GRUPO EDITORIAL DE PSICOFUNDACIÓN

Apuntes de Psicología (2025) 43(1) 107-120

Apuntes de Psicología

https://www.apuntesdepsicologia.com • e-ISSN: 0213-3334 • eISSN: 19896441

Colegio Oficial de Psicología de Andalucía Occidental
Universidad de Cádiz
Universidad de Córdoba
Universidad de Huelva
Universidad de Sevilla

Artículo

Equidad de los Test Psicológicos Desde una Perspectiva de Género: Análisis de Buenas Prácticas en Psicometría

Francisco Rivera

Universidad de Sevilla, España

INFORMACIÓN

Recibido: Septiembre 23, 2024
Aceptado: Diciembre 17, 2024

Palabras clave:
Psicometría
Evaluación psicológica
Equidad
Sesgo de género

RESUMEN

Los test psicológicos son herramientas clave en la evaluación de características cognitivas, sociales, emocionales y comportamentales. En el estudio de las propiedades psicométricas de dichos test se abordan de forma frecuente las evidencias de fiabilidad y validez, pero el análisis de la equidad, específicamente desde una perspectiva de género, sigue siendo un reto. Este artículo analiza las prácticas habituales en el abordaje de género en la construcción y análisis de estos instrumentos. A través de una revisión sistemática de 20 estudios publicados en el año 2023 en revistas especializadas del campo de la evaluación psicológica, se identifican buenas prácticas psicométricas para abordar la equidad de género. Los resultados muestran que, aunque algunos estudios incluyen análisis de funcionamiento diferencial de ítems (*Differential Item Functioning*, DIF) y pruebas de invarianza factorial, son pocos los que desglosan resultados por género o implementan medidas adecuadas y sistemáticas para abordar la equidad. El artículo propone una serie de recomendaciones para garantizar la equidad de género en la psicometría, destacando la importancia de integrar análisis diferenciados por género en todas las etapas del desarrollo y validación de los test psicológicos. Se concluye que una mayor atención a la equidad de género es esencial para evitar sesgos que distorsionen los resultados y asegurar evaluaciones justas.

Equity in Psychological Tests From a Gender Perspective: Analysis of Best Practices in Psychometrics

Keywords:
Psychometrics
Psychological assessment
Equity
Gender Bias

ABSTRACT

Psychological tests are key tools for evaluating cognitive, social, emotional, and behavioral traits. While the study of the psychometric properties of these tests often addresses evidence of reliability and validity, the analysis of equity, particularly from a gender perspective, remains a challenge. This article examines common practices in addressing gender in the construction and analysis of these instruments. Through a systematic review of 20 studies published in 2023 in specialized journals in the field of psychological assessment, psychometric best practices for addressing gender equity are identified. The results show that, although some studies include differential item functioning (DIF) analyses and tests of factorial invariance, few disaggregate results by gender or implement adequate and systematic measures to address equity. The article offers a series of recommendations to ensure gender equity in psychometrics, highlighting the importance of integrating gender-differentiated analyses at every stage of test development and validation. It concludes that greater attention to gender equity is essential to avoid biases that distort results and to ensure fair assessments.

Cómo citar: Rivera, Francisco (2025). Equidad de los test psicológicos desde una perspectiva de género: análisis de buenas prácticas en psicometría. *Apuntes de Psicología, 43*(1), 107-120. https://doi.org/10.70478/apuntes.psi.2025.43.10

Autor de correspondencia: Francisco Rivera, franciscorivera@us.es

Introducción

Los test psicológicos son herramientas esenciales para la evaluación de características cognitivas, sociales, emocionales y comportamentales de las personas, y juegan un papel crucial en una amplia variedad de contextos, como el educativo, laboral y clínico. Desde sus inicios, la psicometría ha buscado desarrollar instrumentos que permitan medir con precisión constructos psicológicos complejos, tales como la inteligencia, la personalidad y las habilidades específicas (Anastasi y Urbina, 1997). Estos instrumentos no solo facilitan la comprensión del comportamiento humano, sino que también proporcionan una base para la toma de decisiones informadas que afectan directamente a la vida de las personas evaluadas.

En el ámbito educativo, los test son utilizados para identificar talentos, necesidades educativas especiales y para orientar el proceso de enseñanza y aprendizaje. La identificación temprana de dificultades de aprendizaje, por ejemplo, permite implementar intervenciones personalizadas que pueden cambiar el curso del desarrollo académico de un o una estudiante (Fernández-Ballesteros, 2008). De igual forma, las pruebas de orientación vocacional y de aptitudes ayudan a los y las estudiantes a tomar decisiones informadas sobre su futuro profesional, lo que puede tener un impacto significativo en su vida laboral y personal.

En el contexto laboral, la psicometría se utiliza para seleccionar personal, evaluar competencias y diseñar programas de desarrollo profesional. Las organizaciones dependen de estos test para seleccionar a los y las candidatas con habilidades y características necesarias para desempeñar un rol específico. Además, las evaluaciones psicométricas pueden ayudar a identificar áreas de mejora en empleados actuales, contribuyendo al desarrollo de planes de formación personalizados que no solo mejoran el rendimiento individual, sino también la eficiencia global de la organización (Muñiz, 2010).

En el ámbito clínico, los test psicológicos son fundamentales para el diagnóstico y tratamiento de trastornos mentales. Las evaluaciones clínicas permiten a los y las profesionales de la salud mental obtener un perfil detallado de sus pacientes, identificar trastornos como depresión, ansiedad, trastornos del espectro autista, entre otros, y desarrollar planes de tratamiento adecuados. Los test como el MMPI-2 o el WAIS-IV son ejemplos de instrumentos clínicos ampliamente utilizados y validados que proporcionan información crítica para la planificación terapéutica (Groth-Marnat y Wright, 2016).

La confianza depositada en estos test, tanto por parte de los y las profesionales como de las personas evaluadas, se fundamenta en la premisa de que los resultados obtenidos son precisos, válidos y equitativos (AERA, APA y NCME, 2014). Sin embargo, esta confianza solo es justificable si los instrumentos cumplen con altos estándares de calidad psicométrica, lo que incluye no solo la precisión y consistencia de las mediciones, sino también su equidad para todos los individuos evaluados.

Principales Propiedades de las Puntuaciones de los Test: Evidencias de Validez, Fiabilidad y Equidad

Para que las puntuaciones derivadas de los test psicológicos, y las conclusiones obtenidas de ello, sean efectivas y éticamente defendibles, deben cumplir con ciertos estándares psicométricos que garanticen su calidad (AERA, APA y NCME, 2014). Tres de las características más críticas asociadas a estas puntuaciones son la validez, la fiabilidad y la equidad. Es fundamental comprender que estas no son propiedades intrínsecas de los test en sí mismos, sino que se refieren específicamente a las puntuaciones que los test generan, las cuales dependen de las poblaciones en las que se aplican y del uso que se les da (Prieto y Delgado, 2010).

Evidencias de Validez

La concepción unitaria que engloba todas las evidencias de validez es el concepto más crítico en la evaluación de las puntuaciones de un test psicológico y se refiere al grado en que estas puntuaciones realmente reflejan el constructo que se pretende medir. La validez no es una propiedad que reside dentro del test de manera universal e inmutable, sino que se refiere a la adecuación de las interpretaciones y usos específicos de las puntuaciones en un contexto dado. Paula Elosúa (2003), en su análisis sobre la validez de los test, subraya que la validez debe ser entendida como un proceso continuo de evaluación que incluye múltiples formas de evidencia, categorizadas en validez interna y validez externa.

En primer lugar, la *validez interna* se refiere a la consistencia y coherencia de las puntuaciones dentro del contexto específico de la evaluación. En esta aproximación se incluye, en primer lugar, la validez de contenido, que evalúa si las puntuaciones reflejan adecuadamente el dominio del contenido que se pretende medir, asegurando que todos los aspectos relevantes del constructo estén representados en las puntuaciones obtenidas. Por otro lado, la validez de constructo se centra en si las puntuaciones se alinean con el constructo teórico que el test pretende medir, considerando la estructura interna del test y cómo esta se relaciona con otras medidas de constructos similares o diferentes.

En segundo lugar, la *validez externa* abarca la generalización de las puntuaciones más allá del contexto específico en el que se aplicaron. Esta incluye la validez de criterio, que examina cómo las puntuaciones de un test predicen criterios externos relevantes, como el rendimiento académico o el desempeño laboral, y la validez ecológica, que se refiere a la aplicabilidad de las puntuaciones en situaciones de la vida real fuera del entorno controlado de la evaluación. La validez externa es esencial para asegurar que las interpretaciones derivadas de las puntuaciones sean útiles y pertinentes en contextos diversos y reales.

Evidencias de Fiabilidad

Por otro lado, las evidencias de fiabilidad de las puntuaciones, en este marco, se refieren a la consistencia de las puntuaciones obtenidas por un test en diferentes ocasiones y condiciones similares (Cronbach, 1951). Al igual que la validez, la fiabilidad no es una

propiedad intrínseca del test, sino que depende de la población en la que se aplique y del contexto específico de uso. Las puntuaciones de un test son consideradas fiables si son consistentes a lo largo del tiempo y en diferentes contextos, lo que implica que las decisiones basadas en estas puntuaciones serán estables y replicables. Los análisis de fiabilidad, como la consistencia interna y la fiabilidad test-retest, permiten evaluar hasta qué punto las puntuaciones mantienen su estabilidad en diferentes aplicaciones del test (Nunnally y Bernstein, 1994).

Evidencias de Equidad

Finalmente, las evidencias de equidad son las que aseguran que un test sea justo y no discriminatorio para todos los grupos de personas que lo utilizan, independientemente de su género, etnia o condición socioeconómica (Zumbo y Chan, 2014). La equidad en psicometría se define, por tanto, como las propiedades psicométricas de las puntuaciones de un test, para una finalidad determinada, de ser igualmente válido, fiable y accesible para todos los grupos que lo utilizan, independientemente de su género, etnia u otras características sociodemográficas (Messick, 1995). Por tanto, es básico enfocar la equidad como un aspecto transversal que se articula de forma intrínseca con las dos fuentes de evidencias anteriores: validez y fiabilidad.

La equidad es crítica porque, sin ella, los test podrían perpetuar desigualdades y estereotipos, afectando negativamente a grupos específicos. Para garantizar la equidad, el equipo de desarrollo de test debe realizar análisis detallados para identificar y corregir cualquier sesgo en los ítems o en la interpretación de los resultados (Camilli y Shepard, 1994). Es importante aclarar que la aplicación de la equidad implica considerar las posibles fuentes de sesgo en todas las etapas del proceso de creación del instrumento, desde la redacción de ítems hasta su adaptación cultural, o el estudio de las propiedades psicométricas en una población específica.

Alcanzar la equidad en los test psicológicos implica, por tanto, asegurarse de que las diferencias observadas en los resultados de los test reflejen diferencias reales en el constructo medido, lo que se denominaría impacto, en lugar de ser producto de sesgos sistemáticos presentes en el test o en su aplicación. Además, en línea con el concepto amplio de validez externa, la equidad implica que las interpretaciones derivadas de las puntuaciones sean apropiadas y útiles para todas las personas evaluadas, sin que ninguna característica sociodemográfica influya de manera injusta en los resultados o en las decisiones basadas en ellos (AERA, APA y NCME, 2014).

Problematización: Equidad en los Test Psicológicos Desde la Perspectiva de Género

La equidad en los test ha sido un tema de creciente preocupación en las últimas décadas, especialmente en lo que respecta a las diferencias de género. Siguiendo un cierto paralelismo con la investigación en general, la integración de la perspectiva de género en la psicometría puede realizarse de dos formas principales (Caprile et al., 2012): por un lado, a través de una evaluación psicométrica que sea consciente del género, en la que este se tenga en cuenta de manera constante durante todo el proceso de validación; por otro lado, mediante una investigación psicométrica enfocada específicamente en el género, en la que este se convierte en el tema central de análisis. Esto implica que, de una forma u otra, todas las investigaciones que aborden de forma parcial o completa la evaluación de los instrumentos de medición deberían incluir el enfoque de género con el objeto de prevenir posibles situaciones de inequidad o sesgo.

Históricamente, muchos instrumentos psicométricos fueron desarrollados en contextos que no consideraban adecuadamente la diversidad de género, lo que ha llevado a sesgos que podrían distorsionar los resultados y las conclusiones extraídas de estos test (Helms, 2006). La falta de equidad de género en los test puede manifestarse en varias formas, incluyendo la construcción de ítems que favorecen a un género sobre otro, la normatización de los test en poblaciones no representativas y la interpretación de los resultados de manera que refuerce o perpetúen estereotipos de género.

Un ejemplo de este problema se encuentra en los test de aptitud y habilidades cognitivas, donde, en ocasiones, se han observado diferencias de género en los resultados que no reflejan necesariamente diferencias reales en las habilidades medidas, sino más bien sesgos en la forma en que los ítems están construidos o en cómo se interpretan los resultados (Hyde, 2005). Estas diferencias pueden deberse, por tanto, a factores como el contenido específico de los ítems, que podría estar más alineado con las experiencias y expectativas de un género en particular o a factores culturales que influyen en cómo los individuos de diferentes géneros responden a las preguntas del test.

Otro ejemplo puede encontrase en la evaluación de violencia de género (Delgado-Álvarez, 2020), donde se enfatiza en aspectos tan esenciales como la validez de constructo en un test tan frecuentemente utilizado como el *Conflict Tactics Scales*. En este caso, el sesgo se inicia desde la propia definición del constructo, ya que parte de la evaluación de la autopercepción en la perpetración y recepción de algunas agresiones, lo que dista bastante de una definición completa de violencia de género. Además, se cuestiona su capacidad para medir la violencia de género, ya que no diferencia entre agresiones proactivas y reactivas, y omiten aspectos cruciales como la agresión sexual. Esto genera resultados inconsistentes con la evidencia actual sobre violencia de género, lo que podría deberse a sesgos en la evaluación. Estos sesgos pueden perpetuar interpretaciones erróneas debido a la falta de exhaustividad y a la incapacidad del test para representar adecuadamente el constructo de violencia de género.

El sesgo de género no solo afecta a la precisión de las mediciones, sino que también tiene implicaciones éticas significativas. Los resultados de los test a menudo se utilizan para tomar decisiones que impactan en la vida de los individuos, como la admisión a programas educativos, la selección para empleos o el diagnósti-

co de trastornos mentales. Si un test presenta sesgos de género, estas decisiones pueden perpetuar desigualdades e injusticias, lo que subraya la importancia de desarrollar y aplicar instrumentos psicométricos que sean verdaderamente equitativos.

Además, la equidad de género en psicometría no solo se refiere a evitar el sesgo, sino también a reconocer y valorar las diferencias de género de manera justa. Esto implica diseñar test que sean igualmente relevantes y accesibles para todas las personas, independientemente de su género, y que los resultados reflejen con precisión las capacidades y características individuales sin influencia de estereotipos o expectativas de género (Helms, 2006).

Perspectiva de Género en la Psicometría: Análisis Histórico

La consideración del género en la psicometría ha evolucionado significativamente desde los primeros desarrollos en la creación de instrumentos de medición psicológica. Inicialmente, los test psicológicos fueron diseñados en contextos en los que predominaban visiones androcentristas, lo que llevó a que las diferencias de género fueran insuficientemente consideradas o bien interpretadas de manera sesgada (Helms, 2006). Este enfoque limitado ha tenido implicaciones profundas en cómo se han diseñado, validado y aplicado los test a lo largo de la historia.

Durante gran parte del siglo XX, la psicometría se enfocó principalmente en desarrollar instrumentos que se consideraban universales, es decir, válidos para cualquier persona, independientemente de su género. Sin embargo, estos test a menudo fueron desarrollados y normatizados en muestras predominantemente masculinas, lo que resultó en instrumentos que no siempre reflejaban con precisión las capacidades y características de las mujeres (Delgado-Álvarez, 2020). Además, la interpretación de las puntuaciones a menudo se hacía desde una perspectiva que asumía que cualquier diferencia de género era un reflejo de diferencias innatas en habilidades o características, sin considerar el impacto de factores socioculturales y de género.

El avance de los estudios feministas y la creciente conciencia sobre la equidad de género llevaron a cuestionar la supuesta neutralidad de los test psicológicos. Investigaciones emergentes comenzaron a señalar que muchas de las diferencias observadas en las puntuaciones entre hombres y mujeres podían estar más relacionadas con el sesgo en los ítems, las condiciones de administración de los test o las expectativas culturales; que con diferencias reales en los constructos medidos (Hyde, 2005). Esto llevó a una reevaluación de cómo se construyen y aplican los test, con un enfoque más explícito en identificar y corregir posibles sesgos de género.

En la actualidad, existe un consenso creciente en la psicometría sobre la necesidad de desarrollar instrumentos que consideren de manera adecuada las diferencias de género. Esto implica no solo evitar la discriminación explícita en los ítems, sino también asegurarse de que las condiciones de evaluación y las interpretaciones de las puntuaciones sean sensibles a las realidades actuales de género. En definitiva, la incorporación de la perspectiva de género en la psicometría moderna busca garantizar que los test sean equitativos, válidos y útiles para todas las personas, sin perpetuar estereotipos de género o desigualdades (AERA, APA y NCME, 2014).

Estudios Sobre Sesgo de Género en Test Psicológicos

El sesgo de género en los test ha sido un tema de investigación y debate significativo en las últimas décadas. El sesgo de género se refiere a la presencia de elementos en un test que sistemáticamente favorecen a un género sobre otro, no por diferencias reales en el constructo medido, sino por cómo están formulados los ítems o cómo se interpretan las puntuaciones (Camilli y Shepard, 1994). Este sesgo puede manifestarse de varias formas, incluyendo diferencias en el contenido de los ítems, las expectativas del evaluador o las condiciones en las que se administra el test.

Varios estudios han documentado la existencia de sesgos de género en una amplia gama de test. Por ejemplo, investigaciones en el ámbito de las pruebas de aptitud y rendimiento académico han mostrado que los ítems que involucran estereotipos de género pueden favorecer sistemáticamente a los hombres o las mujeres, dependiendo del contexto. Un estudio clave de Hyde (2005) destacó cómo las diferencias de género en matemáticas, tradicionalmente atribuidas a diferencias innatas en habilidad, en realidad podrían ser el resultado de sesgos en los test y de factores socioculturales que influyen en cómo los individuos de diferentes géneros abordan estas pruebas.

En el ámbito de la evaluación de la personalidad, Helms (2006) ha encontrado que ciertos ítems pueden ser interpretados de manera diferente por hombres y mujeres, lo que lleva a puntuaciones que no reflejan con precisión el constructo medido para ambos géneros. Por ejemplo, ítems que miden agresividad o asertividad pueden estar sesgados hacia una interpretación que refuerza estereotipos masculinos, lo que podría inflar las puntuaciones de los hombres en estas dimensiones y subestimar las de las mujeres.

Otro campo en el que se ha identificado el sesgo de género es en las evaluaciones de liderazgo, donde los test utilizados a menudo reflejan estereotipos de género en cuanto a las características y comportamientos esperados de las personas que ejercen el liderazgo. Tradicionalmente, estos test han tendido a valorar más altamente atributos asociados con el liderazgo masculino, como la asertividad y la competitividad, mientras que subestiman cualidades relacionadas con el liderazgo femenino, como la colaboración y la empatía. Este sesgo en la evaluación puede llevar a resultados que refuercen la idea de que los hombres son más adecuados para roles de liderazgo, perpetuando la desigualdad de género en posiciones de poder. Sin embargo, estudios han demostrado que cuando estos sesgos son identificados y corregidos, las evaluaciones de liderazgo se vuelven más justas y reconocen una gama más amplia de estilos de liderazgo, independientemente del género del individuo (e.g., Eagly y Carli, 2007).

Los avances en la metodología psicométrica, como el análisis del funcionamiento diferencial de los ítems (*Differential Item Functioning*, DIF), han permitido identificar con mayor precisión dónde y cómo ocurren estos sesgos, y han proporcionado herramientas para corregirlos (Gómez-Benito et al., 2010). A pesar de estos avances, aún queda trabajo para lograr garantizar que todos los test sean verdaderamente equitativos en términos de género. La investigación continua en este ámbito es crucial para seguir identificando y eliminando los sesgos de género en los instrumentos de evaluación psicológica.

El presente estudio se propone abordar la problemática de la equidad en los test desde una perspectiva de género, con el objetivo de identificar y analizar las buenas prácticas en psicometría que pueden contribuir a la reducción del sesgo de género en la construcción, validación y aplicación de estos instrumentos.

Por ello, se propone como objetivo general explorar y proponer buenas prácticas en el desarrollo psicométrico que aseguren que las diferencias observadas en las puntuaciones reflejen de manera precisa diferencias reales en los constructos medidos, y no sean producto de sesgos de género.

Específicamente, se propone realizar una revisión acerca de las metodologías y técnicas psicométricas utilizadas para identificar, prevenir y/o corregir el sesgo de género a partir de una selección aleatoria de artículos científicos publicados en revistas adscritas a Psicología, así como desarrollar un *checklist* que recoja un conjunto de buenas prácticas en el desarrollo, adaptación o evaluación de test que aseguren la equidad de género en las puntuaciones obtenidas.

Metodología

Diseño

El estudio sigue un diseño de revisión sistemática de literatura, centrado en la identificación y análisis de buenas prácticas en materia de equidad en estudios psicométricos. La revisión se ha focalizado en la evaluación de artículos publicados en 2023 en revistas académicas especializadas en psicometría o evaluación de pruebas psicológicas, analizando las técnicas aplicadas en cuestión de equidad, específicamente de género/sexo, en la construcción y validación de pruebas psicológicas. Debido a la dificultad de discernir si en los diferentes artículos evaluados se aborda el sexo o el género de los sujetos, en las siguientes secciones se utilizará la expresión sexo/género para evitar posibles errores de interpretación de resultados.

Procedimiento

El proceso de selección de artículos consistió en una búsqueda exhaustiva en las principales revistas científicas de psicometría. Las revistas incluidas en este estudio fueron seleccionadas por su relevancia en el campo de la evaluación psicológica y la psicometría, así como por su impacto en la publicación de estudios empíricos sobre desarrollo y validación de instrumentos. Las revistas seleccionadas para el análisis fueron:

- *European Journal of Psychological Assessment*
- *Journal of Psychoeducational Assessment*
- *Journal of Psychopathology and Behavioral Assessment*
- *Journal of Personality Assessment*
- *Psychological Assessment*
- *Assessment*
- *Psychometrika*

Se establecieron criterios de inclusión para asegurar que los estudios seleccionados abordaran específicamente la construcción y/o evaluación de instrumentos psicométricos. Se incluyeron aquellos artículos, publicados en el año 2023, que cumplen los siguientes criterios: (1) abordan el desarrollo o la evaluación de pruebas psicométricas y (2) hacen referencia explícita a cuestiones de equidad.

Se establecieron también criterios de exclusión para descartar estudios que no fueran relevantes para los objetivos del presente trabajo: (1) estudios teóricos sin análisis empírico, (2) artículos de revisión que no incluyeran datos cuantitativos o cualitativos sobre la validación de pruebas psicométricas y (3) investigaciones que no hicieran referencia a cuestiones de equidad, en la validación o aplicación de los instrumentos, en función del sexo o el género.

Se revisaron un total de 298 artículos, con la siguiente distribución en función de la revista donde se han publicado: 82 artículos en *Psychological Assessment*, 68 artículos en *Journal of Psychopathology and Behavioral Assessment*, 40 artículos en *Journal of Psychoeducational Assessment*, 35 artículos en *Psychometrika*, 32 artículos en *Assessment*, 32 artículos en *European Journal of Psychological Assessment* y nueve artículos en *Journal of Personality Assessment*.

A partir de la revisión del contenido de los artículos, se seleccionaron, en una primera fase de cribado, 210 artículos por abordar el desarrollo o evaluación de pruebas psicométricas. De estos artículos, 57 se centraron en distintas aproximaciones de la equidad, seleccionado de ellos los 20 estudios que abordaron específicamente la cuestión de sexo/género. De estos 20 artículos, cinco se centraron en el desarrollo y estudio de nuevas escalas, mientras que los 15 restantes abordaron escalas ya existentes.

El proceso de codificación de los artículos seleccionados fue realizado por un único codificador. Se aplicó un *checklist* diseñado para el estudio, que incluía criterios predefinidos relacionados con la equidad de género en psicometría, con una definición clara para aumentar la consistencia en la extracción de información.

Instrumento y Variables

El instrumento utilizado en el estudio fue una lista de verificación o *checklist*, diseñada específicamente para evaluar las buenas

prácticas en la validación de instrumentos psicométricos desde la perspectiva de la equidad de género en el contexto de este artículo (tabla 1). La *checklist* incluye los siguientes elementos clave:

1. Análisis descriptivos de ítems y DIF. Se analizó si los estudios reportan análisis descriptivos de los ítems segregados por sexo/género y/o análisis de DIF para identificar sesgos en los ítems relacionados con el sexo/género.
2. Análisis de fiabilidad. Se revisó si los estudios incluyeron análisis de consistencia interna (e.g., *alfa* de Cronbach) y test-retest, reportando datos desglosados por sexo/género.
3. Estudio de validez interna, en concreto, análisis de estructura interna. En este caso se incluye análisis factoriales exploratorios y confirmatorios, con atención a estudios de invarianza factorial en función del sexo/género.
4. Estudio de validez externa. En este apartado se evaluó si los estudios reportan evidencias de validez predictiva, diferenciando por sexo/género.
5. Otras evidencias de validez.

Tabla 1

Checklist Para Evaluar la Inclusión de la Equidad de Género en los Procesos de Evaluación Psicométrica

1. Análisis descriptivos de ítems

¿Se proporcionan estadísticas descriptivas (media, mediana, desviación estándar, etc.) desglosadas por sexo/género? Sí ☐ No ☐

¿Se evalúan efectos suelo/techo desglosados por sexo/género? Sí ☐ No ☐ No procede ☐

¿Se incluyen análisis inferenciales (e.g., t-test, ANOVA) para identificar diferencias significativas en función del sexo/género? Sí ☐ No ☐

¿Se reportan medidas del tamaño del efecto para diferencias por sexo/género? Sí ☐ No ☐

2. Funcionamiento diferencial del ítem (*Differential Item Functioning*, DIF)

¿Se ha realizado un análisis DIF para evaluar sesgos en función del sexo/género? Sí ☐ No ☐

Si se realizó DIF, ¿se identificaron y reportaron ítems con sesgos de sexo/género? Sí ☐ No ☐ No procede ☐

¿Se realizaron ajustes en los ítems o se eliminaron aquellos con sesgos diferenciales en función del sexo/género? Sí ☐ No ☐ No procede ☐

3. Análisis de fiabilidad

¿Se reportaron coeficientes de fiabilidad basados en consistencia interna (como *alfa* de Cronbach, Omega, etc.)? Sí ☐ No ☐

¿Se desglosaron los coeficientes de fiabilidad por sexo/género? Sí ☐ No ☐ No procede ☐

¿Se realizó un análisis de estabilidad temporal (test-retest) considerando el sexo/género de los participantes? Sí ☐ No ☐

¿Se observó alguna diferencia significativa en los coeficientes de fiabilidad en función del sexo/género? Sí ☐ No ☐ No procede ☐

4. Estudio de validez interna

¿Se realizó un análisis factorial exploratorio (AFE) o confirmatorio (AFC) para evaluar la estructura del test? Sí ☐ No ☐

¿Se incluyeron estudios de invarianza factorial en función del sexo/género? Sí ☐ No ☐ No procede ☐

¿Se estableció que la estructura del test es equivalente en función del sexo/género? Sí ☐ No ☐ No procede ☐

5. Estudio de validez externa

¿Se realizaron análisis de correlaciones o regresiones para evaluar la validez predictiva del test? Sí ☐ No ☐

¿Se incluyeron análisis de diferencias de sexo/género en las correlaciones o resultados de la validez predictiva? Sí ☐ No ☐ No procede ☐

6. Otras evidencias de validez

¿Se evaluó la validez de constructo considerando el sexo/género? Sí ☐ No ☐

¿Se discutieron posibles diferencias en la interpretación del constructo medido por sexo/género? Sí ☐ No ☐ No procede ☐

¿Se evaluó otras evidencias de validez, considerando el sexo/género? Sí ☐ No ☐ No procede ☐

7. Consideraciones generales sobre equidad de género

¿Se consideró el sexo/género como una variable importante en los análisis psicométricos? Sí ☐ No ☐

¿Se identificaron y corrigieron posibles sesgos de sexo/género en el instrumento? Sí ☐ No ☐

Resultados

Los resultados de esta revisión pretenden identificar en qué medida se están implementando buenas prácticas en materia de equidad de género en los test, evaluando 20 artículos psicométricos, publicados en revistas de referencia en esta materia mediante un *checklist* de siete dimensiones que se concretan en 21 ítems diseñado expresamente para este estudio (tabla 1) y mediante el cual se analiza la inclusión del sexo/género en los análisis descriptivos de ítems y el DIF, las evidencias de fiabilidad, validez interna, validez externa y, por último, otras aproximaciones de los test. La tabla 2 presenta los 20 artículos seleccionados sobre los que se ha efectuado el análisis y la tabla 3 presenta la sistematización de datos de los artículos seleccionados para la revisión.

Tabla 2

Artículos Seleccionados Para la Revisión Sistemática

Autoría	Año	Título	Revista
Ahmed, W.	2023	Measuring stress among black adolescents: Validation of perceived stress scale.	*Journal of Psychopathology and Behavioral Assessment*
Alkan, M.F., Sevim, F.O.M. y Evers, A.T.	2023	Factor structure and measurement invariance of the Teacher Autonomous Behavior Scale in Turkey.	*Journal of Psychoeducational Assessment*
Alshayea, A.K.	2023	Development and validation of an Arabic version of the World Health Organization Well-Being Index (WHO-5).	*Journal of Psychopathology and Behavioral Assessment*
Anghel, E., Mahalik, J.R. y Harris, M.P.	2023	Examining the measurement invariance of the Conformity to Masculine Norms Inventory (CMNI-30) by sexual orientation.	*Assessment*
Asgarabad, M.H., Yegaei, P.S., Ho, W.S. y Cheung, H.N.	2023	The gender invariance of Multidimensional Depression Assessment Scale in adolescents.	*Journal of Psychopathology and Behavioral Assessment*
Chen, Y., Li, C., Ouyang, J. y Xu, G.	2023	DIF statistical inference without knowing anchoring items.	*Psychometrika*
Dong, Y., Dumas, D., Clements, D.H., Day-Hess, C. A. y Sarama, J.	2023	Evaluating the consequential validity of the Research-Based Early Mathematics Assessment.	*Journal of Psychoeducational Assessment*
Feinstein, B.A., Khan, A., Chang, C.J. y Miller, S.A.	2023	Use of the Heterosexist Harassment, Rejection, and Discrimination Scale with different sexual orientation, gender, and racial/ethnic groups: An examination of measurement invariance.	*Assessment*
Fino, E., Popusoi, S.A., Holman, A.C., Iliceto, P. y Heym, N.	2023	Dimensionality, factorial invariance, and cross-cultural differential item functioning of the Short Dark Tetrad (SD4) in Italian, Romanian, and UK samples.	*European Journal of Psychological Assessment*
Hsiao, Y.Y., Qi, C.H., Dale, P.S., Bulotsky-Shearer, R. y Wang, Q.	2023	Measuring behavior problems in children from low-income families: A Rasch analysis of the Child Behavior Checklist for ages 1½-5.	*Journal of Psychoeducational Assessment*
Lau, C., Chiesi, F., Fermani, A., Muzi, M., del Moral Arroyo, G., Bruno, F., Ruch, W., Quilty, L.C., Saklofske, D.H. y Canestrari, C.	2023	Measuring gelotophobia, gelotophilia, and katagelasticism in Italy and Canada using PhoPhiKat-30: A multidimensional item response theory and differential item functioning analysis.	*European Journal of Psychological Assessment*
Li, N., Hein, S., Cavitt, J., Chapman, J., Geib, C.F. y Grigorenko, E.L.	2023	Applying item response theory analysis to the SAVRY in justice-involved youth.	*Assessment*
Liu, D., Wang, Y. y Li, C.	2023	Development and validation of the Work Orientation Questionnaire Short-Form (WOQ-SF): Evidence from China.	*European Journal of Psychological Assessment*
Liu, L. y Sun, J.	2023	Gender and age invariance of the Global Belief in a Just World Scale.	*European Journal of Psychological Assessment*
Martin, J.A., Tarantino, D.M. y Levy, K.N.	2023	Investigating gender-based differential item functioning on the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD): An item response theory analysis.	*Psychological Assessment*
Moron, M., Mozgol, L., Gajda, A.N., Rode, M., Biela, M., Stalmach, K., Kuchta, W. y Marsee, M.	2023	Forms and functions of aggression in young adults: The Polish modified version of the Peer Conflict Scale.	*Journal of Psychopathology and Behavioral Assessment*
Ober, T.M., Lu, Y., Blacklock, C.B., Liu, C. y Cheng, Y.	2023	Development and validation of a cognitive load measure for general educational settings.	*Journal of Psychoeducational Assessment*
Prati, G. y Mancini, A.D.	2023	Social and behavioral consequences of the COVID-19 pandemic: Validation of a Pandemic Disengagement Syndrome Scale (PDSS) in four national contexts.	*Psychological Assessment*
Shin, H., Shah, P. y Preston, S.D.	2023	The Reasoning through Evidence versus Advice (EvA) Scale: Scale development and validation.	*Journal of Personality Assessment*
Yaremych, H.E. y Persky, S.	2023	Development and validation of the Parental Food Choice Guilt Scale.	*European Journal of Psychological Assessment*

Tabla 3

Sistematización de Datos de los Artículos Seleccionados Para la Revisión

Autoría	Distribución por sexo/ género	Características relevantes de la muestra	Análisis descriptivos de ítems y DIF	Análisis de fiabilidad	Estudio de validez interna	Estudio de validez externa	Otras evidencias de validez
Ahmed	Hombres: 562 Mujeres: 608	Adolescentes afroamericanos en EE.UU.	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, desglosado por sexo/ género.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.
Alkan et al.	Hombres: 333 Mujeres: 378	Docentes de Turquía.	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE y AFC, con análisis de invarianza en función del sexo/género.	Se reporta relación con otras variables no desglosado por sexo/ género.	No se mencionan.
Alshayea	Hombres: 77 Mujeres: 113	Población adulta de Arabia Saudí	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFC, no desglosado o analizada la invarianza por sexo/ género.	Se reportan evidencias de validez convergente y discriminante (controlando sexo/género), y se reportan diferencias de sexo/género como de validez externa.	No se mencionan.
Anghel et al.	Solo hombres. Hetero: 553, Gay: 160, Bisexual: 163.	Población adulta de EEUU.	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.
Asgarabad et al.	Hombres: 1031 Mujeres: 1850	Población adolescente de Hong Kong, China y Reino Unido.	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE y AFC, con análisis de invarianza en función del sexo/género.	Se reporta relación con otras variables mediante análisis de evidencia convergente y discriminante, no desglosado por sexo/ género.	No se mencionan.
Chen et al.	Hombres: 609 Mujeres: 832	Población genérica de Reino Unido (no especificada)	Análisis DIF mediante modelos MIMIC, sin descriptivos de los ítems segregados por sexo/género.	No se reportan indicadores de análisis de fiabilidad.	No se reportan análisis de validez interna.	No se reportan evidencias de validez externa.	No se mencionan.
Dong et al.	Hombres: 307 Mujeres: 320	Estudiantes de primaria de EE.UU.	Análisis descriptivo de ítems no segregado y DIF por submuestras culturales (no por sexo/género).	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	No se reportan análisis de validez interna.	Se reporta evidencia de validez externa basada en el Índice de Validez de Contenido (CVR), considerando la interacción con el sexo/género.	No se mencionan.
Feinstein et al.	Hombres cis: 340, Mujeres cis: 334, Transgénero y género diverso: 118	Estudio longitudinal de adultos jóvenes pertenecientes a minorías sexuales de Estados Unidos.	No se reporta DIF ni análisis descriptivo segregado.	No se reportan indicadores de análisis de fiabilidad.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.
Fino et al.	Hombres: 300 Mujeres: 300	Estudiantes universitarios de Italia, Rumanía y Reino Unido.	Análisis descriptivo de ítems no segregado y DIF por submuestras culturales (no por sexo/género).	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta la evaluación de la estructura basada en TRI multidimensional, no desglosado o analizada la invarianza por sexo/género.	No se reportan evidencias de validez externa.	Evidencia de validez cruzada con diferentes muestras culturales, no por sexo/ género.
Hsiao et al.	Hombres: 121 Mujeres: 123	Estudiantes de infantil de EE.UU.	No reporta análisis descriptivo segregado. Análisis DIF por sexo/género.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Evaluación unidimensional (modelo Rasch) e información sobre independencia local, no analizada la invarianza por sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.

Autoría	Distribución por sexo/ género	Características relevantes de la muestra	Análisis descriptivos de ítems y DIF	Análisis de fiabilidad	Estudio de validez interna	Estudio de validez externa	Otras evidencias de validez
Lau et al.	Hombres: 62 Mujeres: 264	Población universitaria de Italia.	Análisis DIF con submuestras culturales (no por sexo/género) y descriptivo de ítems no segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE, no desglosado o analizada la invarianza por sexo/ género.	Evidencia de validez externa basada en las diferencias de sexo/género en los factores latentes, mediante regresiones lineales bayesianas.	No se mencionan.
Li et al.	Hombres: 665 Mujeres: 208	Adolescentes de EE.UU. involucrados en el sistema de justicia juvenil.	No reporta análisis descriptivo segregado. Análisis DIF por sexo/género.	No se reportan indicadores de análisis de fiabilidad.	Se reporta la evaluación unidimensional basada en GRM (*Graded Response Model*), no analizada la invarianza por sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.
Liu et al.	Hombres: 330 Mujeres: 417	Población adulta de China.	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE y AFC, con análisis de invarianza en función del sexo/género.	Se reportan evidencias de validez externa (relación con otras variables), no desglosado por sexo/ género.	Se informa sobre la traducción, adaptación y reducción del número de ítems, sin perspectiva de género.
Liu y Sun	Hombres: 1199 Mujeres: 929	Población adulta trabajadora de China	Se reportan diferencias en las variables latentes. No DIF, ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	No se reportan evidencias de validez externa.	No se mencionan.
Martin et al.	Hombres: 7734 Mujeres: 14301	Posgraduados de EEUU.	Se reporta análisis descriptivo y DIF segregado por sexo/género.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Evaluación unidimensional basada en TRI, no analizada la invarianza por sexo/ género.	No se reportan evidencias de validez externa.	No se mencionan.
Moron et al.	Hombres: 228 Mujeres: 466	Población adulta joven de Polonia.	Se analizan diferencias de sexo/género a nivel de factores latentes. Sin descriptivos de los ítems ni DIF segregados por sexo/género.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	Se reportan evidencias de relación con otras variables, etiquetada como validez de constructo, no desglosado por sexo/género.	No se mencionan.
Ober et al.	Hombres: 252 Mujeres: 524	Estudiantes de secundaria de EE.UU.	Se analizan diferencias de sexo/género a nivel de factores latentes, pero sin descriptivos de los ítems ni DIF segregados por sexo/género.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFC, con análisis de invarianza en función del sexo/género.	Se reportan evidencias de validez externa (relación con otras variables) etiquetada como validez de constructo, no desglosado por sexo/género.	Se reporta el Esfuerzo en el Tiempo de Respuesta (RTE), no desglosado por sexo/ género.
Prati y Mancini	Hombres: 672 Mujeres: 671	Población adulta de diversos países (EEUU, Italia, Suiza y Noruega).	No se reporta DIF ni análisis descriptivo segregado.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE y AFC, con análisis de invarianza en función del sexo/género.	Evidencias de relación con otras variables mediante análisis de evidencia convergente y discriminante, no desglosado por sexo/ género.	No se mencionan.

Autoría	Distribución por sexo/ género	Características relevantes de la muestra	Análisis descriptivos de ítems y DIF	Análisis de fiabilidad	Estudio de validez interna	Estudio de validez externa	Otras evidencias de validez
Shin et al.	Hombres: 400 Mujeres: 431	Población adulta de EEUU.	No se reporta análisis descriptivo segregado. Análisis DIF por sexo/género.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE y AFC, con análisis de invarianza en función del sexo/género.	Evidencias de relación con otras variables mediante análisis de evidencia convergente y discriminante, no desglosado por sexo/ género.	Se reporta evidencias de validez cruzada con estudios previos.
Yaremych y Persky	Hombres: 150 Mujeres: 150	Estudiantes de secundaria de EE.UU. y sus padres/madres.	No hay análisis descriptivo de ítems y se reporta DIF por sexo/ género de los progenitores.	Coeficiente de fiabilidad reportado, no desglosado por sexo/género.	Se reporta AFE, no desglosado o analizada la invarianza por sexo/ género.	Evidencias de validez externa (relación con otras variables mediante análisis de evidencia convergente y discriminante), no desglosado por sexo/ género.	Se reporta evidencias basadas en el Modelo de respuesta graduada (GRM), no desglosado por sexo/ género.

Análisis Descriptivos de Ítems y DIF

El análisis descriptivo de los ítems es una herramienta fundamental para comprender el comportamiento de los datos en diferentes grupos. Es recomendable no solo proporcionar medidas como la media, la mediana y la desviación estándar, sino también analizar posibles efectos suelo/techo que podrían indicar sesgos en la respuesta de los ítems. En investigaciones que abordan la equidad por sexo/género, sería relevante no solo presentar estos datos de manera global, sino también desglosarlos por sexo/género, acompañándolos de análisis inferenciales y medidas de tamaño del efecto para determinar si existen diferencias significativas entre hombres y mujeres, pudiendo analizarse también dichas diferencias en las variables latentes (factores). En la revisión realizada de los artículos que incluyeron el abordaje de la equidad en función del sexo/género, únicamente un artículo de los 20 analizados incluyó esta información a nivel de ítems (Martin et al., 2023) y tres de 20 en función de los factores latentes (Liu y Sun, 2023; Moron et al., 2023; Ober et al., 2023).

Por otro lado, el análisis DIF resulta crucial para identificar posibles sesgos en los ítems. El DIF permite detectar si un ítem tiene un rendimiento diferencial en función del sexo/género, lo que indicaría que el ítem mide de manera desigual a diferentes grupos, independientemente de sus habilidades reales. De los 20 estudios revisados, nueve mencionaron la aplicación de análisis DIF, pero solo en seis artículos se consideraron específicamente las diferencias por sexo/género (Chen et al., 2023; Hsiao et al., 2023; Li et al., 2023; Martin et al., 2023; Shin et al., 2023; Yaremych y Persky, 2023).

Análisis de Fiabilidad

La fiabilidad es un aspecto crítico para evaluar la consistencia de las medidas obtenidas a partir de un test. La mayoría de las investigaciones que estudian las propiedades psicométricas de escalas suelen basarse en coeficientes de fiabilidad a partir del enfoque en la consistencia interna, utilizando el *alfa* de Cronbach o el coeficiente Omega, entre otros. Además, se utilizan otros enfoques, como el de la estabilidad temporal (basado en el test-retest, por ejemplo) y en el de formas paralelas, menos frecuente por las dificultades de obtención de formas realmente paralelas. No obstante, un enfoque centrado en la equidad de género debe ir más allá del reporte global de estos coeficientes, en cualquiera de sus aproximaciones, y debería incluir un desglose por sexo/género para determinar si las puntuaciones del test son igualmente fiables tanto en hombres como en mujeres.

De los 20 estudios revisados, 17 reportaron coeficientes de fiabilidad basados en consistencia interna. Sin embargo, solo en un estudio se ofreció un desglose de los coeficientes por sexo/género (Ahmed, 2023), no realizando comparaciones inferenciales en las diferencias entre los coeficientes.

Estudio de Validez Interna

La validez interna se refiere a la estructura del test y cómo los ítems reflejan el constructo teórico que se pretende medir. Este aspecto se suele evaluar mediante análisis factoriales exploratorios (AFE) y confirmatorios (AFC). En el contexto de la equidad de género, la inclusión de estudios de invarianza factorial es particularmente relevante dado que permite evaluar si la estructura del test se puede considerar equivalente para hombres y mujeres.

En la revisión, 18 de los 20 estudios incluidos emplearon análisis de la estructura interna (ya sea mediante aplicación de AFE, AFC o técnicas derivadas de la TRI), de los que 11 incluyeron estudios de invarianza factorial en función del sexo/género (Ahmed, 2023; Alkan et al., 2023; Anghel et al., 2023; Asgarabad et al., 2023; Feinstein et al., 2023; Liu y Sun, 2023; Liu et al., 2023; Moron et al., 2023; Ober et al., 2023; Prati y Mancini, 2023; Shin et al., 2023).

Estudio de Validez Externa

La validez externa se refiere a la capacidad del test para predecir o correlacionarse con otras medidas o indicadores. Para garantizar la equidad, sería recomendable incluir análisis de diferencias de sexo/género en estas correlaciones, lo que podría realizarse a través de comparaciones entre correlaciones por grupo.

En las referencias analizadas, 11 de los 20 estudios proporcionaron evidencias de validez predictiva en sus diferentes aproximaciones: convergente, divergente o de criterio, entre otras. Sin embargo, ninguno incluyó un análisis de diferencias por género en este apartado.

Otras Evidencias de Validez

Además de las formas tradicionales de validez interna y externa, algunos estudios proporcionaron otras formas de evidencia de validez, como la validez de constructo, que ayudan a determinar el grado en el que el test mide la complejidad del constructo analizado. Sin embargo, no se realizaron distinciones explícitas por sexo/género en estas formas de evidencias de validez.

Discusión

Los resultados obtenidos a partir del análisis de los estudios revisados muestran una clara falta de atención a la equidad de género en diversas etapas del desarrollo y validación de test psicológicos, lo que contraviene las recomendaciones de la *American Psychological Association* (APA) y otras organizaciones internacionales desde hace más de una década (AERA, APA y NCME, 2014).

Aunque algunos estudios han hecho esfuerzos por incluir análisis DIF o invarianza factorial en función del sexo/género, la mayoría no abordan sistemáticamente este aspecto. Esta omisión es preocupante, ya que puede llevar a resultados sesgados que perpetúan desigualdades de género, afectando tanto la validez como la fiabilidad de las pruebas psicológicas. Los test son fundamentales para la toma de decisiones en contextos educativos, clínicos, laborales y sociales, por lo que la equidad de género debería ser un pilar central en su desarrollo y validación (Helms, 2006).

Análisis descriptivo de ítems y DIF. El análisis descriptivo de los ítems, desglosado por sexo/género, es crucial para garantizar la equidad en las pruebas psicológicas. Cuando se complementa con el DIF, estas técnicas facilitan la detección de diferencias de impacto entre los grupos evaluados, asegurando que las diferencias observadas en las puntuaciones reflejen diferencias reales en los constructos psicológicos y no sesgos sistemáticos en los ítems (Zumbo y Chan, 2014). Sin embargo, la revisión muestra que solo un pequeño porcentaje de los estudios incluye este tipo de análisis, lo que indica que no se está abordando adecuadamente la influencia del género en la respuesta a los ítems. Esto es un vacío metodológico significativo, ya que el uso limitado del DIF compromete tanto la validez como la fiabilidad de las pruebas para diferentes grupos (Hyde, 2005).

Fiabilidad. A pesar de que la mayoría de los estudios revisados reportaron análisis de consistencia interna a través de coeficientes como el *alfa* de Cronbach, solo uno de ellos desglosó estos coeficientes por sexo/género. Esto representa una omisión crítica, ya que no diferenciar entre hombres y mujeres puede ocultar problemas de fiabilidad específicos para cada grupo. La consistencia interna es clave para asegurar la estabilidad y precisión de las puntuaciones, pero si no se examina de manera diferenciada por sexo/género, no es posible garantizar que los test sean igualmente fiables para ambos géneros (Cronbach, 1951). Esta falta de análisis impide la identificación de posibles sesgos que puedan comprometer la equidad en la interpretación de los resultados.

Validez interna y externa. El análisis factorial exploratorio y confirmatorio es una técnica común para evaluar la validez interna de los test (Elosúa, 2003), pero solo algunos estudios revisados incluyeron análisis de invarianza factorial para determinar si la estructura del test es la misma para hombres y mujeres. Este tipo de análisis es esencial, ya que sin él no es posible garantizar que los test midan de manera equivalente los constructos psicológicos en ambos géneros. A pesar de que la invarianza factorial es una técnica bien documentada y fácilmente aplicable en muchos software psicométricos, su uso sigue siendo limitado en el contexto de la equidad de género, lo que es un área de mejora significativa.

En cuanto a la validez externa, la falta de análisis diferenciados por género es una preocupación importante. Si no se verifica si el test predice con la misma precisión los resultados en hombres y mujeres, existe el riesgo de perpetuar desigualdades en la toma de decisiones que dependen de los resultados de estos test, como la admisión a programas educativos o la selección de personal en el ámbito laboral (Hyde, 2005). La simpleza de realizar este tipo de análisis, como un análisis de moderación o la aplicación de pruebas como la Z de Fisher y el q de Cohen, refuerza la necesidad de incluirlos en los estudios psicométricos para garantizar la equidad de género en los resultados.

Es importante señalar que los resultados de esta revisión se basan exclusivamente en estudios publicados en 2023 en un conjunto específico de revistas. Si bien esta decisión ha permitido ofrecer una fotografía actualizada de las prácticas más recientes en equidad de género dentro de la psicometría, no resulta posible evaluar los avances históricos ni identificar tendencias a lo largo del tiempo. Además, la selección de revistas, aunque representativa del campo, puede no reflejar la totalidad de las investigaciones existentes en otras fuentes o contextos. Por lo tanto, los resultados obtenidos deben interpretarse con cautela, ya que podrían reflejar tanto las prácticas habituales actuales como las limitaciones derivadas del marco temporal y muestral del estudio. Aunque se han identificado evidencias en la aplicación de técnicas que abordan la equidad de género, no puedo determinarse, por tanto, la evolución histórica de estas prácticas ni establecer comparaciones con periodos anteriores. Estudios futuros que amplíen el rango temporal y de revistas seleccionadas podrán ofrecer una visión más robusta y completa de la evolución en la implementación de la equidad en los estudios psicométricos.

En cuanto a otras limitaciones del presente estudio, cabe destacar en primer que el proceso de codificación fue realizado por un único codificador. Si bien se aplicó un protocolo sistemático y detallado para reducir posibles sesgos en la extracción y análisis de datos, la falta de evaluación de la concordancia entre codificadores podría haber influido en la interpretación de los resultados.

Conclusiones

Los resultados de esta revisión destacan la necesidad urgente de integrar de manera más explícita y sistemática la perspectiva de género en los estudios psicométricos, pudiendo partir esta iniciativa desde las propias revistas científicas que publican de forma recurrente este tipo de estudios. Si bien se han observado algunos ejemplos en la inclusión de análisis como el DIF y la invarianza factorial, estos esfuerzos son aún insuficientes para garantizar una fuerte base de evidencias que garanticen la equidad de género en los test utilizados. Es básico además que en el desarrollo de pruebas psicológicas se adopte de forma explícita un enfoque más inclusivo que considere las diferencias de género en todas las fases de construcción y validación de los instrumentos, desde la definición y operativización del constructo hasta el desarrollo de baremos y la interpretación de los resultados.

Para avanzar hacia una mayor equidad en la psicometría, es fundamental aplicar buenas prácticas como las que se sugieren en este estudio. Estas incluyen la inclusión y/o desagregación de los análisis por sexo/género en todas las etapas del desarrollo y validación de los test, la implementación de estudios de invarianza factorial y DIF, y la evaluación de la fiabilidad y validez de los test en función del género, entre otras aproximaciones. Solo mediante la adopción de estas prácticas será posible garantizar que los test no perpetúen estereotipos ni desigualdades de género, y que reflejen de manera justa y precisa las capacidades y características de todas las personas evaluadas (AERA, APA y NCME, 2014).

Conflicto de Intereses

El autor declara que no existe conflicto de intereses.

Financiación

El presente trabajo no recibió financiación específica de agencias del sector público, comercial o de organismos no gubernamentales.

Referencias

Se señalan con * las referencias incluidas en la revisión sistemática.

AERA, APA y NCME (American Educational Research Association, American Psychological Association y National Council on Measurement in Education) (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/spanish_standards_pdf.pdf

Ahmed, Wondimu (2023). Measuring stress among Black adolescents: Validation of perceived stress scale. *Journal of Psychopathology and Behavioral Assessment*, *45*(3), 385-397. https://doi.org/10.1007/s10862-023-10079-z *

Alkan, Muhammet F.; Sevim, Fazilet O.M. y Evers, Arnoud T. (2023). Factor structure and measurement invariance of the Teacher Autonomous Behavior Scale in Turkey. *Journal of Psychoeducational Assessment*, *41*(5), 491-506. https://doi.org/10.1177/07342829231186229 *

Alshayea, Ahmad K. (2023). Development and validation of an Arabic version of the World Health Organization Well-Being Index (WHO-5). *Journal of Psychopathology and Behavioral Assessment*, *45*(2), 192-205. https://doi.org/10.1007/s10862-023-10027-x *

Anastasi, Anne y Urbina, Susana (1997). *Psychological testing* (7th Ed.). Prentice Hall/Pearson Education.

Anghel, Ella; Mahalik, James R. y Harris, Michael P. (2023). Examining the measurement invariance of the Conformity to Masculine Norms Inventory (CMNI-30) by sexual orientation. *Assessment*, *30*(5), 1086-1100. https://doi.org/10.1177/10731911221149085 *

Asgarabad, Mojtaba H.; Yegaei, Pardis S.; Ho, W.S. y Cheung, Ho N. (2023). The gender invariance of Multidimensional Depression Assessment Scale in adolescents. *Journal of Psychopathology and Behavioral Assessment*, *45*(3), 398-412. https://doi.org/10.1007/s10862-023-10040-0 *

Camilli, Gregory y Shepard, Lorrie (1994). *Methods for identifying biased test items* (vol. 4). Sage.

Caprile, María; Addis, Elisabetta; Castaño, Celia; Klinge, Ineke; Larios, Marina; Meulders, Daniele; Vázquez-Cupeiro, Susana (2012). Meta-analysis of gender and science research: Synthesis report. *European Union Publications Office.* https://op.europa.eu/en/publication-detail/-/publication/3516275d-c56d-4097- abc3-602863bcefc8

Chen, Yunxiao; Li, Chengcheng; Ouyang, Jing y Xu, Gongjun (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, *88*(3), 601-626. https://doi.org/10.1007/s11336-023-09930-9 *

Cronbach, Lee J.(1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Delgado-Álvarez, Carmen (2020). La ceguera al género inducida por la ceguera a los estándares de medición. Comentario a Ferrer-Pérez y Bosch-Fiol, 2019. *Anuario de Psicología Jurídica, 30*(1), 93-96. https://doi.org/10.5093/apj2019a8

Dong, Yixiao; Dumas, Denis; Clements, Douglas H.; Day-Hess, Crystal A. y Sarama, Julie (2023). Evaluating the consequential validity of the Research-Based Early Mathematics Assessment. *Journal of Psychoeducational Assessment*, *41*(5), 507-522. https://doi.org/10.1177/07342829231165812 *

Eagly, Alice y Carli, Linda L. (2007). *Through the labyrinth: The truth about how women become leaders*. Harvard Business School Press.

Elosúa, Paula (2003). Sobre la validez de los tests. *Psicothema*, *15*(2), 315-321.

Feinstein, Brian A.; Khan, Aaminah; Chang, Cindy J. y Miller, Steven A. (2023). Use of the Heterosexist Harassment, Rejection, and Discrimination Scale with different sexual orientation, gender, and racial/ethnic groups: An examination of measurement invariance. *Assessment*, *30*(5), 1175-1191. https://doi.org/10.1177/10731911231156135 *

Fernández-Ballesteros, Rocío (2008). Introducción a la evaluación psicológica. En R. Fernández-Ballesteros (Ed.), *Evaluación psicológica: concepto, métodos y estudio de casos* (2ª Ed., pp. 21-45). Pirámide.

Fino, Emanuele; Popusoi, Simona A.; Holman, Andrei C.; Iliceto, Paolo y Heym, Nadja (2023). Dimensionality, factorial invariance, and cross-cultural differential item functioning of the Short Dark Tetrad (SD4) in Italian, Romanian, and UK samples. *European Journal of Psychological Assessment*, *39*(1), 44-56. https://doi.org/10.1027/1015-5759/a000775 *

Gómez-Benito, Juana; Hidalgo, María Dolores y Guilera, Georgina (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31*(1), 75-84. https://www.papelesdelpsicologo.es/pdf/1798.pdf

Groth-Marnat, Gary y Wright, A. Jordan (2016). *Handbook of psychological assessment* (6th Ed.). John Wiley & Sons.

Helms, Janet E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845-859. https://doi.org/10.1037/0003-066X.61.8.845

Hsiao, Yu-Yu; Qi, Cathy Huaqing; Dale, Philip S.; Bulotsky-Shearer, Rebecca y Wang, Qing (2023). Measuring behavior problems in children from low-income families: A Rasch analysis of the Child Behavior Checklist for ages 1½-5. *Journal of Psychoeducational Assessment*, *41*(4), 397-413. https://doi.org/10.1177/07342829231162216 *

Hyde, Janet S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581-592. https://doi.org/10.1037/0003-066X.60.6.581

Lau, Chloe; Chiesi, Francesca; Fermani, Alessandra; Muzi, Morena; del Moral Arroyo, Gonzalo; Bruno, Francesco; Ruch, Willibald; Quilty, Lena C.; Saklofske, Donald H. y Canestrari, Carla (2023). Measuring gelotophobia, gelotophilia, and katagelasticism in Italy and Canada using PhoPhiKat-30: A multidimensional item response theory and differential item functioning analysis. *European Journal of Psychological Assessment*, *39*(2), 79-97. https://doi.org/10.1027/1015-5759/a000787 *

Li, Nan; Hein, Sascha; Cavitt, Joslyn; Chapman, John; Geib, Catherine Foley y Grigorenko, Elena L.L. (2023). Applying item response theory analysis to the SAVRY in justice-involved youth. *Assessment*, *30*(5), 1192-1209. https://doi.org/10.1177/10731911221146120 *

Liu, Doudou; Wang, Yiming y Li, Chaoping (2023). Development and validation of the Work Orientation Questionnaire Short-Form (WOQ-SF): Evidence from China. *European Journal of Psychological Assessment*, *39*(3), 163-177. https://doi.org/10.1027/1015-5759/a000814 *

Liu, Lei y Sun, Jianmin (2023). Gender and age invariance of the Global Belief in a Just World Scale. *European Journal of Psychological Assessment*, *39*(2), 98-108. https://doi.org/10.1027/1015-5759/a000811 *

Martin, Jacob A.; Tarantino, Danielle M. y Levy, Kenneth N. (2023). Investigating gender-based differential item functioning on the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD): An item response theory analysis. *Psychological Assessment*, *35*(3), 263-275. https://doi.org/10.1037/pas0001229 *

Messick, Samuel (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Moron, Marcin; Mozgol, Ludwika; Gajda, Anna N.; Rode, Magdalena; Biela, Marta; Stalmach, Kamila; Kuchta, Weronika; Marsee, Monica y Vagos, Paula (2023). Forms and functions of aggression in young adults: The Polish modified version of the Peer Conflict Scale. *Journal of Psychopathology and Behavioral Assessment*, *45*(2), 206-218. https://doi.org/10.1007/s10862-023-10053-9 *

Muñiz, José (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo, 30*(1), 57-66. https://papelesdelpsicologo.es/pdf/1796.pdf

Nunnally, Jum y Bernstein, Ira (1994) The assessment of reliability. *Psychometric Theory, 3*, 248-292.

Ober, Teresa M.; Lu, Yikai; Blacklock, Chessley B.; Liu, Cheng y Cheng, Ying (2023). Development and validation of a cognitive load measure for general educational settings. *Journal of Psychoeducational Assessment*, *41*(5), 523-538. https://doi.org/10.1177/07342829231169171 *

Prati, Gabriele y Mancini, Anthony D. (2023). Social and behavioral consequences of the COVID-19 pandemic: Validation of a Pandemic Disengagement Syndrome Scale (PDSS) in four national contexts. *Psychological Assessment*, *35*(3), 305-317. https://doi.org/10.1037/pas0001213 *

Prieto, Gerardo y Delgado, Ana R. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31*(1), 67-74. https://papelesdelpsicologo.es/pdf/1797.pdf

Shin, Hwayong; Shah, Priti y Preston, Stephanie D. (2023). The reasoning through Evidence versus Advice (EvA) Scale: Scale development and validation. *Journal of Personality Assessment*, *105*(5), 636-649. https://doi.org/10.1080/00223891.2023.2297266 *

Yaremych, Haley E. y Persky, Susan (2023). Development and validation of the Parental Food Choice Guilt Scale. *European Journal of Psychological Assessment*, *39*(2), 109-122. https://doi.org/10.1027/1015-5759/a000800 *

Zumbo, Bruno D. y Chan, Eric (2014). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. En Bruno D. Zumbo y Eric Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 3-8). Springer International Publishing.