cheema (2017)	485,490	-	Multi-country	No	QUAN	Science literacy
Chen & Jiao (2014)	5,433	-	USA	No	QUAN	Reading literacy
Cheng et al. (2011)	3,941	15.94	Taiwan	Yes	QUAN	Bullying
Cho & Cohen (2010) (Only study 2)	39,614	-	USA	No	QUAN	Mathematics achievement
Demirbağ et al. (2020)	326	-	Turkey	Yes	QUAN	Self-efficacy in STEM

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Doudeen & Annabi (2008)	5,299	-	Jordan	No	QUAN	Mathematics achievement
Ferretti & Giberti (2020)	41,830	-	Italy	No (QUAN) / Yes (QUAL)	MIXED	Mathematics achievement
Grover & Ercikan (2017)	23,207	-	Canada	No	QUAN	Reading literacy
Harris & Carlton (1993)	379,896	-	USA	No	QUAN	Mathematics achievement
Hatlevik et al. (2017)	919	-	Norway	Yes	QUAN	ICT literacy
Heafner & Fitchett (2017)	1,670	-	USA	No	QUAN	History literacy
Heyder et al. (2020)	877	16.09	Germany	No	QUAN	Motivation, self-concept, and fixed mindset in Mathematics and Language
Hofer et al. (2017)	380	16.16	Switzerland	Yes	QUAN	Physics literacy
Huang et al. (2022)	1,439	-	Taiwan	Yes	QUAN	Mathematics achievement
Innabi & Dodeen (2018)	7,774	13.80	Jordan	No	QUAN	Mathematics achievement
Jansen et al. (2014)	6,036	15.49	Germany	Yes	QUAN	Self-concept in Science
Jin et al. (2020)	1,274	-	China	Yes	QUAN	ICT literacy
Kaiser et al. (2020)	6,230	14.59	Germany	Yes	QUAN	Economic literacy
Kalaycioğlu & Berberoğlu (2011)	35,372	-	Turkey	No	QUAN	Science literacy and Mathematics achievement

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Kan & Bulut (2014)	671	-	Turkey	Yes	QUAN	Mathematics achievement
Kaye-Tzadok et al. (2017)	19,182	-	Multi-country	No	QUAN	Subjective well-being and satisfaction with self, school, peers, and family.
Korpershoek et al. (2019)	11,657	14.66	Multi-country	No	QUAN	School motivation
Lane et al. (1996)	3,946	-	USA	Yes	QUAN	Mathematics achievement
Latifi al. (2016)	5,557	-	Pakistan	No	QUAN	Mathematics achievement, and Physics and English literacy
Lau & Yuen (2015)	825	13.16	China	Yes	QUAN	ICT literacy
Le (2009)	83,000	-	Multi-country	No	QUAN	Science literacy
Lee & Geisinger (2014)	1,753	-	South Korea	Yes	QUAN	English literacy
Liou & Lin (2021)	22,673	-	Multi-country	No	QUAN	Self-concept, beliefs about the utility value, and beliefs about intrinsic value of Science
Lohbeck et al. (2016)	4,045	13.14	Germany	Yes	QUAN	Social and learning behavior
Lyons-Thomas et al. (2014)	21,670	-	Multi-country	No	QUAN	Mathematics achievement

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Mahmud & Nur (2018)	84	-	Indonesia	Yes	MIXED	Language learning strategies
Mendes-Barnett & Ercikan (2006)	9,000	-	Canada	No	QUAN	Mathematics achievement
Mireles-Ríos et al. (2020)	783	16.01	USA	Yes	QUAN	Teacher support and perceived discrimination
Mucherah et al. (2012)	1,990	16.50	Kenya	Yes	QUAN	Self-concept
Nalipay et al. (2019)	84,429	15.75	Multi-country	No	QUAN	Students' and parents' enjoyment of reading
Oberrauch & Kaiser (2019)	1,687	13.88	Germany	Yes	QUAN	Economic literacy
Ozdemir & Alshamrani (2020)	1,000	-	Saudi Arabia	No	QUAN	English literacy
Pabian & Vandebosch (2015)	2,128	12.70	Belgium	Yes	QUAN	Bullying
Pae (2012)	44,300	-	South Korea	Yes	QUAN	English literacy
Penner (2003)	1,616	-	Multi-country	Yes	QUAN	Science literacy and Mathematics achievement
Pomplun & Omar (2001)	4,000	-	USA	No	QUAN	Reading literacy
Primi et al. (2017) (Only study 1)	838	15.37	Italy	Yes	QUAN	Cognitive reflection
Raufelder et al. (2015)	1,088	13.70	Germany	Yes	QUAN	Test anxiety, school engagement, and parental support

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Rohatgi et al. (2016)	2,410	14.80	Norway	No	QUAN	Self-efficacy in ICT, ICT use and ICT literacy
Romine et al. (2017)	138	-	USA	Yes	QUAN	Interest in Science careers
Ryan & Fan (1996)	6,277	-	USA	No	QUAN	Mathematics achievement
Schroeders & Jansen (2020)	1,229	15.40	Germany	Yes	QUAN	Self-concept in Science
Schwabe et al. (2014)	4,974	15.83	Germany	No	QUAN	Reading literacy
Schweder & Raufelder (2019)	759	13.56	Germany	Yes	QUAN	Self-directed learning
Seo et al. (2016)	1,233	-	Multi-country	No	QUAN	English literacy
Shamsuddin et al. (2020)	158	-	Malaysia	Yes	QUAN	Mathematics achievement
Shanmugam (2018)	905	-	China	Yes	QUAN	Mathematics achievement
Shin et al. (2015)	5,176	-	South Korea	No	QUAN	Home environments, Science motivation, and interest in Science careers
Shin et al. (2018)	2,171	-	Multi-country	Yes	QUAN	STEM career motivation
Steegh et al. (2020)	445	16.50	Germany	Yes	QUAN	Self-concept in Science and interest in Chemistry
Susongko et al. (2021)	310	-	Indonesia	No	QUAN	Science literacy

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Syafrial et al. (2022)	62	-	Indonesia	Yes	QUAN	Creative thinking during Science learning
Taylor & Lee (2012)	60,000	-	USA	No	QUAN	Reading literacy and Mathematics achievement
Tayyaba (2010)	14,440	-	Pakistan	Yes	QUAN	Mathematics achievement
Tsai & Chang (2018)	1,027	16.79	Taiwan	Yes	QUAN	Ocean literacy
Tsaousis et al. (2020)	6,265	-	Saudi Arabia	Yes	QUAN	Academic achievement
Tvedt et al. (2019)	1,379	-	Norway	Yes	QUAN	Intentions to quit school, emotional engagement with school, boredom in school, teachers' feedback quality and support, and autonomy granting
Walstad & Robson (1997)	4,123	-	USA	No	QUAN	Economic literacy
Wang et al. (2021)	698	17.10	China	Yes	QUAN	Cognitive and metacognitive learning strategies
Wirthwein et al. (2019) (Study 1)	425	16.60	Germany	Yes	QUAN	Achievement goals in school and in Mathematics and German

	<i>N</i>	Mean age	Country of participants	Data collection	Type of study	Target variables
Wirthwein et al. (2019) (Study 2)	1,210	14.36	Germany	Yes	QUAN	Achievement goals in Mathematics, German, English, Physics, Chemistry and History
Woitschach et al. (2019)	12,657	-	Multi-country	No	QUAN	Science literacy
Xia et al. (2015)	979	11.30/ 11.90/ 13.00/ 13.90/ 14.90	USA	Yes	QUAN	Home environment, self-regulation and school attachment
Xu (2017)	1,799	-	China	No	QUAN	Value of Mathematics and expectancy to Mathematics assignments
Yildirim & Büyüköztürk (2018)	1,063,570	-	Turkey	No	MIXED	Mathematics achievement
Young & Sudweeks (2005)	400	-	USA	Yes	QUAN	Self-concept
Yu & McLellan (2019)	536	-	UK	No	QUAN	Achievement goals in specific subjects, academic self-handicapping, and social achievement goals
Zenisky et al. (2004)	360,000 (aprox.)	-	USA	No	QUAN	Science literacy

Note. Sample size and mean age provided only for adolescent participants in studies with parents or teachers.

Supplementary Material X. Target Variables.

Table X.1.

Frequency of Studies that Used Each Target Variable for DIF or Invariance Analyses.

Cognitive	
	<i>f</i>
Mathematics achievement	24
Reading literacy	5
English literacy	5
Science literacy	7
Physics literacy	2
Ocean literacy	2
Economic literacy	3
History literacy	1
Information and Communication Technology (ICT) literacy	4
Academic achievement (general)	1
Non-cognitive	
	<i>f</i>
School attachment	1
School climate	1
Bullying	2
Perceived discrimination	1
Teacher support	4
Teachers' feedback quality	1
Intentions to quit school	1
Emotional engagement with school	1
Boredom in school	1
Subjective well-being	1
Satisfaction with self	1
Satisfaction with family	1
Satisfaction with school	1
Satisfaction with peers	1
Social and learning behavior	1
Social achievement goals	1
Autonomy granting	1
Cognitive reflection	1
Cognitive and metacognitive learning strategies	1
Language learning strategies	1
Creative thinking during Science learning	1
Self-directed learning	1
Self-regulation (academic)	2
Self-concept (general)	3
Self-concept in Science	4
Self-concept in Mathematics	1

	<i>f</i>
Self-concept in Language	1
Self-efficacy in STEM	1
Self-efficacy in ICT	1
Confidence in Mathematics	1
Expectancy to Mathematics assignments	1
Achievement goals in school	1
Achievement goals in specific subjects	2
Fixed mindset in Mathematics	1
Fixed mindset in Language	1
School engagement	2
School motivation	1
STEM career motivation	1
Mathematics motivation	3
Science motivation	1
Language motivation	2
Interest in Science careers	2
Interest in Chemistry	1
Like learning Mathematics	1
Enjoyment of Reading	1
Value of Mathematics	2
Intrinsic value of Science	1
Utility value of Science	1
ICT use for different purposes	1
Test anxiety	1
Academic self-handicapping	1
Stereotype threat at Mathematics	1
Stereotype threat at Language Art	1
Parents' masculinity beliefs	1
Parents' femininity beliefs	1
Parents' ability stereotypes	1
Parental support	2
Home environments	2

Supplementary Material XI. Risk of Bias Assessment.

Table XI.1.

Risk of Bias in Studies.

	Reporting quality	Methodological quality	Overall risk of bias
Abu-Hilal et al. (2019)	Low risk	Low risk	Low risk
Alivernini et al. (2018)	Medium risk	Low risk	Low risk
Ardic & Gelbal (2017)	Medium risk	Low risk	Medium risk
Bedynska et al. (2021)	Low risk	Low risk	Low risk
Bielinski & Davison (2001)	Low risk	Medium risk	Medium risk
Bofah & Hannula (2015)	Low risk	Low risk	Low risk
Bru et al. (2019)	Low risk	Low risk	Low risk
Brunner et al. (2013)	Low risk	Low risk	Low risk
Calderón & González (2021)	Medium risk	Low risk	Low risk
Cascella et al. (2020)	Low risk	High risk	Medium risk
Cascella et al. (2021)	Low risk	Low risk	Low risk
Çelik & Özkan (2020)	Low risk	Medium risk	Medium risk
Chaffee & Plante (2022)	Medium risk	Low risk	Medium risk
Chang (2019)	Low risk	Low risk	Low risk
Cheema (2017)	Low risk	Low risk	Low risk
Chen & Jiao (2014)	Low risk	Low risk	Low risk
Cheng et al. (2011)	Low risk	Low risk	Low risk
Cho & Cohen (2010) (Only study 2)	Medium risk	Low risk	Medium risk
Demirbağ et al. (2020)	Medium risk	Low risk	Medium risk
Doudeen & Annabi (2008)	Medium risk	Low risk	Low risk
Ferretti & Giberti (2020)	Low risk	Low risk	Low risk
Grover & Ercikan (2017)	Low risk	Low risk	Low risk
Harris & Carlton (1993)	Medium risk	Low risk	Medium risk
Hatlevik et al. (2017)	Low risk	Low risk	Low risk
Heafner & Fitchett (2017)	Low risk	Low risk	Low risk
Heyder et al. (2020)	Low risk	Low risk	Low risk
Hofer et al. (2017)	Medium risk	Low risk	Low risk
Huang et al. (2022)	Low risk	Low risk	Low risk
Innabi & Dodeen (2018)	Low risk	Low risk	Low risk
Jansen et al. (2014)	Low risk	Low risk	Low risk
Jin et al. (2020)	Low risk	Low risk	Low risk
Kaiser et al. (2020)	Low risk	Low risk	Low risk
Kalaycioğlu & Berberoğlu (2011)	Low risk	Low risk	Low risk
Kan & Bulut (2014)	Low risk	Low risk	Low risk
Kaye-Tzadok et al. (2017)	Low risk	Low risk	Low risk
Korpershoek et al. (2019)	Medium risk	Low risk	Medium risk
Lane et al. (1996)	High risk	Low risk	Medium risk
Latifi al. (2016)	Low risk	Low risk	Low risk

	Reporting quality	Methodological quality	Overall risk of bias
Lau & Yuen (2015)	Medium risk	Low risk	Low risk
Le (2009)	Low risk	Low risk	Low risk
Lee & Geisinger (2014)	Medium risk	Low risk	Low risk
Liou & Lin (2021)	Low risk	Low risk	Low risk
Lohbeck et al. (2016)	Low risk	Low risk	Low risk
Lyons-Thomas et al. (2014)	Medium risk	Low risk	Medium risk
Mahmud & Nur (2018)	Low risk	Medium risk	Medium risk
Mendes-Barnett & Ercikan (2006)	Medium risk	Low risk	Medium risk
Mireles-Ríos et al. (2020)	Low risk	Low risk	Low risk
Mucherah et al. (2012)	Medium risk	Low risk	Low risk
Nalipay et al. (2019)	Medium risk	Low risk	Medium risk
Oberrauch & Kaiser (2019)	Medium risk	Low risk	Low risk
Ozdemir & Alshamrani (2020)	Low risk	Low risk	Low risk
Pabian & Vande Bosch (2015)	Medium risk	Low risk	Low risk
Pae (2012)	Low risk	Low risk	Low risk
Penner (2003)	Medium risk	Medium risk	Medium risk
Pomplun & Omar (2001)	Medium risk	Low risk	Medium risk
Primi et al. (2017) (Only study 1)	Medium risk	Low risk	Medium risk
Raufelder et al. (2015)	Low risk	Low risk	Low risk
Rohatgi et al. (2016)	Medium risk	Low risk	Medium risk
Romine et al. (2017)	High risk	Low risk	Medium risk
Ryan & Fan (1996)	Medium risk	Low risk	Low risk
Schroeders & Jansen (2020)	Medium risk	Low risk	Low risk
Schwabe et al. (2014)	Low risk	Low risk	Low risk
Schweder & Raufelder (2019)	Low risk	Low risk	Low risk
Seo et al. (2016)	Medium risk	Low risk	Medium risk
Shamsuddin et al. (2020)	Medium risk	Medium risk	Medium risk
Shanmugam (2018)	Medium risk	Low risk	Low risk
Shin et al. (2015)	Medium risk	Low risk	Low risk
Shin et al. (2018)	Medium risk	Low risk	Low risk
Steeh et al. (2020)	Low risk	Low risk	Low risk
Susongko et al. (2021)	Low risk	Medium risk	Low risk
Syafrial et al. (2022)	Medium risk	Medium risk	Medium risk
Taylor & Lee (2012)	Low risk	Low risk	Low risk
Tayyaba (2010)	Medium risk	Low risk	Low risk
Tsai & Chang (2018)	Low risk	Low risk	Low risk
Tsaousis et al. (2020)	Low risk	Low risk	Low risk
Tvedt et al. (2019)	Medium risk	Low risk	Low risk
Walstad & Robson (1997)	Medium risk	Low risk	Low risk
Wang et al. (2021)	High risk	Low risk	Medium risk
Wirthwein et al. (2019) (Study 1)	Medium risk	Low risk	Medium risk
Wirthwein et al. (2019) (Study 2)	Medium risk	Low risk	Medium risk

	Reporting quality	Methodological quality	Overall risk of bias
Woitschach et al. (2019)	Low risk	Low risk	Low risk
Xia et al. (2015)	Low risk	Low risk	Low risk
Xu (2017)	Medium risk	Low risk	Low risk
Yildirim & Büyüköztürk (2018)	Low risk	Medium risk	Low risk
Young & Sudweeks (2005)	Medium risk	Low risk	Low risk
Yu & McLellan (2019)	Low risk	Low risk	Low risk
Zenisky et al. (2004)	Medium risk	Low risk	Low risk

Supplementary Material XIII. DIF/Measurement Invariance Analyses and Instruments' Characteristics.

Table XIII.1.

DIF Detection/Invariance Testing Method(s) Used by Each Study and Instruments' Characteristics.

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Abu-Hilal et al. (2019)	700	Multigroup CFA	36	Poly.	Closed-ended	Not specified
Alivernini et al. (2018)	263,683	Multigroup CFA	12	Poly.	Closed-ended	Not specified
Ardic & Gelbal (2017)	3,124	Multigroup CFA; IRTLR; MH; poly-SIBTEST	8	Poly.	Closed-ended	Not specified
Bedynska et al. (2021)*	1,241	Multigroup CFA	7 / 7	Poly.	Closed-ended	Computer-based
Bielinski & Davison (2001)**	19,790 / 10,973 / 55,674	IRT difficulty parameters comparison	225 / 124 / 81	Not specified	Closed-ended	Not specified
Bofah & Hannula (2015)****	38,806	Multigroup CFA	5 / 6 / 9 / 4 / 4	Poly.	Closed-ended	Not specified
Bru et al. (2019)*	1,265	Multigroup CFA	13 / 9	Poly.	Closed-ended	Computer-based
Brunner et al. (2013)	275,369	Multigroup CFA	-	Not specified	Not specified	Paper-based
Calderón & González (2021)	176,126	Cognitive diagnosis models	46	Both	Closed-ended	Not specified
Cascella et al. (2020)	56,061	IRT + ANOVA	5	Dicho.	Both	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Cascella et al. (2021)***	400 / 400 / 400 / 400 (aprox.)	IRT + ANOVA	31 / 31 / 31 / 31	Dicho.	Closed-ended	Not specified
Çelik & Özkan (2020)	2,409	IRT Rasch analysis	63	Dicho.	Closed-ended	Computer-based
Chaffee & Plante (2022)****	170	Multigroup CFA	11 / 11 / 6 / 6 / 20	Poly. (4) / Not specified (1)	Closed-ended	Paper-based (2) / Mixed (3)
Chang (2019)	1,027	Multigroup CFA; MIMIC	48	Dicho.	Both	Paper-based
Cheema (2017)	485,490	Logistic regression	50	Dicho.	Closed-ended	Not specified
Chen & Jiao (2014) (Only study 2)	5,433	LCA; MH	93	Dicho.	Closed-ended	Not specified
Cheng et al. (2011)**	3,941	IRTLR	19 / 19 / 19	Poly.	Closed-ended	Not specified
Cho & Cohen (2010)	39,614	Mixture IRT model	38	Dicho.	Not specified	Not specified
Demirbağ et al. (2020)	326	Multigroup CFA	12	Poly.	Closed-ended	Not specified
Doudeen & Annabi (2008)	5,299	MH	124	Not specified	Closed-ended	Not specified
Ferretti & Giberti (2020)	41,812 (QUAN) / 18 (QUAL)	IRT Rasch analysis	45	Not specified	Closed-ended	Not specified
Grover & Ercikan (2017)*****	23,207	LCA; LH method; MH	Between 14-59	Dicho.	Both	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Harris & Carlton (1993)*****	379,896	MH	60 / 60 / 60 / 60 / 60 / 60	Dicho.	Closed-ended	Not specified
Hatlevik et al. (2017)	919	Multigroup CFA; MIMIC	14	Dicho.	Closed-ended	Computer-based
Heafner & Fitchett (2017)	1,670	Logistic regression; area measures	12	Dicho.	Closed-ended	Not specified
Heyder et al. (2020)*****	877	Multigroup SEM	4 / 4 / 4 / 4 / 2 / 2	Poly.	Closed-ended	Not specified
Hofer et al. (2017)*	239 / 141	IRT Rasch analysis	22 / 12	Dicho.	Closed-ended	Not specified
Huang et al. (2022)	1,439	Poly-SIBTEST	34	Poly.	Open-ended	Not specified
Innabi & Dodeen (2018)	7,774	MH	197	Dicho.	Both	Not specified
Jansen et al. (2014)**	6,036	Multigroup CFA	6 / 6 / 6	Poly.	Closed-ended	Not specified
Jin et al. (2020)	1,274	IRT difficulty parameters comparison	80	Dicho.	Both	Computer-based
Kaiser et al. (2020)	6,230	IRT difficulty parameters comparison	30	Dicho.	Both	Computer-based
Kalaycioğlu & Berberoğlu (2011)*	35,372	Logistic regression; IRTLRL; MH; Multigroup CFA	30 / 30	Dicho.	Closed-ended	Not specified
Kan & Bulut (2014)*	671	Logistic regression; EIRM	25 / 25	Dicho.	Closed-ended	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Kaye-Tzadok et al. (2017)****	19,182	Multigroup SEM	4 / 9 / 3 / 2 / 3	Poly.	Closed-ended	Mixed
Korpershoek et al. (2019)	11,657	Multigroup CFA	32	Poly.	Closed-ended	Paper-based
Lane et al. (1996)	3,946	Logistic discriminant function analysis	42	Poly.	Open-ended	Not specified
Latifi al. (2016)**	2,067 / 1,756 / 1,734	SIBTEST; EIRM	25 / 30 / 25	Dicho.	Closed-ended	Paper-based
Lau & Yuen (2015)	825	Multigroup CFA	17	Poly.	Closed-ended	Computer-based
Le (2009)	83,000	Lord chi-square statistic	210	Both	Both	Not specified
Lee & Geisinger (2014)	1,753	Logistic regression; MH	33	Dicho.	Closed-ended	Not specified
Liou & Lin (2021)**	22,673	Multigroup CFA	8 / 9 / 9	Poly.	Closed-ended	Not specified
Lohbeck et al. (2016)	4,045	Multigroup CFA	40	Poly.	Closed-ended	Not specified
Lyons-Thomas et al. (2014)	21,670	Area measures; LH method	35	Dicho.	Both	Not specified
Mendes-Barnett & Ercikan (2006)	9,000	SIBTEST	45	Dicho.	Closed-ended	Not specified
Mireles-Ríos et al. (2020)*	783	Multigroup CFA	6 / 11	Poly.	Closed-ended	Not specified
Mucherah et al. (2012)	1,990	MIMIC	135	Poly.	Closed-ended	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Nalipay et al. (2019)	84,429	Multigroup CFA/SEM	6 / 3	Poly.	Closed-ended	Not specified
Oberrauch & Kaiser (2019)	1,687	Lord chi-square	30	Dicho.	Both	Not specified
Ozdemir & Alshamrani (2020)	1,000	Multigroup CFA; Lord chi-square	80	Dicho.	Not specified	Not specified
Pabian & Vandebosch (2015)*	2,128	Multilevel analysis	2 / 2	Poly.	Closed-ended	Paper-based
Pae (2012)**	14,000 / 15,000 / 15,000	IRTLR; MH	38 / 33 / 33	Dicho.	Closed-ended	Not specified
Penner (2003)*	896 / 720	Logits of the gender-specific item-difficulty scores	45 / 36	Dicho.	Closed-ended	Not specified
Pomplun & Omar (2001)	4,000	Multigroup CFA	22	Poly.	Closed-ended	Not specified
Primi et al. (2017) (Only study 1)	838	IRTLR	6	Dicho.	Open-ended	Paper-based
Raufelder et al. (2015)**	1,088	Multigroup CFA/SEM	12 / 10 / 8	Poly.	Closed-ended	Not specified
Rohatgi et al. (2016)**	2,410	Multigroup SEM	13 / 8 / 80	Poly. (1) / Dicho. (1) / Not specified (1)	Closed-ended	Computer-based
Romine et al. (2017)	138	Logistic regression	5	Poly.	Closed-ended	Not specified
Ryan & Fan (1996)***	1,569 / 1,569 / 1,569 / 1,569 (aprox.)	SIBTEST	35 / 35 / 35 / 35	Dicho.	Closed-ended	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Schroeders & Jansen (2020)	1,229	Multigroup CFA	16	Poly.	Closed-ended	Not specified
Schwabe et al. (2014)	4,974	GLMM	100	Dicho.	Both	Not specified
Schweder & Raufelder (2019)**	759	Multigroup CFA/SEM	4 / 5 / 4	Poly.	Closed-ended	Not specified
Seo et al. (2016)	1,233	MH	20	Dicho.	Closed-ended	Not specified
Shamsuddin et al. (2020)	158	IRT Rasch analysis	43	Poly.	Open-ended	Paper-based
Shanmugam (2018)	905	MH	24	Dicho.	Closed-ended	Not specified
Shin et al. (2015)*****	5,176	Multigroup SEM	Between 2-8	Poly. (7) / Dicho. (1)	Closed-ended	Not specified
Shin et al. (2018)	2,171	IRT Rasch analysis	32	Poly.	Closed-ended	Not specified
Steegh et al. (2020)*	445	Multigroup CFA	6 / 4	Poly.	Closed-ended	Computer-based
Susongko et al. (2021)	310	Cognitive diagnosis models	42	Dicho.	Closed-ended	Not specified
Syafrial et al. (2022)	62	IRT Rasch analysis	18	Poly.	Closed-ended	Not specified
Taylor & Lee (2012)*****	15,000 / 15,000 / 15,000 / 15,000	IRT Rasch analysis; poly-SIBTEST	Between 34-46	Poly.	Both	Not specified
Tayyaba (2010)	14,440	IRT Rasch analysis	56	Dicho.	Both	Not specified

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Tsai & Chang (2018)	1,027	IRT Rasch analysis; IRT difficulty parameters comparison	48	Both	Closed-ended	Paper-based
Tsaousis et al. (2020)	6,265	LCA; MIMIC	65	Not specified	Closed-ended	Not specified
Tvedt et al. (2019)*****	1,379	Multigroup SEM	Between 3-5	Poly. (2) / Not specified (4)	Closed-ended (2) / Not specified (4)	Computer-based
Walstad & Robson (1997)	4,123	Area measures	46	Dicho.	Closed-ended	Not specified
Wang et al. (2021)	698	Multigroup CFA	31	Poly.	Closed-ended	Not specified
Wirthwein et al. (2019) (Study 1)**	425	Multigroup CFA	31 / 31 / 31	Poly.	Closed-ended	Not specified
Wirthwein et al. (2019) (Study 2)*****	1,210	Multigroup CFA	31 / 31 / 31 / 31 / 31 / 31	Poly.	Closed-ended	Not specified
Woitschach et al. (2019)	12,657	Logistic regression	92	Dicho.	Both	Not specified
Xia et al. (2015)***	979	Multigroup SEM	13 / 4 / 5 / Not specified.	Poly.	Closed-ended	Paper-based
Xu (2017)	1,799	Multigroup CFA	8	Poly.	Closed-ended	Not specified
Yildirim & Büyüköztürk (2018)	1,063,570	Logistic regression; MH	20	Not specified	Closed-ended	Not specified
Young & Sudweeks (2005)	400	MH; SMD	149	Poly.	Closed-ended	Not specified
Yu & McLellan (2019)**	536	Multigroup CFA	14 / 8 / 6	Poly.	Closed-ended	Paper-based

	<i>N</i>	DIF/invariance method(s)	N° items	Scoring format	Response format	Admin. mode
Zenisky et al. (2004)***	360,000 (aprox.)	SMD	Between 32-42	Both	Both	Not specified

Note. * = Studies with two instruments; ** = Studies with three instruments; *** = Studies with four instruments; **** = Studies with five instruments; ***** = Studies with six or more instruments; Sample size provided only for adolescent participants in studies with parents or teachers; Admin. Mode = Administration mode; CFA = Confirmatory Factor Analysis; IRTL = Item Response Theory Likelihood Ratio; MH = Mantel-Haenszel; SIBTEST = Simultaneous Item Bias Test; MIMIC = Multiple-Indicator Multiple-Cause; LCA = Latent Class Analysis; SEM = Structural Equation Modeling; EIRM = Explanatory Item Response Modeling; GLMM = Generalized Linear Mixed Models; SMD = Standardized Mean Difference; Poly. = Polytomous; Dicho. = Dichotomous.

Supplementary Material XIV. DIF/Measurement Invariance Approaches.

Table XIV.1.

DIF/Measurement Invariance Approach Followed by Each Study.

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Abu-Hilal et al. (2019)	700	QUAN	Multigroup CFA	A*	No
Alivernini et al. (2018)	263,683	QUAN	Multigroup CFA	A*	No
Ardic & Gelbal (2017)	3,124	QUAN	Multigroup CFA; IRTLR; MH; poly-SIBTEST	B	No
Bedynska et al. (2021)	1,241	QUAN	Multigroup CFA	A*	No
Bielinski & Davison (2001)	86,437	QUAN	IRT difficulty parameters comparison	A	No
Bofah & Hannula (2015)	38,806	QUAN	Multigroup CFA	A*	No
Bru et al. (2019)	1,265	QUAN	Multigroup CFA	A*	No
Brunner et al. (2013)	275,369	QUAN	Multigroup CFA	A*	No
Calderón & González (2021)	176,126	QUAN	Cognitive diagnosis models	A	No
Cascella et al. (2020)	56,061	MIXED	IRT + ANOVA	C	Didactical interpretation of results
Cascella et al. (2021)	1,647	QUAN	IRT + ANOVA	C	Comparison of distractor response curves

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Çelik & Özkan (2020)	2,409	QUAN	IRT Rasch analysis	A	No
Chaffee & Plante (2022)	170	QUAN	Multigroup CFA	A*	No
Chang (2019)	1,027	QUAN	Multigroup CFA; MIMIC	A*	No
Cheema (2017)	485,490	QUAN	Logistic regression	A	No
Chen & Jiao (2014) (Only study 2)	5,433	QUAN	LCA; MH	C	Overlap between latent classes and manifest groups (Sources of latent DIF)
Cheng et al. (2011)	3,941	QUAN	IRTLR	A	No
Cho & Cohen (2010)	39,614	QUAN	Mixture IRT model	C	Multilevel analysis of latent groups (Sources of latent DIF)
Demirbağ et al. (2020)	326	QUAN	Multigroup CFA	A	No
Doudeen & Annabi (2008)	5,299	QUAN	MH	C	Content analysis
Ferretti & Giberti (2020)	41,812 (QUAN) / 18 (QUAL)	MIXED	IRT Rasch analysis	C	Interviews
Grover & Ercikan (2017)	23,207	QUAN	LCA; LH method; MH	C	Multinomial logistic regression
Harris & Carlton (1993)	379,896	QUAN	MH	C	ANOVA
Hatlevik et al. (2017)	919	QUAN	Multigroup CFA; MIMIC	B	No

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Heafner & Fitchett (2017)	1,670	QUAN	Logistic regression; area measures	A	No
Heyder et al. (2020)	877	QUAN	Multigroup SEM	C	Multigroup SEM
Hofer et al. (2017)	380	QUAN	IRT Rasch analysis	A	No
Huang et al. (2022)	1,439	QUAN	Poly-SIBTEST	C	Poly-BW indices
Innabi & Dodeen (2018)	7,774	QUAN	MH	C	Content analysis
Jansen et al. (2014)	6,036	QUAN	Multigroup CFA	A*	No
Jin et al. (2020)	1,274	QUAN	IRT difficulty parameters comparison	A	No
Kaiser et al. (2020)	6,230	QUAN	IRT difficulty parameters comparison	A	No
Kalaycioğlu & Berberoğlu (2011)	35,372	QUAN	Logistic regression; IRTLR; MH; Multigroup CFA	C	Content analysis
Kan & Bulut (2014)	671	QUAN	Logistic regression; EIRM	C	Content analysis
Kaye-Tzadok et al. (2017)	19,182	QUAN	Multigroup SEM	A*	No
Korpershoek et al. (2019)	11,657	QUAN	Multigroup CFA	A*	No
Lane et al. (1996)	3,946	QUAN	Logistic discriminant function analysis	C	Analytic Scoring Analysis
Latifi al. (2016)	5,557	QUAN	SIBTEST; EIRM	A	No
Lau & Yuen (2015)	825	QUAN	Multigroup CFA	A	No

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Le (2009)	83,000	QUAN	Lord chi-square statistic	C	Content analysis
Lee & Geisinger (2014)	1,753	QUAN	Logistic regression; MH	C	Propensity scores
Liou & Lin (2021)	22,673	QUAN	Multigroup CFA	A*	No
Lohbeck et al. (2016)	4,045	QUAN	Multigroup CFA	A*	No
Lyons-Thomas et al. (2014)	21,670	QUAN	Area measures; LH method	B	No
Mahmud & Nur (2018)	72 (QUAN) / 12 (QUAL)	MIXED	Not applicable	Not applicable	Interviews
Mendes-Barnett & Ercikan (2006)	9,000	QUAN	SIBTEST	C	Conf. Approach
Mireles-Ríos et al. (2020)	783	QUAN	Multigroup CFA	C	OLS regression models
Mucherah et al. (2012)	1,990	QUAN	MIMIC	A	No
Nalipay et al. (2019)	84,429	QUAN	Multigroup CFA/SEM	C	Multigroup SEM
Oberrauch & Kaiser (2019)	1,687	QUAN	Lord chi-square	A	No
Ozdemir & Alshamrani (2020)	1,000	QUAN	Multigroup CFA / Lord chi-square	C	Content analysis
Pabian & Vandebosch (2015)	2,128	QUAN	Multilevel analysis	C	Multilevel analysis
Pae (2012)	44,300	QUAN	IRTLR; MH	C	Expert appraisal + Multiple regression analysis

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Penner (2003)	1,616	QUAN	Logits of the gender-specific item-difficulty scores	A	No
Pomplun & Omar (2001)	4,000	QUAN	Multigroup CFA	B	No
Primi et al. (2017) (Only study 1)	838	QUAN	IRTLR	A*	No
Raufelder et al. (2015)	1,088	QUAN	Multigroup CFA/SEM	C	Multigroup SEM
Rohatgi et al. (2016)	2,410	QUAN	Multigroup SEM	A*	No
Romine et al. (2017)	138	QUAN	Logistic regression	A*	No
Ryan & Fan (1996)	6,277	QUAN	SIBTEST	A	No
Schroeders & Jansen (2020)	1,229	QUAN	Multigroup CFA	A	No
Schwabe et al. (2014)	4,974	QUAN	GLMM	C	GLMM
Schweder & Raufelder (2019)	759	QUAN	Multigroup CFA/SEM	C	Multigroup SEM
Seo et al. (2016)	1,233	QUAN	MH	A*	No
Shamsuddin et al. (2020)	158	QUAN	IRT Rasch analysis	A	No
Shanmugam (2018)	905	QUAN	MH	B	No
Shin et al. (2015)	5,176	QUAN	Multigroup SEM	A*	No
Shin et al. (2018)	2,171	QUAN	IRT Rasch analysis	B	No
Steeh et al. (2020)	445	QUAN	Multigroup CFA	A*	No

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Susongko et al. (2021)	310	QUAN	Cognitive diagnosis models	C	Content analysis
Syafrial et al. (2022)	62	QUAN	IRT Rasch analysis	A	No
Taylor & Lee (2012)	60,000	QUAN	IRT Rasch analysis; poly-SIBTEST	C	Content analysis
Tayyaba (2010)	14,440	QUAN	IRT Rasch analysis	A	No
Tsai & Chang (2018)	1,027	QUAN	IRT Rasch analysis; IRT difficulty parameters comparison	A	No
Tsaousis et al. (2020)	6,265	QUAN	LCA; MIMIC	C	Relating classes to covariates (Sources of latent DIF)
Tvedt et al. (2019)	1,379	QUAN	Multigroup SEM	A*	No
Walstad & Robson (1997)	4,123	QUAN	Area measures	A	No
Wang et al. (2021)	698	QUAN	Multigroup CFA	A*	No
Wirthwein et al. (2019) (Study 1)	425	QUAN	Multigroup CFA	A*	No
Wirthwein et al. (2019) (Study 2)	1,210	QUAN	Multigroup CFA	A*	No
Woitschach et al. (2019)	12,657	QUAN	Logistic regression	C	Multilevel analysis
Xia et al. (2015)	979	QUAN	Multigroup SEM	A*	No
Xu (2017)	1,799	QUAN	Multigroup CFA	A*	No
Yildirim & Büyüköztürk (2018)	1,063,570	MIXED	Logistic regression; MH	C	Focus groups; Delphi

	<i>N</i>	Type of study	DIF detection method(s)	Approach (Category)	Explanation
Young & Sudweeks (2005)	400	QUAN	MH; SMD	A	No
Yu & McLellan (2019)	536	QUAN	Multigroup CFA	A*	No
Zenisky et al. (2004)	360,000 (aprox.)	QUAN	SMD	C	Content analysis

Note. Sample size provided only for adolescent participants in studies with parents or teachers; Gen. of DIF = Generation of DIF; Explanation = Explanations for DIF/lack of measurement invariance strategies; CFA = Confirmatory Factor Analysis; IRTLRL = Item Response Theory Likelihood Ratio; MH = Mantel-Haenszel; SIBTEST = Simultaneous Item Bias Test; MIMIC = Multiple-Indicator Multiple-Cause; LCA = Latent Class Analysis; SEM = Structural Equation Modeling; EIRM = Explanatory Item Response Modeling; GLMM = Generalized Linear Mixed Models; SMD = Standardized Mean Difference; A* = No examination of sources of DIF because no DIF/invariance found; A = No examination of sources of DIF/lack of measurement invariance; B = Mention of sources of DIF/lack of measurement invariance; C = Examination of sources of DIF/lack of measurement invariance; Conf. approach = Confirmatory approach; OLS = Ordinal Logistic Regression.

Supplementary Material XV. DIF Detection/Invariance Testing Method X Instruments' Characteristics and Sample Size: Cross Tables.

Table XV.1.

DIF Detection/Invariance Testing Method X Scoring Format.

	Dichotomous	Polytomous	Both	TOTAL
Multigroup CFA	4 (13.80%)	25 (86.20%)	0 (0.00%)	29
Multigroup SEM	0 (0.00%)	7 (77.78%)	2 (22.22%)	9
Logistic regression or Ordinal logistic regression	6 (85.72%)	1 (14.28%)	0 (0.00%)	7
Latent class models	2 (100%)	0 (0.00%)	0 (0.00%)	2
Multilevel analysis	0 (0.00%)	1 (100%)	0 (0.00%)	1
MIMIC	2 (66.67%)	1 (33.33%)	0 (0.00%)	3
LH method	2 (100%)	0 (0.00%)	0 (0.00%)	2
Lord chi-square statistic	2 (66.67%)	0 (0.00%)	1 (33.33%)	3
Area measures	3 (100%)	0 (0.00%)	0 (0.00%)	3
IRT log-likelihood test	3 (60.00%)	2 (30.00%)	0 (0.00%)	5
Mixture IRT model	1 (100%)	0 (0.00%)	0 (0.00%)	1
Explanatory Item Response Modeling	2 (100%)	0 (0.00%)	0 (0.00%)	2
IRT Rasch analysis	3 (37.50%)	4 (50.00%)	1 (12.50%)	8
IRT + ANOVA	2 (100%)	0 (0.00%)	0 (0.00%)	2
Cognitive diagnosis models	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
IRT difficulty parameters comparison	2 (66.67%)	0 (0.00%)	1 (33.33%)	3
Logistic discriminant function analysis	0 (0.00%)	1 (100%)	0 (0.00%)	1
Generalized Linear Mixed Models	1 (100%)	0 (0.00%)	0 (0.00%)	1
Logits of the gender-specific item-difficulty scores	1 (100%)	0 (0.00%)	0 (0.00%)	1
Standardized Mean Difference	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
Mantel-Haenszel (MH)	9 (81.82%)	2 (18.18%)	0 (0.00%)	11
SIBTEST or poly-SIBTEST	3 (50.00%)	3 (50.00%)	0 (0.00%)	6
TOTAL	49	48	7	104

Note. Studies with no specification of scoring format for any instrument are not included. CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modeling; MIMIC = Multiple-Indicator Multiple-Cause; SIBTEST = Simultaneous Item Bias Test.

Table XV.2.

DIF Detection/Invariance Testing Method X Response Format.

	Closed-ended	Open-ended	Both	TOTAL
Multigroup CFA	27 (96.43%)	0 (0.00%)	1 (3.57%)	28
Multigroup SEM	9 (100%)	0 (0.00%)	0 (0.00%)	9
Logistic regression or Ordinal logistic regression	7 (87.50%)	0 (0.00%)	1 (12.50%)	8
Latent class models	2 (66.67%)	0 (0.00%)	1 (33.33%)	3
Multilevel analysis	1 (100%)	0 (0.00%)	0 (0.00%)	1
MIMIC	3 (75.00%)	0 (0.00%)	1 (25.00%)	4
LH method	0 (0.00%)	0 (0.00%)	2 (100%)	2
Lord chi-square statistic	0 (0.00%)	0 (0.00%)	2 (100%)	2
Area measures	2 (66.67%)	0 (0.00%)	1 (33.33%)	3
IRT log-likelihood test	4 (80.00%)	1 (20.00%)	0 (0.00%)	5
Mixture IRT model*	-	-	-	-
Explanatory Item Response Modeling	2 (100%)	0 (0.00%)	0 (0.00%)	2
IRT Rasch analysis	6 (66.67%)	1 (11.11%)	2 (22.22%)	9
IRT + ANOVA	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
Cognitive diagnosis models	2 (100%)	0 (0.00%)	0 (0.00%)	2
IRT difficulty parameters comparison	2 (50.00%)	0 (0.00%)	2 (50.00%)	4
Logistic discriminant function analysis	0 (0.00%)	1 (100%)	0 (0.00%)	1
Generalized Linear Mixed Models	0 (0.00%)	0 (0.00%)	1 (100%)	1
Logits of the gender-specific item-difficulty scores	1 (100%)	0 (0.00%)	0 (0.00%)	1
Standardized Mean Difference	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
Mantel-Haenszel (MH)	11 (84.62%)	0 (0.00%)	2 (15.38%)	13
SIBTEST or poly-SIBTEST	4 (66.68%)	1 (16.66%)	1 (16.66%)	6
TOTAL	86	4	19	108

Note. Studies with no specification of response format for any instrument are not included; * = All studies that used this technique did not specify the response format; CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modeling; MIMIC = Multiple-Indicator Multiple-Cause; SIBTEST = Simultaneous Item Bias Test.

Table XV.3.

DIF Detection/Invariance Testing Method X Administration Mode.

	Paper-based	Computer-based	Mixed	TOTAL
Multigroup CFA	4 (40.00%)	5 (50.00%)	1 (10.00%)	10
Multigroup SEM	1 (25.00%)	2 (50.00%)	1 (25.00%)	4
Logistic regression or Ordinal logistic regression*	-	-	-	-
Latent class models*	-	-	-	-
Multilevel analysis	1 (100%)	0 (0.00%)	0 (0.00%)	1
MIMIC	1 (50.00%)	1 (50.00%)	0 (0.00%)	2
LH method*	-	-	-	-
Lord chi-square statistic*	-	-	-	-
Area measures*	-	-	-	-
IRT log-likelihood test	1 (100%)	0 (0.00%)	0 (0.00%)	1
Mixture IRT model*	-	-	-	-
Explanatory Item Response Modeling	1 (100%)	0 (0.00%)	0 (0.00%)	1
IRT Rasch analysis	2 (66.67%)	1 (33.33%)	0 (0.00%)	3
IRT + ANOVA*	-	-	-	-
Cognitive diagnosis models*	-	-	-	-
IRT difficulty parameters comparison	1 (33.33%)	2 (66.67%)	0 (0.00%)	3
Logistic discriminant function analysis*	-	-	-	-
Generalized Linear Mixed Models*	-	-	-	-
Logits of the gender-specific item-difficulty scores*	-	-	-	-
Standardized Mean Difference*	-	-	-	-
Mantel-Haenszel (MH)*	-	-	-	-
SIBTEST or poly-SIBTEST	1 (100%)	0 (0.00%)	0 (0.00%)	1
TOTAL	13	11	2	26

*Note. Studies with no specification of administration mode for any instrument are not included; * = All studies that used this technique did not specify the administration mode; CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modeling; MIMIC = Multiple-Indicator Multiple-Cause; SIBTEST = Simultaneous Item Bias Test.*

Table XV.4.

DIF Detection/Invariance Testing Method X Number of Items (Small $n < 10$; Medium $10 \leq n < 100$; Large $n \geq 100$).

	Small	Medium	Large	TOTAL
Multigroup CFA	15 (42.86%)	20 (57.14%)	0 (0.00%)	35
Multigroup SEM	9 (75.00%)	3 (25.00%)	0 (0.00%)	12
Logistic regression or Ordinal logistic regression	1 (12.50%)	7 (87.50%)	0 (0.00%)	8
Latent class models	0 (0.00%)	3 (100%)	0 (0.00%)	3
Multilevel analysis	1 (100%)	0 (0.00%)	0 (0.00%)	1
MIMIC	0 (0.00%)	3 (75.00%)	1 (25.00%)	4
LH method	0 (0.00%)	2 (100%)	0 (0.00%)	2
Lord chi-square statistic	0 (0.00%)	2 (66.67%)	1 (33.33%)	3
Area measures	0 (0.00%)	3 (100%)	0 (0.00%)	3
IRT log-likelihood test	2 (40.00%)	3 (60.00%)	0 (0.00%)	5
Mixture IRT model	0 (0.00%)	1 (100%)	0 (0.00%)	1
Explanatory Item Response Modeling	0 (0.00%)	2 (100%)	0 (0.00%)	2
IRT Rasch analysis	0 (0.00%)	9 (100%)	0 (0.00%)	9
IRT + ANOVA	1 (50.00%)	1 (50.00%)	0 (0.00%)	2
Cognitive diagnosis models	0 (0.00%)	2 (100%)	0 (0.00%)	2
IRT difficulty parameters comparison	0 (0.00%)	4 (80.00%)	1 (20.00%)	5
Logistic discriminant function analysis	0 (0.00%)	1 (100%)	0 (0.00%)	1
Generalized Linear Mixed Models	0 (0.00%)	0 (0.00%)	1 (100%)	1
Logits of the gender-specific item-difficulty scores	0 (0.00%)	1 (100%)	0 (0.00%)	1
Standardized Mean Difference	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
Mantel-Haenszel (MH)	1 (7.69%)	9 (69.23%)	3 (23.08%)	13
SIBTEST or poly-SIBTEST	1 (16.67%)	5 (83.33%)	0 (0.00%)	6
TOTAL	31	82	8	121

Note. Studies with no specification of number of items for any instrument are not included. CFA = Confirmatory Factor Analysis; SEM = Structural Equation Modeling; MIMIC = Multiple-Indicator Multiple-Cause; SIBTEST = Simultaneous Item Bias Test.

Table XV.5.

DIF Detection/Invariance Testing Method X Sample Size (Small $n < 100$; Medium $100 \leq n < 1000$; Large $n \geq 1000$).

	Small	Medium	Large	TOTAL
Multigroup CFA	0 (0.00%)	11 (36.67%)	19 (63.33%)	30
Multigroup SEM	0 (0.00%)	3 (33.33%)	6 (66.67%)	9
Logistic regression or Ordinal logistic regression	0 (0.00%)	2 (25.00%)	6 (75.00%)	8
Latent class models	0 (0.00%)	0 (0.00%)	3 (100%)	3
Multilevel analysis	0 (0.00%)	0 (0.00%)	1 (100%)	1
MIMIC	0 (0.00%)	1 (25.00%)	3 (75.00%)	4
LH method	0 (0.00%)	0 (0.00%)	2 (100%)	2
Lord chi-square statistic	0 (0.00%)	0 (0.00%)	3 (100%)	3
Area measures	0 (0.00%)	0 (0.00%)	3 (100%)	3
IRT log-likelihood test	0 (0.00%)	1 (20.00%)	4 (80.00%)	5
Mixture IRT model	0 (0.00%)	0 (0.00%)	1 (100%)	1
Explanatory Item Response Modeling	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
IRT Rasch analysis	1 (11.11%)	2 (22.22%)	6 (66.67%)	9
IRT + ANOVA	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
Cognitive diagnosis models	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
IRT difficulty parameters comparison	0 (0.00%)	0 (0.00%)	4 (100%)	4
Logistic discriminant function analysis	0 (0.00%)	0 (0.00%)	1 (100%)	1
Generalized Linear Mixed Models	0 (0.00%)	0 (0.00%)	1 (100%)	1
Logits of the gender-specific item-difficulty scores	0 (0.00%)	1 (100%)	0 (0.00%)	1
Standardized Mean Difference	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
Mantel-Haenszel (MH)	0 (0.00%)	2 (15.38%)	11 (84.62%)	13
SIBTEST or poly-SIBTEST	0 (0.00%)	0 (0.00%)	6 (100%)	6
TOTAL	1	27	84	112

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

Supplementary Material XVI. Characteristics of DIF Detection/Invariance Testing Methods.

Table XVI.1.

Pros and Cons of DIF Detection/Invariance Testing Methods Used and Most Common Instruments' Characteristics and Sample Size.

	Pros	Cons	Scoring format	Response format	Admin. mode	N. items	Sample size
Multigroup CFA	Meaningful interpretation of group differences Verification of shared measurement structure Examination of all aspects of heterogeneity across groups	Less parsimonious Full invariance is unrealistic Does not focus on specific items	Poly.	Closed-ended	Computer-based	Medium	Large
Multigroup SEM	Appropriate in latent variable systems in which the latent variable has a causal influence on observed variables Error-free measures of latent variables	Not provided	Poly.	Closed-ended	Computer-based	Small	Large
Logistic regression or Ordinal logistic regression	Admits non-unidimensionality Allows control for overall proficiency Comparable DIF estimates Admits low number of items Detects DIF across different ability levels Uniform and non-uniform DIF	Inflated Type-I error rate Insensitive effect size	Dicho.	Closed-ended	No info.	Medium	Large
Multilevel analysis	Not provided	Not provided	Poly.	Closed-ended	Paper-based	Small	Large
Mixture IRT model	Combines IRT, multilevel and latent class models	Not provided	Dicho.	Closed-ended	No info.	Medium	Large
Area measures	DIF can be graphically illustrated Estimates are not dependent on sample or test Uniform and non-uniform DIF	Require large sample size	Dicho.	Closed-ended	No info.	Medium	Large
IRT log-likelihood test	Uniform and non-uniform DIF	Problems with large sample size Requires model fit	Dicho.	Closed-ended	Paper-based	Medium	Large

	Pros	Cons	Scoring format	Response format	Admin. mode	N. items	Sample size
Latent class models	Identification of sources of DIF						
	Address DIF caused by an interaction of multiple variables and/or a non-explored manifest variable						
	Understanding of the nature of classes						
Latent class models	Identification of DIF across latent classes	Require unidimensionality	Dicho.	Closed-ended	No info.	Medium	Large
	Invariance of classes across groups						
	Allow for capturing qualitative heterogeneity among groups						
Latent class models	Exploration of within-group heterogeneity						
	Multiple predictor variables						
	DIF on several variables simultaneously						
MIMIC	Dichotomous and polytomous items						
	Uniform and non-uniform DIF						
	Admits small sample size						
MIMIC	Permits latent variables						
	Less estimated parameters						
	Item and bundle level						
MIMIC	Input matrix less demanding						
	High power in difficult situations	Not clear-cut findings on the most preferable estimator	Dicho.	Closed-ended	Both	Medium	Large
	Estimation of IRT parameters (internal validity)						
MIMIC	Relationship between the covariate and the latent variable (external validity)						
	DIF across different levels						
	Admits non-unidimensionality						
MIMIC	Admits non-local dependency						
	DIF detection in matrix-sampled data	Not provided	Dicho.	Both	No info.	Medium	Large
LH method	Dichotomous and polytomous items	Requires unidimensionality and model fit	Poly.	Closed-ended	Paper-based	Medium	Large
	Admits small sample size	Only uniform DIF					
	Equal-interval property						
IRT Rasch analysis							

	Pros		Cons	Scoring format	Response format	Admin. mode	N. items	Sample size
Lord chi-square statistic	It takes more than one parameter into account when calculating DIF Uniform and non-uniform DIF Item and bundle level Fixed and random DIF Differential Facet Functioning		Requires large sample size Rejection of null hypothesis when discrepancy between ICCs is small	Dicho.	Both	No info.	Medium	Large
Explanatory Item Response Modeling	Allows modeling item responses, item-level predictors, person-level predictors, and interactions in the same model Provides a framework for analyses		Only uniform DIF Requires model fit	Dicho.	Closed-ended	Paper-based	Medium	Medium + Large
IRT + ANOVA	Not provided		Not provided	Dicho.	Both	No info.	Small + Medium	Medium + Large
Cognitive diagnosis models	More effective than other methods More sensitive than other methods		Not provided	Both	Closed-ended	No info.	Medium	Medium + Large
IRT difficulty parameters comparison	Not provided		Not provided	Dicho.	Both	Computer-based	Medium	Large
Logistic discriminant function analysis	Uniform and non-uniform DIF Allows non-normality		Not provided	Poly.	Open-ended	No info.	Medium	Large
Generalized Linear Mixed Models	Not provided		Not provided	Dicho.	Both	No info.	Large	Large
Logits of the gender-specific item-difficulty scores	Not provided		Not provided	Dicho.	Closed-ended	No info.	Medium	Medium
Standardized Mean Difference	Not provided		Not provided	Both	Both	No info.	Medium + Large	Medium + Large

	Pros	Cons	Scoring format	Response format	Admin. mode	N. items	Sample size
Mantel-Haenszel (MH)	Dichotomous and polytomous items (with Generalized M-H)						
	Well established model's guideline						
	Most frequently used Good at detecting uniform DIF Highly consistent with other methods More sensitive than other methods Robust to missing data Admits small sample size Useful with large sample size	Only uniform DIF Requires unidimensionality Sensitive to sample size	Dicho.	Closed-ended	No info.	Medium	Large
SIBTEST or poly-SIBTEST	Adaptable to multilevel data						
	Comparable with other methods						
	Item and bundle level More sensitive than other methods Dichotomous and polytomous items (with poly-SIBTEST) Robust with a large percent of DIF items Admits small sample size Uniform and non-uniform DIF Allows multidimensionality	False identification of DBF when groups have different means on latent trait Several separate analyses when more than one bundle is assessed	Dicho. (SIBTEST) / Poly. (Poly-SIBTEST)	Closed-ended	Paper-based	Medium	Large

Note. *N. items Small* = $n < 10$; *N. items Medium* = $10 \leq n < 100$; *N. items Large* = $n \geq 100$; *Sample size Small* = $n < 100$; *Sample size Medium* = $100 \leq n < 1000$; *Sample size Large* = $n \geq 1000$; *FA* = Confirmatory Factor Analysis; *SEM* = Structural Equation Modeling; *MMIC* = Multiple-Indicator Multiple-Cause; *SIBTEST* = Simultaneous Item Bias Test; *No info.* = All studies that used this technique did not specify the administration mode.

**Supplementary Material XVIII. Pros and Cons of Strategies for Examination of Sources
of DIF/Lack of Measurement Invariance.**

Table XVIII.1.

Pros and Cons of Each Strategy for Examination of Sources of DIF/Lack of Measurement Invariance.

Strategies	Pros	Cons
Interviews	Useful to explore cognitive processes	Not provided
Focus groups	Group opinion better than individual	Not provided
Delphi technique	Useful to address complex problems Group opinion better than individual	Dependent on expertise of participants
Didactical interpretation of results	Enriches quantitative data	Not provided
Comparison of distractor response curves	Very informative (visual representation)	Not provided
Propensity scores	Permits hypothesizing a causal inference on DIF Benefits test fairness for test developers and test takers Flexible, easy to apply Applicable to traditional DIF detection methods Not affected by large number of covariates or unassociated covariates Less bias than regression models Robust to model misspecification	Not provided
Confirmatory approach	Permits hypothesizing a causal inference on DIF	No reasons for DIF

Supplementary Material XVII. Li et al. 's (2021) Categorization Across DIF Detection/Invariance Testing Methods.

Table XVII.1.

DIF Detection/Invariance Testing Method X Li et al. 's (2021) Categorization.

	Category A	Category B	Category C	TOTAL
Multigroup CFA	21 (70.00%)	3 (10.00%)	6 (20.00%)	30
Multigroup SEM	5 (55.56%)	0 (0.00%)	4 (44.44%)	9
Logistic regression or Ordinal logistic regression	3 (37.50%)	0 (0.00%)	5 (62.50%)	8
Latent class models	0 (0.00%)	0 (0.00%)	3 (100%)	3
Multilevel analysis	0 (0.00%)	0 (0.00%)	1 (100%)	1
MIMIC	2 (50.00%)	1 (25.00%)	1 (25.00%)	4
LH method	0 (0.00%)	1 (50.00%)	1 (50.00%)	2
Lord chi-square statistic	1 (33.33%)	0 (0.00%)	2 (66.67%)	3
Area measures	2 (66.67%)	1 (33.33%)	0 (0.00%)	3
IRT log-likelihood test	2 (40.00%)	1 (20.00%)	2 (40.00%)	5
Mixture IRT model	0 (0.00%)	0 (0.00%)	1 (100%)	1
Explanatory Item Response Modeling	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
IRT Rasch analysis	6 (66.67%)	1 (11.11%)	2 (22.22%)	9
IRT + ANOVA	0 (0.00%)	0 (0.00%)	2 (100%)	2
Cognitive diagnosis models	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
IRT difficulty parameters comparison	4 (100%)	0 (0.00%)	0 (0.00%)	4
Logistic discriminant function analysis	0 (0.00%)	0 (0.00%)	1 (100%)	1
Generalized Linear Mixed Models	0 (0.00%)	0 (0.00%)	1 (100%)	1
Logits of the gender-specific item-difficulty scores	1 (100%)	0 (0.00%)	0 (0.00%)	1
Standardized Mean Difference	1 (50.00%)	0 (0.00%)	1 (50.00%)	2
Mantel-Haenszel (MH)	2 (15.38%)	2 (15.38%)	9 (69.24%)	13
SIBTEST or poly-SIBTEST	2 (33.33%)	1 (16.67%)	3 (50.00%)	6
TOTAL	54	11	47	112

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

Supplementary Material XIX. GRADE-CERQual Assessment.

Table XIX.1.

CERQual Qualitative Evidence Profile.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Measurement invariance tests (e.g., multigroup CFA) are the most used methods.	Studies 1; 2; 3; 4; 6; 7; 8; 13; 14; 19; 24; 26; 30; 33; 35; 36; 39; 43; 47; 49; 51; 55; 57; 58; 61; 63; 67; 69; 76; 78; 79; 80; 82; 83; 86.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (more than 30 studies with reasonably rich information, two studies with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.
The Mantel-Haenszel procedure is the most used DIF detection method after measurement invariance tests.	Studies 3; 16; 20; 22; 23; 29; 33; 41; 53; 64; 66; 84; 85.	Minor methodological limitations (12 studies with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (13 studies with reasonably rich information).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
The IRT Rasch analysis and logistic regression are the third and fourth most used methods.	Studies 12; 15; 21; 25; 27; 33; 34; 41; 59; 65; 68; 71; 72; 73; 74; 81; 84.	Minor methodological limitations (13 studies with minor and four studies with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (16 studies with reasonably rich information, one study with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.
Methods with more reported pros are MIMIC, Mantel-Haenszel, SIBTEST/poly-SIBTE ST and LCA.	Studies 3; 16; 22; 24; 28; 33; 38; 41; 48; 51; 53; 66; 72; 75.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (12 studies with reasonably rich information).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.
Methods with more reported cons are multigroup CFA and Mantel-Haenszel.	Studies 15; 16; 24; 33; 36; 41; 51; 53.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (8 studies with reasonably rich information).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns about adequacy.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Methods associated with instruments with polytomous items are multigroup CFA, multigroup SEM, multilevel analysis, IRT Rasch analysis, logistic discriminant function analysis, and poly-SIBTEST.	Studies 1; 2; 3; 4; 6; 7; 13; 19; 26; 28; 30; 35; 36; 37; 39; 42; 43; 47; 49; 52; 55; 57; 58; 61; 63; 65; 67; 68; 69; 71; 72; 76; 78; 79; 80; 82; 83; 86.	Minor methodological limitations (more than 30 studies with minor and moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (more than 30 studies with reasonably rich information, two studies with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.
The cognitive diagnosis models and the Standardized Mean Difference are associated with both types of scoring format.	Studies 9; 70; 85; 87.	Moderate methodological limitations (three studies with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only four studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of moderate methodological limitations and substantial concerns about adequacy.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Methods associated with instruments with dichotomous items are LR, LCA, MIMIC, LH method, Lord chi-square statistic, area measures, IRTLR, mixture IRT model, EIRM, IRT + ANOVA, IRT difficulty parameters comparison, GLMM, logits of the gender-specific item-difficulty scores, Mantel-Haenszel, and SIBTEST.	Studies 10; 11; 14; 15; 16; 18; 22; 23; 24; 25; 29; 31; 32; 33; 34; 38; 41; 44; 46; 50; 51; 53; 54; 56; 60; 62; 64; 66; 77; 81.	Minor methodological limitations (28 studies with minor, one study with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (30 studies with reasonably rich information).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
The LH method, the Lord chi-square statistic, IRT + ANOVA, IRT difficulty parameters comparison, GLMM, and SMD are associated with instruments with both closed and open-ended answers.	Studies 5; 10; 11; 22; 31; 32; 40; 44; 50; 62; 74; 85; 87.	Moderate methodological limitations (11 studies with minor, one study with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (13 studies with reasonably rich information).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns regarding methodological limitations.
			Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (more than 60 studies with reasonably rich information, only two with reporting concerns).		
Almost all methods have been associated with instruments with closed-ended answers, except the logistic discriminant function analysis.	Studies 1-4; 6; 7; 9; 12; 13; 15-17; 19-21; 23-27; 30; 33; 34- 39; 41; 43; 46-49; 52-55; 57-61; 63; 64; 66-71; 74-80; 82-86.	Minor methodological limitations (more than 50 studies with minor and only five studies with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (more than 60 studies with reasonably rich information, only two with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Methods associated with a paper-based administration are multilevel analysis, IRTLR, EIRM, IRT Rasch analysis and SIBTEST/ poly-SIBTEST.	Studies 38; 52; 56; 65; 74.	Moderate methodological limitations (four studies with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only five studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of moderate methodological limitations and substantial concerns about adequacy.
Methods associated with a computer-based administration are multigroup CFA, multigroup SEM, and IRT difficulty parameters comparison.	Studies 4; 7; 24; 31; 32; 39; 58; 69; 76	Minor concerns about methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (9 studies with reasonably rich information).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns about adequacy.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall	
						CERQual Assessment of Confidence	Explanation of Judgement
MIMIC has been associated with both types of administration.	Studies 14; 24.	Minor concerns about methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only two studies with reasonably rich information due to missing data).	Low confidence	This finding was graded as low confidence because of substantial concerns about adequacy.
	Studies 1; 2; 7; 9; 11; 12; 13; 14; 15; 16; 17; 18; 19; 21; 22; 23; 24; 25; 27; 28; 31; 32; 33; 34; 36; 37; 38; 39; 41; 43; 44; 46; 47; 50; 51; 53; 54; 55; 57; 60; 61; 64; 65; 66; 68; 70; 71; 72; 73; 74; 75; 77; 78; 79; 80; 81; 84; 86.	Minor methodological limitations (more than 50 studies with minor and 6 studies with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (more than 50 studies with reasonably rich information, only two studies with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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. Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection. Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.						
Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence Explanation of Judgement
Multigroup SEM, multilevel analysis, and IRT + ANOVA have been associated with instruments with less than 10 items.	Studies 10; 26; 35; 49; 52; 57; 58; 63; 67; 76; 82.	Moderate methodological limitations (10 studies with minor and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (11 studies with reasonably rich information).	This finding was graded as moderate confidence because of moderate methodological limitations.
GLMM and SMD have been associated with instruments with 100 or more items.	Studies 62; 85.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only two studies with reasonably rich information).	This finding was graded as low confidence because of substantial concerns about adequacy.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
EIRM, IRT + ANOVA, cognitive diagnosis models, and SMD have been associated with both medium and large sample sizes.	Studies 9; 10; 11; 34; 38; 70; 85; 87.	Substantial methodological limitations (six studies with minor, one study with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (8 studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of substantial methodological limitations and moderate concerns about adequacy.
		Minor methodological limitations (more than 50 studies with minor, three studies with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (60 studies with reasonably rich information and one study with substantial reporting concerns).		
Almost all methods have been associated with large sample sizes, except logits of the gender-specific item-difficulty scores, which has been associated with medium sample sizes.	Studies 2-8; 12; 14-18; 20-23; 25; 28; 29-38; 40-44; 46; 48-53; 55; 57; 58; 60-62; 64; 67; 68; 72-77; 80; 81; 83; 84; 87.	Minor methodological limitations (more than 50 studies with minor, three studies with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (60 studies with reasonably rich information and one study with substantial reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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. Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection. Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.							
Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Methods mostly used by studies that examined sources of DIF are LR, LCA, multilevel analysis, Lord chi-square statistic, IRTLR, mixture IRT model, IRT + ANOVA, Mantel-Haenszel, and SIBTEST/ poly-SIBTEST.	Studies 10; 11; 16; 18; 20; 22; 23; 28; 29; 33; 34; 40; 41; 46; 51; 52; 53; 72; 75; 81; 84	Moderate methodological limitations (19 studies with minor, one study with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (21 studies with reasonably rich information).	Moderate confidence	This finding was graded as moderate confidence because of moderate methodological limitations.
The most used technique to explain sources of DIF/lack of measurement invariance is content analysis.	Studies 20; 29; 33; 34; 40; 51; 70; 72; 87.	Moderate methodological limitations (8 studies with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (9 studies with reasonably rich information).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns regarding methodological limitations and adequacy.

Objective: To describe how measurement invariance 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. Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection. Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.						
Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence Explanation of Judgement
Qualitative techniques used to explain sources of DIF/lack of measurement invariance are content analysis, interview, focus groups, didactical interpretation of results, expert appraisal, and Delphi technique.	Studies 10; 20; 21; 29; 33; 34; 40; 51; 53; 70; 72; 84; 87.	Moderate methodological limitations (10 studies with minor, two studies with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (13 studies with reasonably rich information).	This finding was graded as moderate confidence because of moderate concerns regarding methodological limitations.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Other techniques used to explain sources of DIF/lack of measurement invariance are MG SEM, MLA, analytic scoring analysis, ANOVA, distractor response curves, confirmatory approach, GLMM, multinomial LR, MLR, OLSR, poly-BW indices, and propensity scores.	Studies 11; 22; 23; 26; 37; 41; 46; 47; 49; 52; 53; 57; 62; 63; 81.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (15 studies with reasonably rich information and one study with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Studies Contributing to the Review Finding		Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Techniques to explain sources of DIF/lack of measurement invariance with more advantages are propensity scores and Delphi technique.	Studies 41; 84.	Moderate methodological limitations (one study with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only two studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of moderate methodological limitations and substantial concerns about adequacy.
Techniques to explain sources of DIF/lack of measurement invariance with more disadvantages are the confirmatory approach and Delphi technique.	Studies 46; 84.	Moderate methodological limitations (one study with minor and one study with moderate methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Substantial concerns about adequacy (only two studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of moderate methodological limitations and substantial concerns about adequacy.

Objective: To describe how measurement invariance 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.

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Causes of DIF/lack of measurement invariance have been searched on item domains, individual/cultural variables, item characteristics and cognitive strategies to respond to items	Studies 10, 11, 20, 21, 22, 23, 26, 28, 29, 33, 34, 37, 40, 41, 46, 47, 49, 51, 52, 53, 57, 62, 63, 70, 72, 81, 84, 87.	Minor methodological limitations (25 studies with minor, two studies with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (27 studies with reasonably rich information and one study with reporting concerns).	High confidence	This finding was graded as high confidence because of minor concerns regarding all dimensions.

Objective: To describe how measurement invariance 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. Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection. Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.						
Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence
Item domains that favors girls: reading comprehension, interpretation and estimation of data, patterns, and ratio/proportion, unapplied. Item domains that favors boys: graph, grammar, vocabulary, scientific design, probability, logarithms, exponents, decimal numbers, geometry, measurement, speed and velocity, technology, and Earth and space science, applied real life problems.	Studies 10, 11, 20, 23, 29, 33, 37, 40, 46, 51, 53, 70, 72, 87.	Moderate methodological limitations (12 studies with minor, one study with moderate, and one study with substantial methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Minor concerns about adequacy (13 studies with reasonably rich information and one study with reporting concerns).	This finding was graded as moderate confidence because of moderate concerns regarding methodological limitations. Moderate confidence

Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.

Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.

Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Cognitive strategies that favor girls: fixed or algorithmic answers based on their knowledge. Cognitive strategies that favor boys: judgments and high level of mental processing.	Studies 20, 23, 29, 33, 37, 40, 46, 53, 72.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (8 studies with reasonably rich information and one study with reporting concerns).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns about adequacy.
Item characteristics that favor girls: constructed-response, open-response. Item characteristics that favor boys: multiple-choice, visual stimuli.	Studies 22, 23, 33, 37, 40, 46, 62, 72, 87.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies).	Moderate concerns about adequacy (8 studies with reasonably rich information and one study with reporting concerns).	Moderate confidence	This finding was graded as moderate confidence because of moderate concerns about adequacy.

Objective: To describe how measurement invariance 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.							
Perspective: Experience and expertise of researchers in the field of measurement invariance/DIF detection.							
Included studies: Studies that studied measurement invariance/DIF and/or applied certain qualitative methods in order to gain insight into the causes of the gender gap in educational testing contexts.							
Review Finding	Studies Contributing to the Review Finding	Assessment of Methodological Limitations	Assessment of Relevance	Assessment of Coherence	Assessment of Adequacy	Overall CERQual Assessment of Confidence	Explanation of Judgement
Girls are more vulnerable to mother pressure regarding test anxiety, more insecure and less self-confident in mathematics, and four times more likely to answer items correctly when their country increases in its human development index. Boys benefit more from father support regarding test anxiety.	Studies 21, 26, 49, 53, 57, 63, 81.	Minor methodological limitations (all studies with minor methodological limitations).	Minor concerns about relevance (adequate population, phenomenon of interest and setting).	Minor concerns about coherence (data consistent within and across studies)	Substantial concerns about adequacy (only 7 studies with reasonably rich information).	Low confidence	This finding was graded as low confidence because of substantial concerns about adequacy.

Note. Numbers assigned to each study are based on alphabetic order. CFA = Confirmatory Factor Analysis; IRTLR = Item Response Theory Likelihood Ratio; SIBTEST = Simultaneous Item Bias Test; MMIC = Multiple-Indicator Multiple-Cause; LCA = Latent Class Analysis; MG SEM = Multigroup Structural Equation Modeling; EIRM = Explanatory Item Response Modeling; GLMM = Generalized Linear Mixed Models; SMD = Standardized Mean Difference; MLA = multilevel analysis; LR = logistic regression; MLR = multiple regression; OLS = Ordinal Logistic Regression.

Note. Numbers assigned to each study are based on alphabetic order. CFA = Confirmatory Factor Analysis; IRTLR = Item Response Theory Likelihood Ratio; SIBTEST = Simultaneous Item Bias Test; MIMIC = Multiple-Indicator Multiple-Cause; ICA = Latent Class Analysis; MG SEM = Multigroup Structural Equation Modeling; EIRM = Explanatory Item Response Modeling; GLMM = Generalized Linear Mixed Models; SMD = Standardized Mean Difference; MLA = multilevel analysis; LR = logistic regression; MLR = multiple regression; OLS = Ordinal Logistic Regression.

MATERIAL SUPLEMENTARIO DEL CAPÍTULO V

Database available for download here: <https://osf.io/uqje6/>

Script Step 1, 1 Class (MPlus)

```
TITLE: LCA Bullying Step 1 Classes 1

DATA: FILE = C:\Users\Desktop\bullbase.dat;

VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;

        AUXILIARY = GENDER;

        USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;

        CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;

        CLASSES = c(1);

ANALYSIS: TYPE = MIXTURE;

STARTS = 1000 100;

STITERATIONS = 50;

PLOT: TYPE = PLOT3;

        SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);

OUTPUT: TECH10 TECH11 TECH14;
```

Script Step 1, 2 Classes (MPlus)

```
TITLE: LCA Bullying Step 1 Classes 2

DATA: FILE = C:\Users\Desktop\bullbase.dat;

VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;

        AUXILIARY = GENDER;

        USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;

        CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;

        CLASSES = c(2);

ANALYSIS: TYPE = MIXTURE;

STARTS = 1000 100;

STITERATIONS = 50;

PLOT: TYPE = PLOT3;

        SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);

OUTPUT: TECH10 TECH11 TECH14;
```

Script Step 1, 3 Classes (MPlus)

```
TITLE: LCA Bullying Step 1 Classes 3

DATA: FILE = C:\Users\Desktop\bullbase.dat;

VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;

          AUXILIARY = GENDER;

          USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;

          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;

          CLASSES = c(3);

ANALYSIS: TYPE = MIXTURE;

STARTS = 1000 100;

STITERATIONS = 50;

PLOT: TYPE = PLOT3;

      SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);

OUTPUT: TECH10 TECH11 TECH14;
```

Script Step 1, 4 Classes (MPlus)

```
TITLE: LCA Bullying Step 1 Classes 4

DATA: FILE = C:\Users\Desktop\bullbase.dat;

VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;

          AUXILIARY = GENDER;

          USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;

          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;

          CLASSES = c(4);

ANALYSIS: TYPE = MIXTURE;

STARTS = 1000 100;

STITERATIONS = 50;

PLOT: TYPE = PLOT3;

      SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);

OUTPUT: TECH10 TECH11 TECH14;
```

Script Step 1, 5 Classes (MPlus)

```
TITLE: LCA Bullying Step 1 Classes 5
DATA: FILE = C:\Users\Desktop\bullbase.dat;
VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;
          AUXILIARY = GENDER;
          USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;
          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;
          CLASSES = c(5);
ANALYSIS: TYPE = MIXTURE;
STARTS = 1000 100;
STITERATIONS = 50;
PLOT: TYPE = PLOT3;
      SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);
OUTPUT: TECH10 TECH11 TECH14;
```

Script Multigroup, Heterogeneous (MPlus)

```
TITLE: Multigroup LCA Bullying (Heterogeneous Model)
DATA: FILE = C:\Users\Desktop\bullbase.dat;
VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;
          USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;
          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;
          KNOWNCLASS = G (GENDER = 1 GENDER = 2);
          CLASSES = G(2) L(4);
ANALYSIS: TYPE = MIXTURE;
MODEL: %OVERALL%
      L ON G;
```

Script Multigroup, Homogeneous (MPlus)

```
TITLE: Multigroup LCA Bullying (Homogeneous Model)

DATA: FILE = C:\Users\Desktop\bullbase.dat;

VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;
          USEVARIABLES = BU1 BU2 BU3 BU4 BU5 BU6;
          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;
          KNOWNCLASS = G (GENDER = 1 GENDER = 2);
          CLASSES = G(2) L(4);

ANALYSIS: TYPE = MIXTURE;

MODEL: %OVERALL%

      L ON G;

      %G#1.L#1%
      [BU1$1] (1);
      [BU2$1] (2);
      [BU3$1] (3);
      [BU4$1] (4);
      [BU5$1] (5);
      [BU6$1] (6);

      %G#2.L#1%
      [BU1$1] (1);
      [BU2$1] (2);
      [BU3$1] (3);
      [BU4$1] (4);
      [BU5$1] (5);
      [BU6$1] (6);

      %G#1.L#2%
      [BU1$1] (7);
      [BU2$1] (8);
      [BU3$1] (9);
      [BU4$1] (10);
      [BU5$1] (11);
      [BU6$1] (12);
```

%G#2.L#2%

[BU1\$1] (7);
[BU2\$1] (8);
[BU3\$1] (9);
[BU4\$1] (10);
[BU5\$1] (11);
[BU6\$1] (12);

%G#1.L#3%

[BU1\$1] (13);
[BU2\$1] (14);
[BU3\$1] (15);
[BU4\$1] (16);
[BU5\$1] (17);
[BU6\$1] (18);

%G#2.L#3%

[BU1\$1] (13);
[BU2\$1] (14);
[BU3\$1] (15);
[BU4\$1] (16);
[BU5\$1] (17);
[BU6\$1] (18);

%G#1.L#4%

[BU1\$1] (19);
[BU2\$1] (20);
[BU3\$1] (21);
[BU4\$1] (22);
[BU5\$1] (23);
[BU6\$1] (24);

%G#2.L#4%

[BU1\$1] (19);
[BU2\$1] (20);
[BU3\$1] (21);
[BU4\$1] (22);

```
[BU5$1] (23);  
[BU6$1] (24);
```

```
OUTPUT: TECH15;
```

Script Multinomial Logistic Regression (MPlus)

```
TITLE: LCA Bullying Step 2  
  
DATA: FILE = C:\Users\Desktop\bullbase.dat;  
  
VARIABLE: NAMES = GENDER BU1 BU2 BU3 BU4 BU5 BU6;  
          CATEGORICAL = BU1 BU2 BU3 BU4 BU5 BU6;  
          CLASSES = C(4);  
  
ANALYSIS: TYPE = MIXTURE;  
  
STARTS = 1000 100;  
  
STITERATIONS = 50;  
  
MODEL: %OVERALL%  
       C ON GENDER;  
  
PLOT: TYPE = PLOT3;  
      SERIES = BU1(1) BU2(2) BU3(3) BU4(4) BU5(5) BU6(6);  
  
OUTPUT: TECH10 TECH11 TECH14;
```


MATERIAL SUPLEMENTARIO DEL CAPÍTULO VI

PROTOCOLO DE ENTREVISTA COGNITIVA

Información para la entrevistadora

Información general y fases de la entrevista

Es una entrevista semiestructurada, sin un guión cerrado, por lo que debe prestarse atención al discurso “espontáneo” de las personas entrevistadas, con el objetivo de dirigir la indagación hacia los contenidos relevantes descritos en la sección “*Consideraciones e instrucciones para la entrevistadora*”. El protocolo debe utilizarse a modo de guía.

La entrevista se dividirá en varias fases:

- **Fase de presentación y personalización de la entrevista.** La entrevistadora se presentará a la persona entrevistada, y a esta última se le dirá que si quiere puede identificarse con un seudónimo. Esta fase servirá para que de aquí en adelante la situación entre entrevistadora y persona entrevistada sea cómoda. En esta fase también la entrevistadora se encargará de que se ocupen las posiciones adecuadas en la sala, no sentándose “cara a cara” a la persona entrevistada para rebajar la presión. Las posiciones deberán ser óptimas para la grabación del audio de la entrevista.
- **Fase de establecimiento de la comunicación.** La entrevistadora tratará de generar un ambiente distendido y agradable para la persona entrevistada, formulando frases y preguntas sencillas como “¿qué tal?”, o “*muchas gracias por haberte ofrecido a participar*”. El objetivo es que la interacción sea lo más natural posible.
- **Fase de explicación del objetivo de la entrevista.** La entrevistadora facilitará la “Hoja Informativa del Estudio” a la persona entrevistada y la leerá en voz alta, asegurándose de que la persona entrevistada comprende y acepta todo lo que ahí se expone. La

persona entrevistada, a continuación, firmará el consentimiento informado y la documentación que fuera necesaria para recibir la compensación económica. Si surgiera alguna duda respecto al procedimiento de la entrevista, se resolvería en esta fase.

- **Fase de cumplimentación de la escala.** La persona entrevistada contestará a la escala (escala ST038 del Cuestionario del Estudiante del PISA 2018, sobre experiencias de haber sufrido *bullying*), junto a la escala ST034 del mismo cuestionario sobre sentido de pertenencia al centro escolar, y a algunas preguntas sobre datos sociodemográficos (curso, edad, género y estatus de inmigración). La administración será en un ordenador, para imitar las condiciones de administración de la escala en el estudio original.
- **Fase de indagación.** Al principio de esta fase, la entrevistadora avisará de que a partir de ese momento el audio será grabado, y pondrá en marcha los dispositivos de grabación de audio. La entrevistadora realizará las preguntas de indagación, siguiendo el protocolo, haciendo hincapié en que la persona tome un rol de “*story-teller*” o “*cuentacuentos*”, y relate libremente aquello que considere que puede aportar luz sobre las respuestas que ha dado a los ítems de la escala.
- **Fase de cierre de la entrevista.** La entrevistadora resolverá cualquier situación tensa que se haya podido dar y comunicará que la entrevista ha finalizado. Dará las gracias a la persona entrevistada y le trasladará una impresión positiva del proceso, preguntándole después a la persona entrevistada qué impresión ha tenido ella de la entrevista. Si es necesario, se podrá volver a las preguntas neutrales para lograr un cierre adecuado, y la entrevistadora despedirá a la persona.

Consideraciones e instrucciones para la entrevistadora

A lo largo de la entrevista, la entrevistadora deberá considerar estas cuestiones: 1) se debe pasar por todas las fases de la entrevista; 2) no debe interrumpir a la persona mientras ésta

intervenga; 3) no debe hacer preguntas que orienten las respuestas; 4) debe establecer pausas tras las respuestas; 5) debe escuchar activamente y omitir “orientaciones de respuesta” (e.g., asentir con la cabeza, interjecciones como “*um*”...); 6) debe asegurarse de si la persona entrevistada ha tenido problemas o dudas al responder; y 7) al final de la entrevista debe tener recogida información sobre todos los aspectos relevantes descritos a continuación:

- Situaciones en las que pensaban al contestar a los ítems (principalmente, si piensan en situaciones de bullying relacional, directo, o de todos los tipos).
- Cómo son los “alumnos” en los que han pensado al contestar las preguntas (principalmente, si son sus amigos o no).
- Si creen que los comportamientos representados en los ítems podrían calificarse como *bullying*.

Si la entrevistadora considera que no ha obtenido suficiente información sobre estas cuestiones, puede recurrir a las habituales estrategias de indagación para profundizar en las respuestas, repitiendo las últimas palabras o frases de la persona entrevistada, o formulando frases tales como “*¿Puedes repetirme eso último?*”; “*Me parece que eso no lo he entendido...*”, “*Cuéntame más sobre...*”.

Actuación frente a situaciones complicadas

En caso de que durante la entrevista se detecte un caso nuevo de acoso escolar o *bullying*, o “indicios” del mismo, la entrevistadora pondrá en marcha el protocolo de actuación en supuestos de acoso escolar establecido por la Junta de Andalucía a través de la Orden de 20 de junio de 2011, elaborada de acuerdo con lo especificado en los Decretos 327/2010 y 328/2010 de 13 de julio, mediante los que se aprueban los Reglamentos Orgánicos de los institutos de educación secundaria.

Descripción del guión de la entrevista

Después de presentar el cuestionario “ad hoc” al que cada persona entrevistada deberá contestar previamente en un ordenador, se presentan las preguntas de indagación. Cada pregunta de indagación cuenta con una serie de posibles aclaraciones que pueden hacerse después, en caso de que exista la necesidad de aclarar la pregunta, de insistir más en ella, o de profundizar en el tema.

Información previa para la persona entrevistada

La entrevistadora facilitará y leerá en voz alta a la persona entrevistada la “Hoja Informativa del Estudio”, que se transcribe a continuación:

La investigadora María del Carmen Navarro González, miembro del grupo de investigación HUM-624 de la Universidad de Granada, está realizando un estudio para conocer las percepciones de las y los estudiantes de Educación Secundaria Obligatoria (E. S. O.), sobre el clima escolar. El estudio está financiado por el Ministerio de Universidades con la Ley 14/2011 FPU. Tu participación consistirá en responder a unas preguntas sobre cómo te sientes, tus ideas y opiniones, etc., acerca de las relaciones en el aula y en el colegio. Si tienes más preguntas o quieres conocer más del estudio, puedes ponerte en contacto con María del Carmen (mcnavarro@ugr.es).

Responderás a algunas preguntas sobre datos personales que son imprescindibles para los objetivos del estudio. Todas tus respuestas serán confidenciales, “anonimizadas” mediante un código, y almacenadas en condiciones de seguridad. Sólo la investigadora tendrá acceso a tus datos, que no están ligados a tus respuestas en el mismo archivo. Los datos serán tratados de acuerdo con la legislación vigente: Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos y Garantías de Derechos Digitales. Puedes abandonar en cualquier momento el estudio, si así lo deseas, sin que esto suponga ninguna consecuencia negativa para ti. Durante aproximadamente una hora, se te harán una serie de preguntas que deberás contestar con la mayor sinceridad posible.

Tus respuestas son muy importantes para el estudio, ya que con ellas nos podremos acercar más a la realidad vivida por las y los adolescentes en las aulas. Participar en este

estudio no te causará ningún perjuicio. Además, si deseas participar, tu participación se compensará con 15 euros.

El audio de tus respuestas será grabado y sólo tendrá acceso a la grabación el personal relacionado con el proyecto. Debes saber que los derechos de acceso, supresión, rectificación, oposición, limitación del tratamiento, portabilidad y olvido serán responsabilidad de la investigadora María del Carmen Navarro González, a quien se podrá contactar para resolver cualquier duda o problema. También puedes obtener más información sobre tus derechos sobre el tratamiento de datos personales en la Secretaría General de la Universidad de Granada, o poniéndote en contacto por correo electrónico con la Oficina de Protección de Datos: protecciondedatos@ugr.es.

Guión de la entrevista

ST034. Sentido de pertenencia al centro escolar. Administración por ordenador

En relación con tu centro, ¿en qué medida estás de acuerdo con las siguientes afirmaciones?

(Por favor, selecciona una respuesta en cada fila)

	Totalmente en desacuerdo	En desacuerdo	De acuerdo	Totalmente de acuerdo
Me siento marginado en el centro.				
Hago amigos fácilmente en el centro.				
Me siento integrado en el centro.				
Me siento incómodo y fuera de lugar en mi centro.				
Caigo bien a otros alumnos.				
Me siento solo en el centro.				

ST038. Experiencias de haber sufrido *bullying*. Administración por ordenador

Durante los últimos 12 meses, ¿con qué frecuencia has vivido las siguientes experiencias en el centro?

(Algunas experiencias también pueden darse en las redes sociales)

(Por favor, selecciona una respuesta en cada fila)

	Nunca o casi nunca	Varias veces al año	Varias veces al mes	Una o más veces a la semana
Otros alumnos me han excluido a propósito.				
Otros alumnos se han reído de mí.				
Otros alumnos me han amenazado.				
Otros alumnos me han quitado o han roto mis cosas.				
Otros alumnos me han golpeado o empujado.				
Otros alumnos han difundido rumores horribles sobre mí.				

Indagación

“Ahora me gustaría que hablemos de tus respuestas a las preguntas del cuestionario web sobre tus experiencias en el centro”.

Pruebas de indagación:

[La entrevistadora siempre debe tener en cuenta que si en la respuesta a una pregunta de indagación o “probe” (general o específica) ya obtiene información pensada para otra pregunta de indagación, esta última no se plantea a la persona entrevistada, salvo que quede algún aspecto por indagar. En este caso, la forma para introducir la indagación adicional puede ser: *“me has hablado antes de...”, “¿a qué te referías?”*, o *“¿qué querías decir con...?”*].

- [Debe ser siempre la primera pregunta de indagación] *Cuéntame, ¿qué has pensado mientras respondías a las preguntas? ¿Qué has recordado? ¿Qué situaciones se te han venido a la mente?*

Posibles aclaraciones:

Preguntas sobre la ST034:

- *Acerca de las primeras preguntas sobre cómo te sientes en el centro, que eran...*
[Leer los ítems de la escala ST034 rápidamente para recordarlos], *¿qué has pensado mientras las contestabas?*

Preguntas sobre la ST038:

Caso 1: Da una respuesta muy vaga y en todos los ítems ha marcado “Nunca o casi nunca”.

- *Dime qué es lo que se te ha pasado por la cabeza cuando has contestado que no has vivido ninguna de estas experiencias.* [Preguntar por situaciones

específicas descritas en algún ítem concreto sólo si la persona dice algo que llame la atención referente a alguno de los ítems (por ejemplo, si dice que alguna vez lo sufrió pero que no lo sufre habitualmente, o si dice algo que deje entrever o dé lugar a sospechar que ha sufrido algún comportamiento abusivo pero no lo ha reportado en el test)].

Caso 2: Da una respuesta muy vaga y en al menos un ítem ha marcado “Varias veces al año”, “Varias veces al mes”, o “Una o más veces a la semana”.

- [Sólo en caso de que haya al menos un ítem en el que haya marcado “Nunca o casi nunca”. Si no, ir directamente a la pregunta siguiente]. *Dime qué es lo que se te ha pasado por la cabeza cuando has contestado que no has vivido la(s) experiencia(s) descrita(s) en el/los ítem(s)...* [Leer el ítem o los ítems en los que haya marcado “Nunca o casi nunca”. Procedimiento a seguir igual que en el Caso 1, con el añadido de preguntar específicamente por ítems que describen situaciones que podrían ser “congruentes” con aquellas que sí dice haber sufrido (por ejemplo, si dice haber sufrido algún comportamiento de bullying relacional, preguntar por el resto de ítems de bullying relacional), y con el añadido de preguntar específicamente por el ítem 1 si es el único en el que ha marcado “Nunca o casi nunca”].
- *Dime qué es lo que se te ha pasado por la cabeza cuando has contestado que has vivido la experiencia descrita en el ítem...* [Leer el ítem o uno de los ítems en los que haya marcado “Varias veces al año”, “Varias veces al mes”, o “Una o más veces a la semana”, priorizando aquellos en los que la frecuencia de ocurrencia sea mayor, y luego ir preguntando por los demás, uno por uno].

- [En cualquiera de los dos casos. Si hay “incongruencias” entre las respuestas a la escala ST034 y las respuestas a la escala ST038, o entre los propios ítems de la escala ST034] Cuéntame más sobre la respuesta que diste en la pregunta... [Recordar uno por uno aquellos ítems de ambas escalas entre los que existan las incongruencias, empezando por los ítems de la escala ST034].
- [En principio, preguntar sólo en el Caso 2. Podría preguntarse en el Caso 1 si se da la circunstancia de que la persona no reporta haber sufrido un comportamiento abusivo en el test, pero después lo admite] *¿En quiénes has pensado cuando has leído en las preguntas la palabra “alumnos”? Por ejemplo, en...* [Leer un ítem en el que haya marcado “Varias veces al año”, “Varias veces al mes”, o “Una o más veces a la semana”, priorizando aquel en el que la frecuencia de ocurrencia sea mayor].

Posibles aclaraciones:

- *¿Esos “alumnos” en los que has pensado son amigos tuyos o amigas tuyas?*
- [En caso de que el término “alumnos” le haga pensar en amigos suyos] *¿Y ese tipo de comportamientos son habituales dentro de tu grupo de amigos?*
- *Sobre el primer conjunto de preguntas que contestaste, ¿qué significa para ti “estar integrado en el centro”? ¿y estar “marginado”?*
- [Debe ser siempre la última pregunta de indagación] *Por último, ¿crees que los comportamientos representados en las seis últimas preguntas que contestaste son bullying o acoso escolar?* [Mostrar de nuevo las preguntas de la escala ST038 para que las recuerden]

Posibles aclaraciones:

- [En caso de que hable sobre lo que otra gente podría pensar] *Muy buena observación, pero me interesa saber si tú personalmente los consideras bullying o no.*

