REVISIÓN SISTEMÁTICA DE RECOMENDACIONES PARA LA CONSTRUCCIÓN Y ADAPTACIÓN DE INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA

SYSTEMATIC REVIEW OF RECOMMENDATIONS FOR THE DEVELOPMENT AND ADAPTATION OF PSYCHOLOGICAL ASSESSMENT TOOLS

Deleersnyder, Guido[1]; Fritz, M. Sol; Laguna, Sofía[2]; Mikulic, Isabel M.[3]

RESUMEN

Dentro de la psicometría, las revisiones bibliográficas de artículos instrumentales suelen resaltar múltiples falencias metodológicas y de reporte de datos. Por ello, es necesario recopilar sugerencias concernientes a estos aspectos. **Objetivos:** Examinar recomendaciones para la construcción y adaptación de instrumentos de evaluación psicológica y conocer las tendencias metodológicas dentro del área. **Metodología:** Se siguieron los lineamientos PRISMA. Se realizó una búsqueda en la base PubMed y el repositorio de CONICET de todos los trabajos publicados hasta mayo de 2024. Criterios de inclusión: guías prácticas, revisiones bibliográficas y estudios que establecieran recomendaciones; escritos en español, portugués o inglés; publicados en revistas con referato. **Resultados:** Se obtuvieron 2159 artículos. Ocho cumplieron con los criterios establecidos. Se evaluó la metodología, el riesgo de sesgo y la transparencia del reporte de datos. **Discusión:** Se identifican y describen distintas recomendaciones. A su vez, se delimitan direcciones futuras para el área y próximas líneas de trabajo.

Palabras clave:
Construcción, Adaptación, Psicometría, Guías, Revisión sistemática.

ABSTRACT

Regarding psychometrics, reviews of instrumental articles often highlight multiple methodological and data reporting shortcomings. Therefore, it is necessary to collect suggestions concerning these aspects. **Aims:** To examine recommendations for the construction and adaptation of psychological assessment instruments identifying methodological trends within the area. **Methodology:** The PRISMA Statement was followed. A search of all the articles published up to May 2024 was carried out in the PubMed database and the CONICET repository. Criteria to be included: practical guides, literature reviews and studies that established recommendations; written in Spanish, Portuguese or English; published in refereed journals. **Results:** 2159 articles were obtained. Eight met the established criteria. The methodology, risk of bias, and transparency of data reporting were assessed. **Discussion:** Different recommendations are identified and described. In addition, future directions for the area and following research focuses are established.

Keywords:
Construction, Adaptation, Psychometrics, Guidelines, Systematic review.

[1]Universidad de Buenos Aires (UBA), Facultad de Psicología, Práctica de Investigación, Evaluación Psicológica en Contexto. Email deleersnyderguido@psi.uba.ar
[2]Universidad de Buenos Aires (UBA), Facultad de Psicología, Teoría y Técnica de Exploración y Diagnóstico Psicológico Modulo I - Cátedra I.
[3]Universidad de Buenos Aires, Facultad de Psicología, Práctica de Investigación, Evaluación Psicológica en Contexto. Universidad de Buenos Aires (UBA), Facultad de Psicología, Teoría y Técnica de Exploración y Diagnóstico Psicológico Modulo I - Cátedra I.

INTRODUCCIÓN

La psicometría es una rama de la psicología experimental que se ocupa de medir y cuantificar conceptos psicológicos tales como inteligencias, personalidad, habilidades, atributos, características, competencias y capacidades cognitivas, entre otros (Martínez Corso & Villota Burgos, 2022). Para llevar a cabo estas tareas se propone construir, adaptar y/o validar técnicas sustentadas en la estadística que midan de manera indirecta, pero válida y confiable, dichas variables (Gallo, 2018; Martínez Corso & Villota Burgos, 2022).

La medición es una parte crucial de la ciencia ya que permite comparar resultados (Johansson et al., 2023). Así, el modelo científico plantea que los requisitos indispensables de una técnica de evaluación psicológica son la validez y confiabilidad (Fernández Ballesteros, 2013; Lira & Caballero, 2020). Validez refiere al grado de evidencia y teoría que respaldan las interpretaciones de las puntuaciones del test en los contextos de uso. Por su parte, la confiabilidad es definida como la consistencia de las puntuaciones al replicar el procedimiento de evaluación (Ahmed & Ishtiaq, 2021).

Es sabido que la mayoría de los instrumentos utilizados en psicología provienen de países anglosajones (Lira & Caballero, 2020). Por ello, los informes de los procesos metodológicos de construcción, adaptación y validación son necesarios para garantizar que la interpretación de los resultados sea válida, confiable y replicable en el contexto para el cual la técnica se utiliza (International Test Commission [ITC], 2014; Johansson et al., 2023; Lira & Caballero, 2020). Existe vasta evidencia de que el uso de instrumentos bien construidos e interpretados presenta sustanciales beneficios tanto para los profesionales que los administran como para los usuarios, ya que permiten tomar mejores decisiones respecto al proceso de evaluación (American Educational Research Association [AERA], 2018). Es fundamental resaltar la importancia de dichos procesos en el contexto local, donde se detecta la amplia utilización de instrumentos extranjeros sin la adecuada adaptación y validación (Mikulic et al., 2022).

La estandarización y exactitud son cuestiones esenciales en todas las etapas de la evaluación, desde el desarrollo y administración de las técnicas, hasta la puntuación, análisis e interpretación, así como el reporte (ITC, 2014). Sin embargo, no pareciera existir un consenso acerca de cómo informar las propiedades psicométricas de las técnicas de manera sistemática. En una revisión realizada sobre 700 publicaciones sobre construcción, adaptación y validación de técnicas, se detectaron reportes adecuados del proceso metodológico solamente en 17 de ellas, lo cual representa un 2,4% (Guillermin et al., 1993).

Al realizar publicaciones sobre técnicas de evaluación psicológica será fundamental que estas sean lo más claras posibles y en todo momento se rijan por los lineamientos ya consensuados al respecto. La falta de esfuerzos por mejorar las prácticas de medición y evaluación asociadas a la calidad psicométrica del instrumental evaluativo utilizado acrecienta el riesgo de generar problemas a largo plazo. Uno de ellos es la obtención de resultados no comparables ni replicables, lo cual lleva a interpretaciones confusas y retraso en los avances científicos que pueden acabar perjudicando a la sociedad (Johanssen et al., 2023). Por este motivo, es de suma importancia mejorar los reportes de las investigaciones que se realizan.

Es el creciente número de falencias identificadas en los artículos que contienen datos reportados de revisiones de instrumentos de evaluación, lo que requiere revisar tanto los criterios metodológicos como el reporte de información en sí mismo. Por ello, los objetivos de esta revisión sistemática consisten en: a) examinar normativas y recomendaciones establecidas para la construcción y adaptación de instrumentos de evaluación psicológica y b) conocer las tendencias metodológicas dentro del área, identificando no solo recomendaciones sobre los pasos del proceso sino también referentes al reporte adecuado de la metodología, estadísticos a utilizar, tamaños muestrales y toda otra norma relevante.

METODOLOGÍA

Procedimiento

Para este trabajo se siguieron las directrices de la declaración *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (PRISMA) (Page et al., 2021). La estrategia de búsqueda consistió en el rastreo a través de las bases de datos PubMed y el repositorio del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) durante el mayo de 2024. Se combinaron los términos de búsqueda "(Palabras clave: (psicometría OR evaluación OR test OR item OR scale OR psychometry OR Assesment) AND (construcción OR adaptación OR validación OR development OR adaptation OR validation) AND (guía OR recomendaciones OR statement OR criterios OR guideline OR recommendation OR criteria)".

En lo que respecta a los criterios de inclusión, se limitó la búsqueda a: 1) guías prácticas, revisiones bibliográficas y estudios que establecieran recomendaciones para la construcción y adaptación de instrumentos de evaluación psicológica; 2) escritos en español, portugués o inglés; 3) publicados en revistas con referato.

Para determinar la elegibilidad de los artículos, se revisó inicialmente los títulos, resúmenes y palabras claves de todos los estudios identificados. Los casos de discrepancia se resolvieron mediante la reevaluación por parte de tres investigadores pertenecientes al equipo de investigación. Luego, se indagaron las referencias de los artículos seleccionados con el fin de identificar publicaciones que hubieran sido omitidas en el proceso de búsqueda.

Con respecto a la estrategia de extracción de datos, cuando fue posible se exportaron archivos en formato .csv de las búsquedas realizadas, caso contrario se extrajeron manualmente los datos de cada resultado obtenido. Se utilizó una planilla de cálculo para consignar: título, autores, año de publicación, idioma, revista, instrumento de medición y resultados principales de los estudios hallados. Por último, estas distintas bases fueron unificadas en una única planilla donde se procedió a realizar el cribado.

La calidad y el riesgo de sesgo fueron evaluados analizando la metodología de los estudios hallados. Particularmente, se tuvo en cuenta la metodología que se implementó para desarrollar los trabajos publicados, cómo se conformaron los grupos de expertos a cargo de estos (en los casos en que fuera pertinente), y si se explicitó el grado de consenso sobre las distintas recomendaciones emitidas.

RESULTADOS

Producto de la búsqueda inicial, se obtuvieron un total de 2159 artículos. Luego del proceso de cribado se encontraron ocho artículos que cumplieron con los criterios de inclusión establecidos (Figura 1). Se procedió a evaluar las características metodológicas, buscando establecer su riesgo de sesgo.

Figura 1

Diagrama de flujo del proceso de selección de artículos.

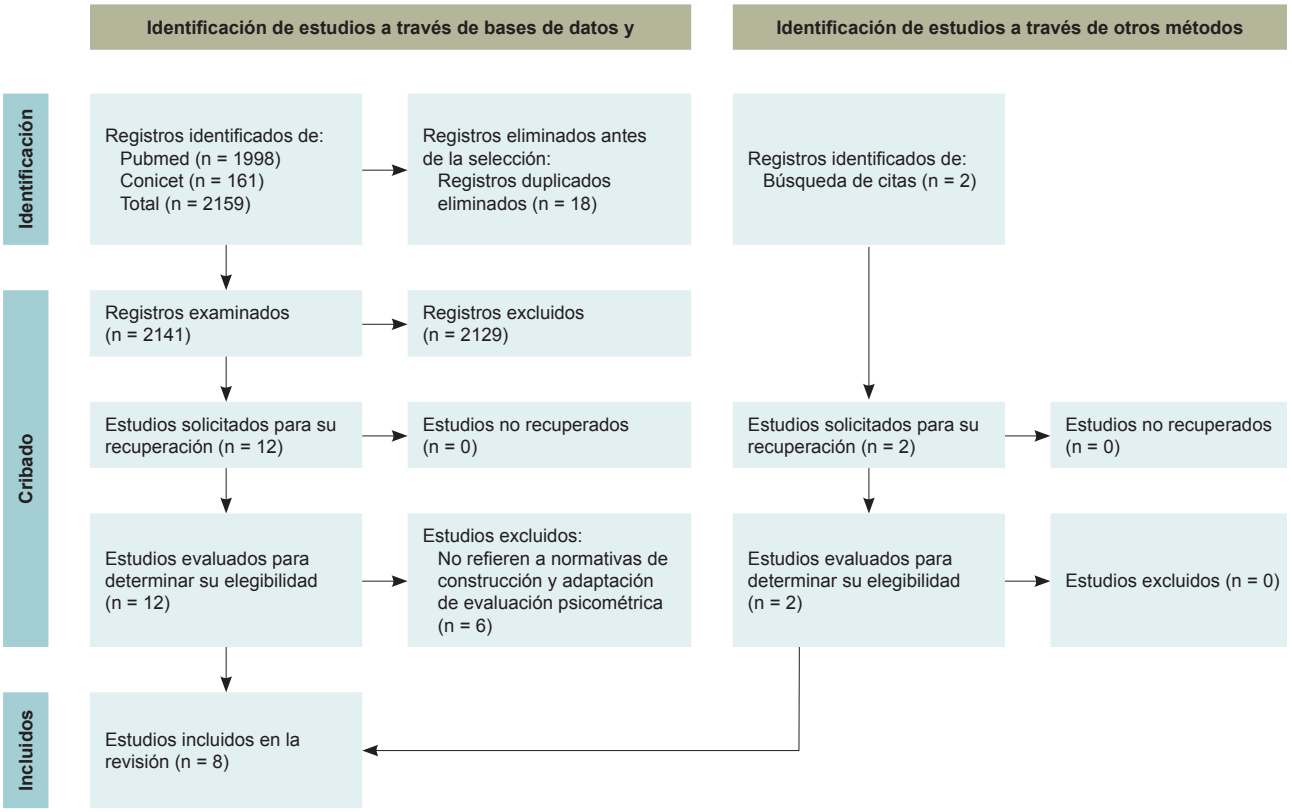

Tabla 1

Tabla de contenido de principales características de los estudios hallados

Autor	Año	Temática	Generales/ Específicos	Características
Caicedo Cavagnis et al.	2018	Entrevistas Cognitivas	Generales	Se establecen criterios estandarizados a cumplir al momento de realizar una entrevista cognitiva durante construcción o adaptación de instrumentos.
Hall et al.	2018	Adaptación de instrumentos relacionados a la audición	Específicos	Realiza una checklist de recomendaciones para adaptar instrumentos de evaluación de tipo autoinforme acerca de cambios en problemas de audición. Enfatizan el hecho de considerar la influencia de diferencias socioculturales.
Iliescu et al.	2024	Adaptación de Instrumentos	Generales	Se desarrolla la checklist TARES, basada en las recomendaciones de la ITC. La misma establece aspectos a tenerse en cuenta en artículos de adaptación.
Kottner et al.	2011	Reporte de confiabilidad y acuerdo	Generales	Proponen la guía GRRAS para el reporte de confiabilidad y acuerdo en el uso de instrumentos o diagnósticos dentro de estudios. Enfatizan la necesidad de reportar la confiabilidad interjueces, tanto dentro de los estudios que validan estas medidas como al momento de aplicarlos dentro de otras investigaciones.
Muniz et al.	2013	Adaptación de instrumentos	Generales	Sugieren recomendaciones para la adaptación de instrumentos, detallándose los distintos pasos a realizar.
Reichenheim et al.	2014	Validez de constructo en instrumentos epidemiológicos	Generales	Realizan recomendaciones para evaluar la validez de constructo en instrumentos epidemiológicos y cómo reportarlos adecuadamente.
Richaud de Minzi	2008	Tendencias en Psicometría	Generales	Establece los procedimientos que predominan dentro de las principales corrientes metodológicas de construcción de instrumentos.
Streiner & Kottner	2014	Reporte de datos en construcción de instrumentos	Generales	Recomendaciones que conciernen al desarrollo de un artículo de construcción de instrumentos, desde la redacción y reporte de datos hasta los estadísticos a realizar y su reporte.

Tabla 2
Reporte de datos de los estudios analizados

Trabajo	Reporte de Datos		
	Fuentes Utilizadas	**Inclusión/Exclusión**	**Grado de acuerdo/certeza**
Caicedo Cavagnis et al. 2018	No se reporta	No se reporta	No se reporta
Hall et al. 2018	No se reporta	No se reporta	No se reporta
Iliescu et al. 2024	Se reporta en forma parcial	No se reporta	Se reporta en forma parcial
Kottner et al. 2011	Se reporta en forma parcial	No se reporta	Se reporta en forma parcial
Muniz et al. 2013	Se reporta en forma parcial	No se reporta	No se reporta
Reichenheim et al. 2014	Se reporta en forma parcial	No se reporta	No se reporta
Richaud de Minzi 2008	No se reporta	No se reporta	No se reporta
Streiner & Kottner 2014	Se reporta en forma parcial	No se reporta	No se reporta

Los ocho artículos hallados pueden ser agrupados para su análisis dependiendo de si buscan realizar recomendaciones amplias, aplicables a múltiples poblaciones y tipos de instrumentos, o si se refieren a una población particular. Bajo esta distinción, como puede observarse en la Tabla 1, ocho artículos abordan temáticas generales (Caicedo Cavagnis et al, 2018; Iliescu et al., 2024; Kottner et al. 2011; Muniz et al., 2013, Reichenheim et al., 2014; Richaud de Minzi, 2008; Streiner & Kottner, 2014) y uno trabaja una población específica (Hall et al., 2018).

En la Tabla 2 se menciona la calidad del reporte de las fuentes consultadas, así como criterios de inclusión o exclusión e información referente al armado de consensos al momento de emitir recomendaciones, o el grado de certeza que se tiene en las mismas en base a las fuentes existentes.

1. RECOMENDACIONES GENERALES

Caicedo Cavagnis et al. (2018)
Este artículo se centra en desarrollar recomendaciones para llevar a cabo entrevistas cognitivas, las cuales buscan identificar y corregir problemas al momento de construir instrumentos de evaluación. El objetivo de estas es entender por qué y cómo una persona responde a los reactivos que se le presentan, proporcionando evidencias de validez basadas en el proceso de respuesta de los evaluados.
Los autores identifican distintas etapas relevantes de la entrevista cognitiva, incluyendo: la planificación, la conducción de la entrevista y el análisis de los datos. Se describen varias técnicas utilizadas en las mismas, como el pensamiento en voz alta, las preguntas de sondeo y las entrevistas retrospectivas.

Iliescu et al. (2024)
En este trabajo, se presentan los Estándares de Reporte de Adaptación de Pruebas - *Test Adaptation Reporting Standars* (TARES). Los mismos fueron desarrollados para mejorar la transparencia y precisión de la documentación de adaptaciones de pruebas. La necesidad de desarrollar los TARES surge de una falta de reportes adecuados, lo que afecta la calidad y utilidad de la investigación.
Los TARES fueron desarrollados por un grupo internacional de expertos bajo la dirección de la International Test Commission (ITC). Los autores reportan haber realizado una revisión bibliográfica, de la cual destacan algunos materiales como las recomendaciones establecidas en la segunda edición de la guía para traducción y adaptación de tests de la ITC y una checklist complementaria a estas guías publicada por Hernández et al. (2020). Se realizaron varias rondas de revisión y retroalimentación para asegurar que los estándares fueran comprensivos y aplicables a diversas disciplinas. Sin embargo, no se menciona detalladamente los criterios de inclusión y exclusión en la búsqueda realizada.
En base a estas recomendaciones, los autores proceden a precisar los distintos elementos de una adaptación y cómo debieran realizarse y reportarse en cada momento. Cada recomendación cuenta con una explicación sobre el razonamiento que la sustenta. También se indica que, si algunas de las recomendaciones no fueran aplicables a la adaptación llevada a cabo, debiera explicarse el por qué. Puntualizan cómo debe redactarse y qué debe contener cada apartado. Hacen especial énfasis en los resultados, los cuales deben contener la confiabilidad, validez y estandarización, así como la comparación de estos con los de la técnica original. Por último, se informa que, al momento de adaptar una técnica, debe reportarse las distintas equivalencias con la técnica original: de constructo, metodológicas y de ítem. También mencionan la importancia de indicar los límites de dichas equivalencias y cómo superarlos de ser necesario.

Kottner (et al. 2011)

En este estudio lo autores proponen la *Guidelines for Reporting Reliability and Agreement Studies* (GRRAS) o Guía para el Reporte de Estudios de Confiabilidad y consenso. Los autores plantean que la relevancia de generar esta guía proviene de la necesidad de proporcionar formas adecuadas de reportar el error inherente en diagnósticos, puntuaciones o mediciones. Destacan que, generalmente, se desconocen en los estudios publicados la fiabilidad y la concordancia entre los usuarios de escalas, instrumentos o clasificaciones son generalmente desconocidas. Se resalta la necesidad de reportar esta información dentro de estudios que implementen cualquier forma de medición.

Para el desarrollo de la guía se contactó a 13 expertos en confiabilidad y consenso, de los cuales solo ocho participaron del desarrollo de la GRRAS. Producto del trabajo de dichos expertos, se realiza una serie de recomendaciones para producir este tipo de estudios. Con respecto a los participantes, se resalta que, dentro de la metodología, no solo se describa la población de interés sino también el tipo de administrador destino del instrumento. Esta población debiera reseñarse según las cualificaciones de los evaluadores, sus antecedentes clínicos, sus conocimientos, su grado de experiencia y su formación, ya que estas características pueden tener un gran impacto en las estimaciones de fiabilidad y concordancia.

A su vez se recomienda reportar estos datos con respecto a los evaluadores que participen del estudio. También debiera describirse detalladamente el proceso de administración y si los evaluadores/administradores participantes trabajaron de forma independiente o conjunta. Los autores también resumen los métodos pertinentes para evaluar la confiabilidad y acuerdo dependiendo del tipo de variable utilizada. Finalmente, destacan la importancia de realizar estudios que midan la confiabilidad y acuerdo entre administradores y entre codificadores.

Muniz et al. (2013)

Este artículo presenta las directrices de la International Test Commission (ITC) para la traducción y adaptación de pruebas psicológicas y educativas. El objetivo de estas es que en el producto final del proceso de adaptación se consigne el máximo nivel de equivalencia lingüística, cultural, conceptual y métrica que sea posible en relación con la versión original.

Para llevar a cabo la revisión se formó un grupo de trabajo multidisciplinar, compuesto por representantes de varias asociaciones de psicólogos. Las directrices fueron concebidas como un patrón para guiar a los investigadores y profesionales. Se buscó abarcar la totalidad de fases y cuestiones a considerar durante el proceso de traducción. Los autores resumen dichas fases en: a) consideraciones legales previas que afectan a la propiedad intelectual; b) valoración del constructo en la población diana; c) diseños de adaptación que tengan en cuenta las características lingüísticas, psicológicas y culturales del texto adaptado, así como su adecuación práctica; d) la importancia de las pruebas piloto; e) la selección cualitativa y cuantitativa adecuada de la muestra de adaptación; f) la importancia de los estudios de equivalencia; g) la delimitación del grado de comparabilidad entre puntuaciones; h) la importancia de unas correctas condiciones de aplicación e interpretación; e i) la información exhaustiva sobre los cambios llevados a cabo en el test adaptado.

Reichenheim et al. (2014)

El trabajo parte de las directrices de la iniciativa *Consensus-based Standards for the selection of Health Measurement Instruments* (COSMIN). Estos criterios engloban a la validez, confiabilidad y sensibilidad de las técnicas. Los autores afirman que, si bien son criterios importantes, también es relevante el estudio de otras propiedades, tales como la validez estructural. No se reporta un criterio de inclusión o exclusión de bibliografía.

Se elabora una guía de siete pasos: 1) corroborar la estructura dimensional, 2) evaluar la fuerza de indicadores componentes relativos con los patrones de las cargas y errores de medición; 3) examinar la redundancia en el contenido a través de la correlación de errores de medición, 4) corroborar la validez factorial convergente y discriminante, 5) evaluar la capacidad de discriminación del ítem e intensidad de indicadores en relación con rasgos latentes; 6) examinar las propiedades de las puntuaciones brutas, 7) evaluar la estructura factorial y la invarianza entre grupos. Si bien los pasos son enumerados, plantean que se trata de un proceso iterativo en la práctica, donde muchas veces, según sea el caso, se debería alternar el orden.

Los autores destacan que al momento de corroborar la estructura factorial, el mejor camino es comenzar con un análisis factorial exploratorio (AFE) cuando se construye una técnica nueva, pero que en el caso de adaptaciones es coherente comenzar con un análisis factorial confirmatorio (AFC). De presentarse demasiadas cargas cruzadas o residuales que volvieran insostenible la estructura factorial se debiera encarar el análisis desde un marco totalmente. Y simultáneamente analizar estos nuevos modelos mediante AFC. También se menciona que existen modelos exploratorios más novedosos, que adecuan los modelos factoriales sin un AFC, como el Modelo de Ecuaciones Estructurales Exploratorio, el cual se clasifica como análisis factorial exploratorio/confirmatorio (AFE/C).

Por otro lado, para evaluar la carga factorial de los ítems y errores de medición, se recomienda reportar las cargas factoriales de cada ítem. Se afirma que, aunque actualmente no hay un punto exacto de corte, sería optimo considerar valores mayores a 0.7 como deseables, y que algunos autores consideran aceptables cargas de hasta 0.3 con otros estableciendo ese punto en 0.5. Además, siempre deberían reportarse la unicidad de los ítems (δi), ya que esta expresa la cantidad de información, varianza, que queda fuera de la estructura factorial. No existe un consenso respecto a cuando considerar la unicidad como alta, pero se sugiere que valores superiores a 0.6 debieran analizarse con cautela e items con residuales de 0.7 o más podrían ser candidatos para ser suprimidos o sustituidos.

Sumado a esto, se remarca que si bien una representación teórica de la estructura factorial del instrumento puede indicar cómo deberían relacionarse los ítems con los factores,

a nivel práctico esto suele implicar conectar conjuntos de indicadores mutuamente excluyentes con factores específicos. Esto conlleva una pérdida de la concentricidad, indicando que un ítem carga en más de un factor. De igual importancia, las cargas cruzadas de ítems tienden a reducir los valores en general, lo que implica una confiabilidad factor-específica del ítem inferior a la deseable. Por ello, examinar las cargas cruzadas también es necesario.
Una posible solución a esto es eliminar los ítems que presentan anomalías, pero esto debe emplearse con precaución, ya que eliminar ítems empíricamente representativos puede remover de la escala el contenido de dicha variable. Al corroborar la validez convergente y discriminante en base a los factores, utilizan el análisis de la varianza promedio extraída. Para evaluar la capacidad de discriminación del ítem e intensidad de indicadores en relación con rasgos latentes, en el caso de ítems categóricos, se toma a la TRI como recomendación que permite relacionar las características de los ítems y sujetos con la probabilidad de presentar una determinada categoría de respuesta. A su vez, en el caso de utilizar puntuaciones directas, los autores enfatizan la necesidad de verificar cómo estos se relacionan con el puntaje del factor y analizar sus propiedades psicométricas.
Finalmente, mencionan que la evaluación de la invarianza puede lograrse mediante un AFC de grupos múltiples, modelos de múltiples indicadores, múltiples causas o modelos TRI.

Richaud de Minzi (2008)
El artículo tiene como objetivo examinar los avances en la psicometría en los últimos años. Para ello, se realiza una revisión narrativa, identificando corrientes teóricas y algunas metodologías de trabajo. Desarrollan la Teoría Clásica de los Tests (TCT), la Teoría de la Generalizabilidad (TG) y la Teoría de Respuesta al Ítem (TRI). No se explicita una metodología específica de búsqueda o evaluación de la evidencia.
El artículo enfatiza particularmente los avances y beneficios que propone la TRI, dado que permite una evaluación más precisa y adaptativa al considerar la dificultad de los ítems y la habilidad del individuo. Además, esta línea teórica ha llevado al desarrollo de tests adaptativos informatizados, que ajustan los ítems presentados al examinado en función de sus respuestas. A su vez, se desarrolló el enfoque de Generalizabilidad, que consta de la implementación en conjunto de TRI y TG haciendo supuestos de distribución acerca de las facetas de medición relevantes. Por otro lado, se resalta los avances en el estudio de la validez de constructo, la importancia de utilizar distintos métodos para realizar los análisis, y la implementación de métodos más sofisticados como los son las ecuaciones estructurales y el análisis factorial confirmatorio.
Por último, la autora hace referencia a los Tests Referidos al Criterio (TRC). Estos instrumentos representan procedimientos para evaluar el rendimiento y/o conducta de los sujetos con relación a dominios de contenidos bien definidos, en vez de por referencia a la conducta de otros sujetos. Se remarca que, además de dominio de conductas, puede hablarse intercambiablemente de objetivos, destrezas y competencias. A su vez se establece la importancia de que el dominio esté bien definido, siendo variables la amplitud y los contenidos de este. La autora destaca que la utilidad de estos tests en contextos educativos y clínicos, donde es crucial determinar si se han alcanzado ciertos estándares.

Streiner & Kottner (2014)
El objetivo de este trabajo es presentar una guía para redactar artículos que aborde la construcción y el desarrollo de instrumentos de evaluación en el ámbito de la salud, y cómo reportar adecuadamente sus características psicométricas. Los autores comienzan enumerando los múltiples usos que tienen estas herramientas en la práctica clínica, y las dificultades para evaluar su calidad debido a reportes inadecuados de sus propiedades psicométricas.
La propuesta consiste en construir un documento breve y sencillo que indique qué secciones deben contener y cómo dichos trabajos deben ser escritos. No se detalla una metodología específica de búsqueda o evaluación de la evidencia, ni se identifican todas las fuentes. Para cada sección se detalla la información relevante a incluir, junto con recomendaciones específicas que se destacarán a continuación.
En principio se plantea la importancia de la claridad y suficiencia del título, especificando si evalúa propiedades psicométricas de una técnica ya existente o de un nuevo instrumento, y si el foco está en la validez o en la confiabilidad. Por su parte, se detalla todo lo que debe incluir el resumen. Se brindan recomendaciones para la elección de palabras clave adecuadas que permitan hallar el trabajo.
Para la introducción, en caso de tratarse de una construcción, se recomienda incluir la justificación de la necesidad de realizarla. El constructo debe ser explicado y descripto. Si la construcción se basa en una delimitación teórica de constructo, debe describirse la teoría. En cambio, si es basada en investigación u observaciones clínicas, debe describirse el proceso realizado e idealmente el proceso de evaluación de inclusión de los ítems.
En la metodología, el instrumento debe ser descripto incluyendo cantidad de ítems, nombre y cantidad de ítems de cada subescala si corresponde, formato de respuesta, cantidad de respuestas alternativas, existencia de ítems inversos, en caso de haber subescalas numéricas su rango, y cómo la escala en su totalidad es puntuada. Para la muestra debe brindarse suficiente información como tipo de población, criterios de selección aplicados, fecha de inicio y final de reclutamiento. El tamaño muestral debe ser debidamente justificado. Es necesario detallar en la sección de procedimientos el proceso de redacción y selección de los ítems.
En la sección de resultados, se presentan diversas recomendaciones de acuerdo con el objetivo que se haya planteado. A nivel general, afirman que los resultados son estimados y deben reportarse como tales, junto con los intervalos de confianza correspondientes. A su vez, enfatizan la importancia de considerar que la validez y la confiabilidad no son propiedades fijas de la escala en sí misma, sino que refieren a los resultados obtenidos con un instrumento de evaluación en determinado grupo de personas en un contexto específico. También brindan re-

comendaciones sobre la selección de estadísticos para el reporte. Se indica que la validez se refiere al grado en que la evidencia acumulada apoya la interpretación propuesta de las puntuaciones de la prueba para los propósitos establecidos, siendo un concepto unitario que no puede dividirse en tipos. Finalmente, se enfatiza la necesidad de reportar los métodos estadísticos y software utilizados.
En la discusión, se recomienda no afirmar que se logró establecer la validez o confiabilidad de un instrumento, por lo anteriormente mencionado. Por otra parte, todos los estudios tienen limitaciones y deben ser adecuadamente reportadas.

2. RECOMENDACIONES PARA POBLACIONES ESPECÍFICAS

Hall et al. (2018)
El objetivo principal del artículo es proponer una guía de buenas prácticas para traducir y adaptar cualquier cuestionario de autorreporte relacionado con problemas de la audición, con el fin de permitir comparaciones entre poblaciones divididas por idioma o cultura. Para realizar estas recomendaciones los autores recurren a su experticia en el campo y artículos influyentes en el área. Se describe una serie de pasos relevantes para asegurar una traducción de alta calidad. Que sea equivalente al cuestionario original y que además tenga en cuenta las diferencias culturales. Se realiza una tabla y una checklist detallando los pasos.
Los autores incluyen ejemplos publicados que ilustran cómo se han implementado y reportado estos pasos en estudios anteriores. A su vez, presentan una lista de verificación de los ítems de reporte preferidos para ayudar a los investigadores a tomar decisiones informadas sobre la realización o la omisión de cualquier ítem, y recomiendan usar esta lista para documentar estas decisiones en cualquier publicación resultante. Por último, se sugieren licencias de acceso abierto para publicar las adaptaciones realizadas, siempre que esto sea posible.

DISCUSIÓN

Al analizar la bibliografía recabada, es posible determinar algunas tendencias dentro de este campo tan específico. En primer lugar, se detecta que predomina la producción de guías que abordan cuestiones amplias en lo referente a la construcción y adaptación de instrumentos, en detrimento del desarrollo de recomendaciones referentes a poblaciones o variables específicas. De los ocho estudios hallados solo uno refiere una población específica, adaptación de cuestionarios de autorreporte sobre problemas de audición (Hall et al. 2018). Además, cabe mencionar que, a pesar de abordar temáticas específicas, las recomendaciones realizadas en este estudio en lo referente a la construcción y adaptación de técnicas difiere escasamente con las mencionadas por los estudios de temáticas generales.
Los estudios que establecen recomendaciones específicas son de gran importancia dadas las particularidades y desafíos que determinadas poblaciones o tipos de variables puedan presentar. Por lo que es de vital importancia que dichas recomendaciones tengan un alto nivel de especificidad, ya que de otra forma el contenido de estas se vuelve redundante con respecto a guías generales y delega a criterio de los distintos autores al momento de tener en cuenta las peculiaridades de estos grupos o constructos. A su vez, se detecta una falta de abordaje de temáticas específicas de relevancia dentro del medio, por ejemplo: recomendaciones particulares para desarrollar instrumentos de evaluación dirigidos a poblaciones con bajos recursos, rurales, transgénero o nivel educativo bajo. Siendo de vital importancia que estas cuestiones sean abordadas en futuros estudios.
Con respecto a la información recabada por los autores, se encontraron distintas recomendaciones sobre la metodología a implementarse. Casi todos los estudios producen distintas sugerencias para la producción de artículos de construcción, adaptación o validación de técnicas. Las mismas abordan cuestiones que van desde la redacción del título, los distintos apartados de un artículo de estas características hasta incluso cuestiones referentes a los derechos de autor. Distintos trabajos optan por abordar aspectos referentes a la adaptación de instrumentos de medición tanto de forma global como en lo que respecta a poblaciones específicas (Hall et al., 2018; Iliescu et al., 2024; Muniz et al., 2013). En este aspecto se encuentra un alto grado de acuerdo entre los distintos autores, donde pareciera destacarse la influencia de la International Test Comission (ITC) al momento de generar estándares en esta área y unificar el criterio entre investigadores.
Por otro lado, en lo que respecta a las recomendaciones referentes a métodos específicos para la medición de propiedades psicométricas, se aprecia una heterogeneidad de temas abordados. Los distintos estudios hallados trabajan temáticas como las distintas teorías de los test (Reichenheim et al., 2014; Richaud de Minzi, 2008), aspectos referentes a la confiabilidad (Iliescu et al., 2024; Kottner et al. 2011; Reichenheim et al., 2014; Streiner & Kottner, 2014), a la validez (Iliescu et al., 2024; Reichenheim et al., 2014; Streiner & Kottner, 2014) y técnicas específicas como las entrevistas cognitivas (Caicedo et al., 2018).
En particular, al momento de abordar la validez se encuentra una aparente diferencia de perspectivas. Ciertos estudios se enfocan en métodos para evaluar tipos específicos de validez como la validez de constructo (Richaud de Minzi, 2008) o validez estructural (Reichenheim et al., 2014). Por otra parte, Streiner & Kottner (2014) postulan que, si bien la validez puede ser estudiada en sus distintos aspectos, este es un concepto unitario que no puede dividirse en tipos. Estas posturas no son necesariamente opuestas, sino más bien complementarias, ya que los autores que abordan métodos específicos no argumentan que estos sean estudiados por separado. Dicho esto, pareciera ser buena práctica el analizar los distintos aspectos de la validez para luego interpretarla en su totalidad.
En lo que respecta a la confiabilidad se encuentra mayor diversidad de opiniones, particularmente en lo referente a la utilización del Alpha de Cronbach. Algunos autores ex-

plicitan que este método no es el más adecuado (Streiner & Kottner, 2014), mientras que otros recomiendan reportarlo en conjunto con otros estadísticos (Iliescu et al, 2024). Esta última postura pareciera ser la adoptada en el medio ante la falta de un criterio unificado al respecto. Por otro lado, Kottner (2011) y Streiner & Kottner (2014) remarcan no solo la importancia de reportar estadísticos de consistencia de la prueba, sino también de considerar detallar la cantidad de administradores que participan de un estudio, sus características y analizar la confiabilidad y acuerdo entre administradores y entre codificadores.

Sumado a estos aportes, debe destacarse que Streiner & Kottner (2014) remarcan la importancia de comprender a la validez y confiabilidad no solo como propiedades fijas de los instrumentos, sino como valores que refieren a los resultados obtenidos con un instrumento de evaluación en determinado grupo de personas en un contexto específico. Por lo que no debiera afirmarse que un instrumento es válido y confiable por fuera de la población y situación en la que ha sido estudiado. Por otro lado, Reichenheim y colaboradores (2014) afirman que las propiedades psicométricas de un instrumento deberían mantenerse estables a través de distintas poblaciones. Esto debiera ser estudiado con mayor especificidad en investigaciones futuras, que permitieran identificar cuál de las perspectivas posee mayor sustento empírico, dado que ambas presentan fuentes que las respaldan.

Además, puede mencionarse que, dentro de los estudios encontrados, se identificaron dos checklist: TARES (Iliescu et al., 2024) y GRRAS (Kottner et al. 2011). Estas herramientas, dirigidas directamente a investigadores y evaluadores, facilitan el desarrollo de estudios de buena calidad y permiten una mejora en el reporte de datos. Iliescu y colaboradores (2024) afirman que en caso de no tener información que reportar en algún ítem se debiera informar que no hay datos disponibles, beneficiando la transparencia y reproductibilidad de los hallazgos.

A su vez, las checklist son elementos de mucha utilidad para los revisores y editores de artículos científicos, facilitando la adopción de criterios comunes al momento de evaluar manuscritos.

Un aspecto a tener en cuenta es la importancia de la traducción de estos materiales a distintos idiomas, con el fin de que puedan ser adoptados por los profesionales de otras regiones con mayor facilidad.

Por otro lado, se debe destacar una alarmante tendencia dentro de los artículos encontrados a no reportar, o hacerlo de forma parcial, la bibliografía recabada y la metodología implementada para el desarrollo de dichos estudios. De los trabajos hallados, tres no reportan una metodología específica de cribado o de la evaluación implementada para realizarlo (Caicedo et al., 2018; Hall et al. 2018; Richaud de Minzi, 2008). Los restantes, cinco, reportan de forma incompleta algunas de las fuentes consultadas, sin detallar los criterios para la inclusión o la exclusión de otros materiales (Ilescu et al., 2024; Kottner et al. 2011; Muniz et al., 2013; Reichenheim et al., 2014; Streiner & Kottner, 2014). Se reportan solo algunas fuentes relevantes o se citan al momento de emitir recomendaciones, pero sin indicar cómo fueron escogidas o si existió algún criterio de selección. Dentro de este grupo, algunos autores reportan la conformación de los grupos expertos encargados de elaborar recomendaciones (Kottner et al. 2011; Iliescu et al., 2024). La falta de reporte adecuado de las fuentes consultadas y cómo incluir o excluir estos datos implica que el riesgo de sesgo sea elevado, siendo necesario que futuros trabajos reporten los criterios de inclusión y exclusión de bibliografía a ser consultada, al igual que un listado detallado de la misma.

En relación con este último tema, sería recomendable la adopción de metodologías de trabajo similares a las implementadas en distintas guías de buenas prácticas clínicas. En ellas se clasifican los distintos niveles de evidencia consultada en base a su robustez y, a su vez, se rotulan las diversas recomendaciones indicando los niveles de certeza. Esto permite tener una mayor comprensión de la solidez de los datos que sustentan las distintas recomendaciones emitidas. A su vez, aumenta la transparencia sobre el hecho de que se tiene distintos niveles de evidencia y consenso entre los expertos sobre las temáticas abordas (Coello et al., 2016; Løvsletten et al., 2024).

LIMITACIONES

Este estudio cuenta con una serie de limitaciones que deben tenerse en cuenta. En primer lugar, esta revisión sistemática no cubre todas las bases de datos existentes. Algunas bases no han sido incluidas en el proceso de revisión dado que las mismas no cuentan con herramientas de extracción de la búsqueda, lo que imposibilita su inclusión en revisiones sistemáticas de gran amplitud. En este aspecto, es imprescindible que las distintas bases de datos incluyan herramientas como la posibilidad de extraer los resultados de las búsquedas realizadas para así poder realizar estudios secundarios. Por otro lado, ciertas bases de datos poseen acceso restringido mediante *paywalls,* lo que impidió el acceso a las mismas. Siendo así, subsiguientes estudios o actualizaciones de esta revisión debieran de considerar su inclusión.

Sumado a esto, otras bases de datos no han sido incluidas dada la inviabilidad de analizar una cantidad más extensa de registros. En pos de superar este inconveniente se sugiere que futuros estudios amplíen la búsqueda aquí realizada ya sea en forma de nuevas revisiones o actualizaciones de esta revisión. De esta manera, se busca que progresivamente se pueda cribar toda la información existente referente a esta temática.

Por último, gran cantidad de los estudios recolectados poseen faltantes en el reporte de su metodología, por lo que el riesgo de sesgo que poseen es elevado. Siendo así debiera de interpretarse la información y recomendaciones que se han podido extraer de ellas con suma cautela. Es de gran importancia que futuros trabajos de este tipo reporten con mucho más detalle sus fuentes y niveles de certeza al momento de emitir recomendaciones.

CONCLUSIONES

Los resultados de este estudio permiten identificar distintas recomendaciones para la producción de artículos referentes a la construcción, adaptación o validación de instrumentos de evaluación psicológica. Contar con un recuento de este tipo de trabajos es de utilidad para observar los consensos del área, la robustez de la evidencia referente a distintas etapas de la evaluación psicológica y áreas en las que se requiere recolectar mayor evidencia. Producto del trabajo realizado se llega a la conclusión de que es necesario un énfasis en el reporte del proceso de selección de datos que son utilizados en este tipo de trabajos, siendo esta la principal falencia de los estudios hallados. Esto genera un elevado riesgo de sesgo en los estudios presentados. Por otro lado, se identifica que pareciera existir un mayor consenso y consistencia en lo referente a las generalidades de la adaptación de instrumentos, existiendo una mayor heterogeneidad de métodos y recomendaciones referentes a la medición de la confiabilidad y validez. Esto indica la necesidad de desarrollar guías o checklist que brinden a los autores mayor claridad al momento de desarrollar instrumentos. En este sentido, es necesario que se realicen este tipo de trabajos en el medio latinoamericano, para así desarrollar guías que tengan en cuenta las particularidades del contexto y las poblaciones locales.

REFERENCIAS

Ahmed, I., & Ishtiaq, S. (2021). Reliability and validity: importance in medical research. *Methods*, *12*(1), 2401-2406. https://doi.org/10.47391/JPMA.06-861

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (Eds.). (2018). *Estándares para Pruebas Educativas y Psicológicas*. American Educational Research Association.

Caicedo Cavagnis, E. E., & Zalazar Jaime, M. F. (2018). Entrevistas cognitivas: revisión, directrices de uso y aplicación en investigaciones psicológicas. *Avaliação Psicológica*, 17(3),362-370. https://doi.org/10.15689/ap.2018.1703.14883.09

Coello, P. A., Arguis Molina, S., Atienza Merino G., Beltrán Calvo, C., Bernabeu Wittel, M., Blas Diez, M. P., Briones Pérez de la Blanca, E., Calle Urra, J. E., Díaz del Campo Fontecha, P., Estrada Sabadell, M. D., Etxeandia Ikobaltzeta, I., Gaminde Inda, I., Gavín Benavent, P., Gracia San Román, J. Segur Caixa, A., Ibargoyen Roteta, N., López-Torres Hidalgo, J., Lorenzo Martínez, S., León I. M.,… del Mar Trujillo Martín, M. (2016) *Guías de Práctica Clínica en el Sistema Nacional de Salud. Manual Metodológico*.: Instituto Aragonés de Ciencias de la Salud (IACS).

Fernández Ballesteros, R. (2013). *Evaluación Psicológica Conceptos, métodos y estudio de casos (2.a ed.)*. Pirámide. ISBN digital: 978-84-368-2870-2.

Gallo, I. G. (2018). Aportes de la psicometría al ejercicio profesional e investigativo en ciencias de la salud. *MedUNAB*, *21*(2), 6-7. https://www.redalyc.org/journal/719/71964815001/html/

Guillemin, F., Bombardier C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol, 46*(12), 1417-1432.

Hall D. A., Zaragoza Domingo S, Hamdache L. Z., Manchaiah V., Thammaiah S., Evans C., & Wong L. L. N. (2018). A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures. *International Journal of Audiology*, 57(3), 161-175. https://doi.org/10.1080/14992027.2017.1393565.

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. Psicothema, 32(3), 390-398. https://doi.org/10.7334/psicothema2019.306

Iliescu, D., Bartram, D., Zeinoun, P., Ziegler, M., Elosua, P., Sireci, S., Geisinger, K. F., Odendaal A., Oliveri, M. E., Twig J., & Camara, W. (2024). The Test Adaptation Reporting Standards (TARES): reporting test adaptations. *International Journal of Testing*, 24(1), 80-102. https://doi.org/10.1080/15305058.2023.2294266

International Test Commission (2014). Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. *International Journal of Testing, 14(*3), 195-217. https://doi.org/10.1080/15305058.2014.918040

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and reliable? Basic and expanded recommendations for psychometric reporting and quality assessment.* https://doi.org/10.31219/osf.io/3htzc

Kottner J., Audigé L., Brorson S., Donner A., Gajewski B.J., Hróbjartsson A, Roberts C., Shoukri M. & Streiner D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1):96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Lira, M. T., & Caballero, E. (2020). Adaptación transcultural de instrumentos de evaluación en salud: Historia y reflexiones del por qué, cómo y cuándo. *Revista Médica Clínica Las Condes*, *31*(1), 85-94. https://doi.org/10.1016/j.rmclc.2019.08.003

Løvsletten, P. O., Wang, X., Pitre, T., Ødegaard, M., Veroniki, A. A., Lunny, C., Tricco A. C., Agoritsas T. & Vandvik, P. O. (2024). A systematic survey of 200 systematic reviews with network meta-analysis (published 2020-2021) reveals that few reviews report structured evidence summaries. *Journal of Clinical Epidemiology*, 173, 111445.

Martínez Corso, R. C., & Villota Burgos, H. H. (2022). La psicometría. *Revista SIGMA, 18*(1), 23-29. https://revistas.udenar.edu.co/index.php/rsigma/article/view/7951

Mikulic, I. M., Crespi, M., Caballero, R., Vizioli, N. A., & Deleersnyder, G. (2022). Medidas de Evaluación de la Inteligencia Emocional en Argentina. Una Revisión Sistemática. *Escritos de Psicología, 15*(2), 159-170. https://dx.doi.org/10.24310/espsiescpsi.v15i2.15127

Muniz, J., Elosua, P., & Hambleton, R. K. (2013). International Test Commission Guidelines for test translation and adaptation. *Psicothema*, 25(2), 151-157. https://doi.org/10.7334/psicothema2013.24

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., Cho, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Welch, V. A., Whiting, P. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, 74(9), 790-799. https://doi.org/10.1016/j.recesp.2021.06.016

Reichenheim, M. E., Hökerberg, Y. H. M., & Moraes, C. L. (2014). Assessing construct structural validity of epidemiological measurement tools: a seven-step roadmap. *Cadernos de Saúde Pública*, 30, 927-939.

Richaud de Minzi, M. C. R. (2008). Nuevas tendencias en psicometría. *Revista Evaluar*, 8(1), 01-19. https://doi.org/10.35670/1667-4545.v8.n1.501

Streiner, D. L., & Kottner, J. (2014). Recommendations for reporting the results of studies of instrument and scale development and testing. *Journal of advanced nursing*, *70*(9), 1970-1979. https://doi.org/10.1111/jan.12402

Fecha de recepción 14 de septiembre de 2024
Fecha de aceptación 31 de octubre de 2024