UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

EVALUACIÓN PSICOLÓGICA DE LA SIMULACIÓN DE ENFERMEDAD MENTAL:

REVISIÓN SISTEMÁTICA DE LA LITERATURA CIENTÍFICA ENTRE 2012 Y 2016

Tesis sometida a la consideración de la Comisión del Programa de Especialidades Médicas para optar al grado y título de Especialidad en Psicología Clínica

MPSC. JUAN CARLOS MIRANDA OROZCO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2017

A mi madre Cecilia y a mi esposa Rebeca
mi motivación y apoyo incondicional durante este proceso.

Agradezco a mi familia, compañeros de residencia Carlos, Leonardo, Marcia, Nelse y Viviana, a mi coordinadora Karen y al equipo docente.

"Esta tesis fue aceptada por la Comisión del Programa de Especialidades Médicas de la Universidad de Costa Rica, como requisito parcial para optar al grado y titulo de Especialidad en Psicología Clínica"

Dra. Karen Quesada Retana

Directora de Tesis y Coordinadora de Posgrado en Psicología Clínica

---

MSc. Alfonso Villalobos Pérez

Lector de tesis

MPsc. Juan Carlos Miranda Orozco

Candidato

TABLA DE CONTENIDOS

RESUMEN

Bajo el enfoque de Evaluación basada en Evidencia (EBE) el presente estudio buscó evaluar la evidencia psicométrica aportada por diferentes instrumentos propuestos para la detección de simulación de enfermedad mental. Se planteó una metodología de Revisión Sistemática de estudios empíricos en seis bases de datos científicas en inglés y español para el último quinquenio. Se revisaron 2634 artículos de los cuales tras la aplicación de los criterios de filtro se obtuvo 17 investigaciones aplicadas sobre el tema, la mayoría de ellas fueron estudios de caso y control, estudios de simulación y de grupos conocidos, se identificaron dieciocho herramientas que aportaron evidencia psicométrica diversa principalmente a nivel de validez predictiva, se discuten los resultados y se plantean conclusiones en torno al tema.

Palabras clave: Simulación; Evaluación psicológica; Propiedades psicométricas; Validez incremental; Utilidad Clínica

ABSTRACT

From an Evidence-Based Assessment (EBA) approach, the present study sought to evaluate the psychometric evidence provided by different instruments proposed for the detection of mental illness malingering. A Systematic Review methodology of empirical studies of the last five years was proposed in six English and Spanish scientific databases. A total of 2634 papers were reviewed, 17 applied studies about the subject were obteined after application of the filter criteria, most of them were case and control studies, simulation studies and known groups compared; eighteen identified test provided diverse psychometric evidence mainly at predictive validity level, the results are discussed and conclusions are drawn around the theme..

Keywords: Malingering; Assessment; Psychometrics; Incremental validity; Clinical Utility

LISTA DE TABLAS

INTRODUCCIÓN

CONCEPTO DE SIMULACIÓN

Rogers (1997) propone que el estilo de respuesta de las personas podría clasificarse de manera general en seis tipos: a) fingimiento, b) defensividad, c) respuesta irrelevante, d) respuesta aleatoria, e) respuesta honesta y f) respuesta hibrida. En este sentido el autoreporte de un sujeto puede estar permeado por la distorsión de los hechos y esto aplica en diferentes contextos entre ellos los de evaluación.

Según Scott (2006) cuando el estilo de respuesta no es honesta la distorsión del discurso puede ser o no deliberada, por ejemplo cuando se da una omisión o confabulación involuntaria de la información o al ocultar, exagerar o inventar información activamente por una motivación intrínseca o extrínseca.

Como forma consciente de distorsión de la autopresentación el fingimiento puede clasificarse de dos formas: la primera es la disimulación que implica minimizar síntomas que sí existen (deseabilidad social) y en contraparte la segunda es la simulación que implica la exageración o fabricación de síntomas que no existen (Scott, 2006).

Respecto a esta última categoría Mora (2013) hace una revisión de los distintas formas en las que se puede clasificar el tipo de simulación. Puede ser a partir de si la persona ha tenido contacto con la enfermedad real o no, si se simula a partir de la fabricación de síntomas o de su exageración, si el sujeto expresa pasiva o activamente síntomas, si el sujeto es más o menos consciente de que está simulando los síntomas y de acuerdo a las finalidades y motivaciones de la simulación.

Lo anterior nos habla de un concepto complejo, que de acuerdo con Inda, Lemos, López y Alonso (2005) tiene diferentes connotaciones en la literatura internacional, entre ellas están:

> …actitudes de encubrimiento (en el inglés británico, descritas como dissimulation o deception), de fingimiento o engaño (en el inglés americano,

faking), o bien de invención consciente y deliberada de un trastorno mental o físico (en inglés, malingering), o de una incapacidad producida por un accidente o enfermedad, que en realidad no fueron causantes de esta, y de la que se deriva alguna ventaja personal. (p. 99).

En el presente documento asumiremos ésta última acepción del término simulación. Teniendo esto claro existen diferentes modelos explicativos de esta actitud, entre ellos el modelo patógeno, el criminalístico y el adaptativo (Rogers, 1997; González, Santamaría y Capilla, 2012):

- Modelo patógeno. En esencia el modelo parte de que la simulación implica un síntoma o condición patológica progresiva de manera que la expresión de dicho síntoma exagerado inicialmente será cada vez menos controlada hasta ser involuntaria, definición más cercana al trastorno facticio dentro de la lógica de los manuales diagnósticos vigentes.
- Modelo criminológico. Por su parte este modelo asume que la simulación responde a una motivación por mentir y por tanto es propia de personalidades antisociales o psicopáticas, se trata pues de personas "malas", en "malas" circunstancias, que hacen cosas "malas".
- Modelo adaptativo. Bajo este modelo se considera la simulación como una decisión costo-beneficio donde se utiliza la enfermedad para obtener recompensas externas y que dependiendo del medio del sujeto esta reacción sería adaptativa de ahí su nombre.

De ahí que algunas características prototípicas de la simulación entendida desde este último modelo serían: a) la simulación es una estrategia de afrontamiento a la adversidad, b) se usa para tener ventajas en contextos desfavorables, c) ante la adversidad se intenta un beneficio externo y d) no existe una mejor forma de lograr lo que se requiere (González, et al., 2012).

La simulación también es incluida en las clasificaciones internacionales. De acuerdo con Chica-Urzola, Escobar-Córdoba y Folino (2005) en la Clasificación

Internacional de las Enfermedades Mentales (CIE-10) se contempla en el apartado "Otros procesos de la CIE-10 frecuentemente asociados con alteraciones mentales y del comportamiento", específicamente en el código Z76.5 de "enfermos fingidos (simuladores conscientes)" lo cual incluye según los autores personas que aparentan enfermedades por motivos obvios.

Por su parte la Asociación Americana de Psiquiatría (APA) propone la siguiente definición de la simulación: "…fabricación consciente o exageración gruesa de síntomas físicos y psicológicos por una meta externa." (Rogers, 1997, p.11). Dicha acción puede estar motivada por la iniciativas externas como evadir el trabajo, obtener una compensación financiera, evadir una persecución por causa penal u obtener drogas entre otras.

Con respecto a las estrategias que podrían ser asumidas por una persona al simular Conroy y Kwartner (2006) mencionan las siguientes que se corresponden con la tipología propuesta por Rogers (1997) para el instrumento SIRS:

- Síntomas raros
- Adherencia indiscriminada a síntomas
- Síntomas obvios en lugar de sutiles
- Síntomas improbables
- Síntomas inusualmente extremos o severos
- Combinación improbable de síntomas
- Estereotipos erróneos
- Inconsistencias entre síntomas reportados y observados

A pesar de su utilidad clínica-fenomenológica esta concepción parte de un concepto taxonómico de la simulación para el cual hay menos evidencia que para la concepción de la simulación como una variables dimensional (Walters, et al., 2008). Para el mismo Rogers (1997) la simulación de psicopatología podría graduarse en tres niveles principales, leve, moderado y severo descritos a continuación:

- Fingimiento leve. En este caso hay evidencia inequívoca de fingimiento pero sobre todo a partir de la exageración, el grado de distorsión aún así es mínimo y juega un papel pequeño en el diagnóstico.
- Fingimiento moderado. En este se pasa de la exageración a la fabricación de síntomas, intenta activamente hacerse ver más afectado de lo que está y esto puede reducirse a pocos síntomas críticos o representar un rango de distorsiones menores.
- Fingimiento severo. En la simulación severa por su parte la fabricación de síntomas es exagerada desde el punto de vista de que la autopresentación es fantástica.

En correspondencia con esta visión dimensional de la simulación González, et al., (2012, p. 143) indican que como resultado de una evaluación de simulación se podría clasificar a la persona según el grado de probabilidad de simulación y mencionan cinco a saber:

- Clara respuesta honesta (menos del 10% de falsos negativos)
- Probable respuesta honesta (menos de un 25% de falsos negativos)
- Indeterminado
- Probable respuesta simulada (menos de un 25% de falsos positivos)
- Clara respuesta simulada (menos de un 10% de falsos positivos)

El tema de probabilidad está asociado al de epidemiología, por ejemplo se ha descubierto que la simulación se da en cualquier rango de edad y es mas común en ciertos contextos como lo son los militares, prisiones, fábricas y contextos judiciales no existiendo mayor consenso entre los diferentes estudios sobre el porcentaje en que se dan estos casos entre otras razones por la dificultad diagnóstica que implica.

Se ha encontrado simulación por ejemplo en 13% de los casos de salas de emergencias, mientras que otros estudios reportan tasas del 10 al 12% en pacientes psiquiátricos a la vez que del 32% de los casos en contextos forenses (Adetungi, et al., 2006).

Por su parte, desde una perspectiva criterial en el Manual Diagnóstico y Estadístico de los Trastornos Mentales - Versión IV- Revisado (DSM-IV-TR) se indica que debería sospecharse de simulación cuando se cumplen los siguientes criterios: "1) se presenta en un contexto medico-legal, 2) hay discrepancia acusada entre el estrés o la alteración explicados por la persona y los datos objetivos de la exploración médica y 3) falta de cooperación durante la valoración diagnóstica e incumplimiento del régimen de tratamiento prescrito y 4) presencia de trastorno antisocial de la personalidad." (González, et al., 2012, p.25).

A nivel de diagnóstico diferencial la simulación debería distinguirse de diferentes condiciones siendo siempre la principal la patología real que se sospecha puede ser simulada. González, et al., (2012) reconocen como las patologías más comunes de diagnóstico diferencial de la simulación las siguientes:

- Fibromialgia
- Esguince cervical y cervicalgia crónica
- Lumbalgia crónica
- Daño cerebral traumático
- Trastornos por ansiedad
- Trastornos afectivos
- Trastornos somatomorfos
- Psicopatía y trastornos de la personalidad
- Trastornos psicóticos
- Trastornos adaptativos

Otras condiciones en la línea de la expresión sintomática irreal que deben considerarse a nivel de diagnóstico diferencial son el trastorno facticio, el síndrome de Ganser y los trastornos somatomorfos y de conversión (Adetungi, et al., 2006).

De acuerdo con los autores la diferencia atribuida a un desorden facticio respecto a la simulación es que en el primero las ganancias de la expresión sintomática no son externas sino internas. En el caso de los trastornos somatomorfos y de conversión la diferencia fundamental radica en que la expresión de síntomas es

más bien involuntaria o inconsciente. Por su parte el síndrome de Ganser se utiliza para describir cualquier conducta que simule específicamente psicosis o demencia.

Mora (2013) compara las diferencias de la simulación con el trastorno facticio y el fingimiento según las principales clasificaciones internacionales (véase tabla 1):

Tabla 1. Diferencias de la simulación con el trastorno facticio y el fingimiento

	Simulación	Trastorno facticio	Fingimiento
Diferencia según el DSM IV	Producción intencionada de síntomas físicos o psicológicos, desproporcionados o falsos motivados por incentivos externos (entre ellos escapar de una condena criminal)	"Producción intencionada o fingimiento" de los síntomas que está motivada por el deseo de asumir un "papel de enfermo"	Fabricación deliberada o exageración de los síntomas psicológicos o físicos sin ninguna suposición acerca de sus metas
Diferencia según el CIE-10	Producción intencional o fingimiento de síntomas o incapacidades somáticas o psicológicas motivadas por incentivos o estrés externo. Entre los motivos más frecuentes está la de eludir acciones de la justicia.	No aparece como “Trastorno Facticio”, se ubica mayormente en el apartado de fingimiento y simulación ya descritos.	Se fingen síntomas de forma repetida y consistente en ausencia de un trastorno, enfermedad o incapacidad mental confirmados. Trastorno caracterizado por buscar el papel de enfermo.

Tomado de: Mora (2013), p. 37 y 38.

DETECCIÓN DE LA SIMULACIÓN

A nivel de evaluación Chica-Urzola, et al. (2005) rescatan la utilidad de la técnica de entrevista para la detección de personas que buscan engañar al profesional en salud mental como es el caso de la simulación y para ello menciona algunas recomendaciones puntuales a tomar en cuenta a la hora de la entrevista para descartar simulación, las cuales se describen a continuación:

- Atenuación de la vergüenza
- Inducción a la fanfarronería
- La exageración
- La amplificación del síntoma
- La negación de lo general
- El supuesto de conducta sospechosa
- Normalización

Así mismo los autores enumeran una serie de características comportamentales, establecidas por el entrevistador y verbales que podrían hablarnos de un sujeto simulador y que se mencionan a continuación (véase tabla 2).

No obstante Masip (2005) argumenta que se debe desconfiar de las llamadas claves para la detección de la simulación no basadas en el contenido, entre ellas claves fisiológicas, comportamentales y verbales, ya que estas no siempre cuentan con suficiente confiabilidad y validez para considerarse predictores de la simulación y podrían por tanto inducir a error al evaluador.

El autor revisa diferentes estudios experimentales en los que analiza la precisión de observadores para determinar la credibilidad de los testimonios de sujetos declarantes que podrían ser mentirosos u honestos a partir de su comportamiento, los diferentes estudios revisados por el autor muestran que en alrededor del 50% de los casos el observador logra acertar la actitud real del declarante lo que no supera al azar como técnica para su detección, dicha tasa de acierto no demostró ser mejor en evaluadores "expertos" en relación a otros "no expertos".

Es por ello que según González, et al., (2012) un principio de la detección de simulación es que entre más inconsistentes sean los resultados del evaluado en diferentes pruebas independientes más probable será este fenómeno. En esta línea existen modelos multidimensionales de la simulación que hacen referencia a que como criterios diagnósticos deben incluirse no solo la presencia de un incentivo externo, sino también evidencia a nivel de evaluaciones neuropsicológicas, de autoinformes, de exámenes físicos/médicos y de una diagnóstico diferencial.

Tabla 2. Características indicadoras de simulación en la entrevista

Comportamentales	Establecidas por el entrevistador	Verbales
Ausencia o disminución del contacto visual	Posibilidad de una ganancia o evitación de sanción	No hay precisión en el testimonio
Menos expresión de brazos y manos	Información contradictoria de familiares y documentación	Aumento de tensión al dar detalles de quien miente
Las palmas de las manos no se encuentran visibles	Presencia de una trastorno de personalidad	Hincapié en la veracidad
Dedos doblados hacia la mano	Alteración en escalas de patología o validez	Finge que no puede recordar
Piernas dobladas, recogidas o cruzadas	Ausencia de deterioro en estudio retrospectivo	Demuestra altos estándares a nivel de principios y valores
Movimientos asimétricos por ejemplo de la voz		Respuestas no espontáneas, sino ensayadas
Distracción		Confesión tácita
Recitación de cuadros clínicos atípicos		Evasión de temas críticos
Poca relación entre entrevista, exploración física y autoreporte		Usar mismos términos y palabras del interlocutor
Solicitar intervenciones riesgosas y difíciles		Términos de jerga técnica
No adherencia al tratamiento		Contradicciones con datos personales e historia de vida
Imagen que busca mostrar signos de enfermedad		Solicita que le repitan o aclaren las preguntas
Negativista ante la confrontación		Admite síntomas absurdos o ajenos al cuadro clínico
Intranquilidad		Ante preguntas de repuesta obvia contesta lo contrario
Niega responsabilidad y expresa defensividad		Ante preguntas exactas da respuestas aproximadas
Tics y gestos de desaprobación asociados a la mentira		Ante acusación se muestra evasivo
Interrumpe la entrevista antes de terminarla		

Adaptado de: Chica-Urzola, et al., (2005)

A nivel evaluativo de acuerdo con González, et al., (2012) un protocolo de evaluación de la simulación debería considerar cuatro aspectos fundamentales que son: a) la determinación de la existencia de un incentivo externo, b) la presencia de

inconsistencias a nivel clínico, c) una evaluación multimétodo-multisistema y d) la convergencia de datos de diferentes fuentes y métodos.

Como puede observarse la información de fuentes colaterales reviste una particular importancia para la corroboración de la simulación de Psicopatología. Conroy y Kwartner (2006, p. 33) indican que fuentes claves de información para la detección de la simulación pueden ser:

- Notas escolares
- Expedientes de salud mental
- Expedientes médicos
- Historial de arrestos
- Historial correccional o disciplinario
- Entrevistas a personas que han tenido contacto con el evaluado

Más allá de las claves conductuales, verbales y de fuentes colaterales ya mencionadas, la detección de la simulación generalmente pasa por la evaluación mediante entrevistas y autoinformes (Inda, et al., 2005). De acuerdo con los autores, algunos de estos instrumentos no son específicos para medir simulación pero cuentan con escalas para ello, entre los más importantes están el Inventario Multifásico de la Personalidad de Minnesota (MMPI), Cuestionario de Dieciséis Factores de Personalidad (16 PF), el Inventario de Evaluación de la Personalidad (PAI) y el Inventario Clínico Multiaxial de Millon (MCMI).

Por su parte existen otros instrumentos especializados para la detección de la simulación como el Test de Simulación (M Test), la Escala de Probabilidad de Simulación (MPS), el Test de Evaluación Forense de Síntomas de Miller (M-FAST), el Inventario Estructurado de Simulación de Síntomas (SIMS) y la Entrevista Estructurada de Síntomas Reportados (SIRS). Así mismo hay otros para la medición de tipos específicos de simulación, entre ellos el Test de Simulación de Problemas de Memoria (TOMM) que es específico para medir simulación de problemas de memoria.

EVALUACIÓN BASADA EN EVIDENCIA

Como dicen Hunsley y Mash (2007) con respecto a las personas que se someten a una evaluación psicológica: "como receptores de servicios psicológicos, estos individuos merecen, por supuesto, nada menos que lo mejor que la ciencia psicológica tiene para ofrecer" (p. 31). No obstante los autores advierten que no siempre la práctica clínica es coherente con la evidencia científica disponible.

Nuestro Código de Ética Profesional es consciente de esta posibilidad y por ejemplo entre sus prohibiciones al profesional en Psicología propone en su art. 13. "E) Crear falsas expectativas que después sea incapaz de satisfacer profesionalmente F) Fomentar que se sobrevalore la eficacia de los servicios que se presta" (CPPCR, 2011, p. 92). Así mismo en su art. 23 advierte que se deberá evitar el uso de métodos, técnicas y/o instrumentos ajenos a la Psicología pero que de usarse "…deberán ser recursos debidamente respaldados, contrastados y cuya efectividad haya sido comprobada" (CPPCR, 2011, p.92).

Y es que la necesidad de guías que vengan a estandarizar la práctica profesional evaluativa ha venido siendo reconocida por diferentes organizaciones a nivel internacional como la American Educational Research Association, American Psychological Association y National Council on Measurement in Education (AERA, APA, NCMUE, 1999), la Comisión Internacional de Tests (ITC, 2000), el Consejo Federal de Psicología de Brasil (CFP, 2008), la Comisión de Tests del Colegio Oficial de Psicólogos de España (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez (2011) e incluso la Organización Internacional para la Estandarización (ISO, 2011).

Dichas organizaciones han desarrollado iniciativas concretas para por un lado determinar la utilidad y calidad científica que deben tener los test para efecto de su construcción y revisión y, por otro lado para regular, estandarizar y optimizar la forma en que los profesionales utilizan dichos instrumentos y toman decisiones a partir de ellos. En Costa Rica el CPPCR ha desarrollado diferentes iniciativas para informar y orientar al agremiado en cuanto a lo que -desde su perspectiva- considera buenas prácticas evaluativas, ejemplos de ello son el documento de recomendaciones de la Fiscalía para la aplicación de pruebas psicológicas (Fiscalía CPPCR, 2011) y más

recientemente el titulado “Pautas para la elaboración de informes psicológicos” (Comité Consultivo CPPCR, 2014).

No obstante pese a que en la actualidad se cuenta con múltiples documentos potencialmente orientadores sobre las buenas prácticas en evaluación psicológica, que la producción científica en Psicometría es abundante a nivel internacional y que con el internet está información es cada vez más accesible a los y las profesionales, no se puede asumir que toda tiene la misma calidad o utilidad científica, ni que por estar accesible pueda ser revisada en su totalidad por los profesionales (Beltrán, 2005; Sánchez-Meca y Botella, 2010).

Por tanto resulta importante contar con mecanismos que permitan filtrar la información en función de encontrar la mejor evidencia disponible y es así como desde el enfoque de la Práctica Basada en Evidencia (PBE) se han desarrollado metodologías de investigación secundaria como lo son las Revisiones Sistemáticas (RS) y los Metanálisis (MA), éstas “constituyen una metodología de investigación que tiene como objetivo acumular de forma sistemática y objetiva las evidencias obtenidas en los estudios empíricos sobre un mismo problema” (Sánchez-Meca y Botella, 2010, p. 8).

Estos autores incluyen la revisión de tamizajes y diagnósticos como parte de los objetos de estudio posibles desde la PBE en los que por ejemplo se puede revisar estudios para evaluar la precisión de test psicológicos en la detección de problemas o temas de interés específicos. Es así como ha surgido la Evaluación Basada en Evidencia (EBE) la cual es definida por Hunsley y Mash (2007) como un acercamiento a la evaluación clínica orientado por la investigación y la teoría con objetivos como: “guiar la selección de constructos a ser evaluados para propósitos específicos de evaluación, los métodos y medidas a ser utilizadas en la evaluación, y la manera en que el proceso de evaluación se desarrolla” (p. 30).

Posteriormente los autores distinguen entre la EBE y los Instrumentos Basados en Evidencia (IBE) partiendo de la diferencia ampliamente discutida por diferentes autores (Meyer, et al., 2001; Urbina, 2007) entre la evaluación psicológica y el *testing*, donde se recalca la mayor complejidad de la primera respecto a la segunda implicando entre otras cosas la integración e interpretación de la información.

El enfoque de IBE en la práctica aplicada se reduce a la posibilidad de contar con criterios fundamentados para elegir métodos o instrumentos de evaluación para propósitos y contextos específicos. Al respecto Weiner (2003) propone contemplar la adecuación psicométrica de los test, su relevancia para responder las preguntas de referencia (utilidad clínica) y la validez incremental que aporta a la evaluación entre otros aspectos.

Con respecto a la adecuación psicométrica Hunsley y Mash (2007) mencionan que se trata de que el instrumento sea estandarizado, tenga normas representativas y un apropiado nivel de validez y confiabilidad para un propósito y población específicos. Existen diferentes propuestas de criterios para determinar la adecuación psicométrica de instrumentos de evaluación, entre ellas se pueden mencionar las de Bickman et al. (1999), la de Hunsley y Mash (2008b) y la de Cohen et al. (2008) todas citadas en Hunsley y Mash (2011), la Carretero-Dios y Pérez (2005) y la de Hernández, Ponsoda, Muñiz, Prieto y Elosua (2016).

Con la validez incremental se busca aclarar si un método o instrumento aporta a la predicción de un criterio evaluado más allá de lo que pueden predecir otras fuentes de datos que responden al mismo criterio (Hunsley y Mash, 2011). Según los autores con esto se contesta a dos preguntas fundamentales: ¿puede un instrumento X medir mejor un constructo que otras medidas alternativas? Y ¿es eficiente obtener datos extra a los recolectados tras medir el constructo con la herramienta X?

Con respecto a la utilidad clínica esta parte de un análisis funcional del instrumento en el que se busca determinar la medida en que el uso de los datos de la evaluación lleva a mejoras en la precisión, resultados o eficacia de los servicios clínicos y el funcionamiento del cliente, entre ellos el cambio asociado a la sensibilidad al tratamiento (Hunsley y Mash, 2007).

EVIDENCIA EN LA EVALUACIÓN DE LA SIMULACIÓN

Desde la EBE Hunsley y Mash (2007) consideran la evaluación de la simulación uno de los tópicos de más alto riesgo por sus implicaciones:

> La detección de la simulación se ha convertido en una tarea importante en muchos ámbitos clínicos y forenses. Los investigadores han examinado la habilidad de medidas especialmente desarrolladas para la simulación y de escalas de validez incluidas en mediciones de amplio espectro para identificar con precisión individuos que parecen fingir malestar clínicamente significativo. (Hunsley y Mash, 2007, p. 44).

En el caso del estudio sobre la evaluación de la simulación se cuenta con múltiples antecedentes investigativos de diseño experimental, entre ellos Ensayos Aleatorios Controlados (EACs) que de acuerdo con Petticrew y Roberts (2006) en la llamada "jerarquía de evidencia" ocupan un lugar privilegiado como evidencia científica en cuanto a su alcance para responder a diferentes preguntas de investigación solo por debajo de las RS. A continuación se mencionarán algunos estudios de este tipo.

Sullivan y Richer (2002) en un estudio con 60 estudiantes universitarios de primer año a los que se ubicó en uno de tres grupos experimentales: simuladores, simuladores advertidos sobre la detección de simulación y controles, se les aplicó la lista de chequeo de síntomas neurológicos, el cuestionario GHQ-30 de salud general y las escalas de depresión, ansiedad y estrés DASS como autoreportes estandarizados de quejas subjetivas, encontrando que dichas escalas fueron vulnerables a la simulación y que la medida de advertir sobre la detección de la simulación en la evaluación si bien mostró una tendencia hacia la reducción del fingimiento no logró ser significativa.

Esto si fue logrado en un estudio con un diseño experimental similar donde King y Sullivan (2009) dividieron en los mismos tres grupos a 67 estudiantes de Psicología de primer año donde a ambos grupos de simuladores se les brindó un incentivo económico por simular con credibilidad y al grupo advertido de la posibilidad de detección se le amenazó con la pérdida de créditos académicos de ser detectados. Les fueron aplicados el test PAI y la Lista de Chequeo de 90 síntomas - Revisada (SCL-90) y encontraron que el grupo de simuladores advertidos obtuvo

puntuaciones significativamente más bajas que el grupo de simuladores sin advertencia de detección y muy cercanas a la del grupo control.

Algunos estudios han estado más orientados a explorar las propiedades psicométricas de instrumentos potencialmente útiles para detectar simulación.

En un estudio referente desde el enfoque de la EBE Bagby, Marshall y Bacchiochi (2005) exploraron la validez incremental y utilidad de la escala de Simulación de Depresión del MMPI-2 respecto a otras escalas del instrumento reconocidas para la detección del fingimiento como la F, la Fb, y la Fp, para lo cual analizaron su capacidad diferencial para discriminar en una muestra de protocolos del MMPI-2 que incluía pacientes deprimidos y profesionales de salud mental entrenados para fingir depresión. Los autores encontraron que si bien la escala de Simulación de Depresión obtuvo diferencias estadísticamente significativas para detectar con una precisión ligeramente mayor a los simuladores esta diferencia no se tradujo en una utilidad clínica relevante respecto a las escalas más tradicionales con las que se comparó al contar estas con una sensibilidad similar a ellas.

En esta misma línea en su estudio de validez predictiva Killgore y DellaPietra (2000) reconocen la utilidad del Índice de Ítems Raramente Perdidos (RMI por sus siglas en inglés) del subtest de memoria lógica a largo plazo por reconocimiento de la Escala de Memoria de Wechsler-III (WMS-III) al encontrar que los 6 ítems que componen la escala lograron una sensibilidad del 97% y una especificidad del 100% al discriminar entre un grupo de 51 pacientes con discapacidad neurológica y 36 voluntarios sanos que simularon lesión cerebral y bajo desempeño cognitivo.

Por su parte en un estudio en el que se aplicó a 519 adultos sin quejas por lesión cerebral los instrumentos: Test de Memoria de Palabras (WMT por sus siglas en inglés), Evaluación Computarizada del Sesgo de Respuesta (CARB por sus siglas en inglés) y el Test de Simulación de Problemas de Memoria (TOMM) Gervais, Rohling, Green y Ford (2004) encontraron que dichos test difieren en su sensibilidad para detectar sesgo de respuesta y esfuerzo subnormal siendo que el test WMT logró una mayor sensibilidad para detectar esfuerzo subnormal y sesgo de respuesta que el TOMM y el CARB en ese orden. Los autores advierten sobre el riesgo de asumir dichas medidas como equivalentes para este objetivo de evaluación y a nivel de utilidad

clínica cuestionan el aporte del TOMM en una batería que contemple los otros dos instrumentos.

Siempre en el tema de simulación de la lesión cerebral Sullivan (2000) realizó un Meta-análisis sobre instrumentos neuropsicológicos útiles para su medición en el que encontró que en el desempeño en test de memoria de reconocimiento los simuladores obtuvieron puntuaciones que se encontraban entre una desviación estándar y desviación estándar y media por debajo de sujetos control y personas con daño cerebral. La autora advierte que si bien test de atención, función ejecutiva y procesos viso-espaciales pueden ser de aporte en la evaluación de la simulación tienen un poder predictivo mayor los test de fingimiento y los de reconocimiento.

FORMULACIÓN DEL PROBLEMA

¿Los casos de presunta simulación de enfermedad mental peritados psicológicamente con distintas estrategias metodológicas e instrumentos cuentan con evidencia psicométrica diferencial para fundamentar la toma de decisiones en relación al caso?

OBJETIVO GENERAL

Evaluar la evidencia psicométrica aportada por diferentes herramientas de evaluación propuestas para la detección de simulación de enfermedad mental en estudios aplicados del periodo 2012-2016 en idioma español e inglés.

OBJETIVOS ESPECÍFICOS

Resumir la evidencia relevante disponible sobre las principales herramientas para la evaluación de simulación propuestas por la literatura científica.

Contrastar entre sí la evidencia psicométrica aportada por los distintos métodos e instrumentos para la evaluación psicológica de la simulación de enfermedad mental.

Determinar la validez incremental con la que cuentan o no las diferentes herramientas revisadas en relación a mediciones alternativas.

Estimar la utilidad clínica de las diferentes herramientas revisadas para los objetivos, poblaciones y contextos para los que se propone la evaluación de la simulación.

MÉTODO

Al tratarse de una Revisión Sistemática, el presente estudio utilizó una metodología de investigación secundaria de estudios observacionales primarios de tipo psicométrico con un análisis cuantitativo básico y cualitativo. En los términos propuestos por Beltrán (2005) se trata de un estudio integrativo, observacional y retrospectivo.

A continuación se presentan los criterios de selección, codificación y análisis de los estudios revisados.

RECOPILACIÓN DE DATOS

Se utilizó una estrategia de búsqueda de la literatura utilizando diferentes fuentes formales secundarias de artículos primarios en inglés y español disponibles para descarga de texto completo. Específicamente las bases de datos que se contemplaron para este estudio fueron las siguientes:

- Base de datos de investigación EBSCOhost
- Biblioteca Científica Electrónica en línea Scielo
- Plataforma de búsqueda ProQuest
- Portal Bibliográfico de la Universidad de la Rioja Dialnet
- Red de Revistas de América Latina y el Caribe, España y Portugal (Redalyc)
- Sitio web de investigación científica y medica ScienceDirect

Sobre la elección de las palabras clave utilizadas para la búsqueda de artículos, ésta se dio a partir de una revisión inicial de la literatura en la que se identificaron estos buscadores como los principales en relación al tema propuesto. Se descartaron otros buscadores cercanos al tema como “fingimiento” porque como se mencionó antes, este constructo tiene una connotación diferente, mientras que otros buscadores como “validez” y “confiabilidad” eran normalmente englobados en otros buscadores

escogidos como “propiedades psicométricas”. Finalmente se utilizaron los siguientes buscadores:

- Simulación / Malingering
- Evaluación / Assessment
- Propiedades psicométricas/Psychometrics
- Validez incremental / Incremental validity
- Utilidad Clínica / Clinical Utility

Estos buscadores se incluyeron en las bases de datos tanto en idioma español como en idioma inglés. Para cada base de datos se asoció el primer buscador “simulación/malingering” con cada una de las cuatro palabras clave restantes.

DEPURACIÓN DE ARTÍCULOS

Para la selección de los diferentes estudios se propusieron los siguientes criterios de inclusión en orden de prioridad:

1. Artículos de acceso libre a texto completo en las bases de datos consideradas.
2. Estudios publicados entre enero 2012 y diciembre 2016 incluyéndolos.
3. Documentos escritos en idioma inglés o español solamente.
4. Que sean estudios empíricos o primarios publicados por el autor original/es.
5. Que sean reportes finales de estudios, no avances de investigación.
6. Que cuenten con título, autor/es, resumen, método, resultados y referencias.
7. Estudios aplicados que aclaren muestra e instrumentos utilizados.
8. Incluir en el título, resumen o palabras clave al menos dos de las palabras clave propuesta para este estudio siendo una de ellas simulación / malingering.
9. Que el tipo de simulación investigada en el estudio sea de enfermedad mental.

Como podrá notarse se restringió la inclusión de otras publicaciones diferentes a los estudios empíricos disponibles en texto completo, como resúmenes, conferencias, libros, otras Revisiones Sistemáticas y Meta-análisis y otros documentos potencialmente valiosos, no obstante esto se hizo con el objetivo de favorecer la

disponibilidad de información relevante y suficiente en el documento para el llenado de los protocolos de registro propuestos (véase anexos 1 y 2).

Para garantizar que los artículos cumplieran con algunos criterios de inclusión como el contar con texto completo descargable, periodo e idioma se aplicaron los filtros automáticos respectivos cuando las bases de datos contaban con dicha opción. Para otros criterios de inclusión como el que fueran estudios primarios, que fueran reportes finales de investigación, que aclararan muestra e instrumentos y que incluyeran las palabras clave de la Revisión Sistemática se procedió a la lectura del título, resumen y palabras clave del artículo.

Para los demás criterios de inclusión (que cuenten con estructura de reporte de investigación y que el tipo de simulación fuera de enfermedad mental) así como cuando los pasos previos no permitían esclarecer el cumplimiento de los otros criterios de inclusión se procedió a una lectura superficial del artículo hasta esclarecer el cumplimiento o no del criterio.

La lectura del resumen del artículo o en caso de ser necesario una lectura superficial del documento permitió determinar si los artículos que cumplían con los criterios de inclusión resultaban contar con información relevante y suficiente para pasar a la etapa de codificación, es así como se filtraron los artículos a los que se les dio una lectura analítica del texto completo y se les aplicó el Protocolo de registro de variables de la publicación para artículos elegidos (véase anexo 1).

PROCESO DE CODIFICACIÓN

Para la codificación de la literatura revisada se optó por seguir la propuesta de Sánchez-Meca y Botella (2010) quienes recomiendan para estudios de Revisión Sistemática y Meta-análisis la inclusión de variables en diferentes categorías que respondan a las necesidades del estudio, en este caso se consideraron las categorías de instrumentos, participantes, metodológicas y extrínsecas. A continuación se mencionan las variables moduladoras propuestas para esta investigación:

- Variables de los instrumentos. Incluye el método de evaluación, el nombre de los instrumentos revisados, normas, consistencia interna, consistencia inter-

evaluador, consistencia test-retest, validez de contenido, validez de constructo, generalización de la validez y utilidad clínica.

- Variables de los participantes. Toma en cuenta la edad de la muestra, porcentaje masculino y femenino y el país de la recolección de datos.
- Variables metodológicas. Tipo de muestra, tamaño muestral, criterio de selección de la muestra, instrumentos utilizados, método de evaluación del instrumento.
- Variables extrínsecas. Nombre del artículo, año de publicación del estudio, nombre de los autores y nombre de la revista.

Para una descripción más detallada de las opciones de codificación específicas de cada variable véase el Protocolo de registro de variables de la publicación para artículos elegidos (anexo 1) que cubre las variables extrínsecas, del contexto, de los participantes y metodológicas; y el Protocolo de registro de variables del método para instrumentos elegidos (véase anexo 2) para las variables de la herramienta de evaluación.

El primer protocolo se aplicó a todos los artículos seleccionados para fase de codificación, mientras que el segundo se aplicó a las herramientas de evaluación de la simulación propuestas en los diferentes artículos seleccionados. Entonces la información para el llenado de dicho protocolo se obtuvo partir de los datos aportados sobre tales herramientas en los diferentes artículos revisados que la mencionaban.

Las variables moduladoras correspondientes al primer protocolo y sus opciones de codificación fueron elegidas por este autor con la premisa de que fueran variables influyentes en la variación de los resultados de los diferentes estudios. Mientras que para la codificación de las variables de la herramienta de evaluación se tradujo y adaptó a las necesidades del tema de revisión la Escala de Valoración de Instrumentos de Evaluación propuesta por Hunsley y Mash (2008b) citada en Hunsley y Mash (2011) (véase anexo 3).

Perestelo-Pérez (2013) recomienda el uso de protocolos de codificación previamente publicados con el objetivo de reducir el sesgo del evaluador en esta fase del estudio, es por ello que se optó por utilizar la propuesta por Hunsley y Mash, no obstante la misma autora reconoce la necesidad de contemplar modificaciones a los protocolos publicados para dar cuenta de las particularidades de cada Revisión Sistemática.

Es así como se mantuvieron como las propuestas por los autores originales todas las variables excepto una: "sensibilidad al tratamiento" incluida en la versión original de la escala y que fue eliminada para el presente protocolo (véase anexo 3) por las características del constructo explorado, ya que la simulación en la práctica profesional no se asocia comúnmente a tratamientos específicos por no tratarse de una enfermedad mental como tal sino una actitud ante la evaluación.

En términos generales cada uno de los criterios explorados por el protocolo de revisión de la herramienta de evaluación puede clasificarse de alguna de las siguientes formas (Hunsley y Mash, 2011):

- Menos que adecuado: el instrumento no cuenta con el nivel mínimo de rigor
- Adecuado: el instrumento cuenta con un mínimo nivel de rigor científico
- Bueno: el instrumento cuenta en general con soporte científico sólido
- Excelente: hay un soporte de evidencia extensiva, de alta calidad
- Inviable: no se ha realizado o publicado aún investigación psicométrica
- No aplicable: las propiedades psicométricas no fueron relevantes para el test

Para conocer los criterios específicos de cada variable para una puntuación de menos que adecuado, adecuado, bueno, excelente, inviable o no aplicable véase anexo 3.

ESTRATEGIA DE ANÁLISIS

Como se mencionó en la descripción de la metodología del presente estudio se trata de una Revisión Sistemática, la cual se diferencia de un Meta-análisis en que la estrategia de análisis utilizada es de corte primordialmente cualitativa o cuantitativa

más básica, excluyendo por ejemplo análisis del tamaño del efecto (Beltrán, 2005; Petticrew y Roberts, 2006; Sánchez-Meca y Botella, 2010).

Una vez registrada la información en los protocolos de registro se procedió a la reducción de datos de cada variable moduladora contemplada en el estudio, para lo cual se procedió a generar estadísticos descriptivos de la mayoría de variables mediante el Paquete Estadístico para Ciencias Sociales (SPPSS) versión 20 para Mac (véase anexo 4).

Una vez generados los estadísticos mencionados se procedió a realizar las correlaciones entre diferentes variables relevantes para el estudio y su respectivo análisis a la luz de la teoría.

RESULTADOS

SOBRE LA RECOPILACIÓN DE DATOS

Como resultado de la búsqueda sistemática en torno al tema se revisaron 2634 referencias en seis bases de datos, que tras la aplicación de los criterios de inclusión del estudio llevaron a la obtención de diecisiete artículos relevantes. En la tabla 3 se puede ver el detalle de los resultados de la revisión para cada base de datos.

Tabla 3. Resumen de resultados de la recopilación de datos y depuración de artículos.

Base de datos	Idioma	Combinación de buscadores*	Resultados obtenidos	Artículos elegibles	Artículos repetidos
Scielo	Inglés	Primera	1	0	0
		Segunda	0	0	0
		Tercera	0	0	0
		Cuarta	0	0	0
	Español	Primera	30	0	0
		Segunda	4	0	0
		Tercera	0	0	0
		Cuarta	0	0	0
ProQuest	Inglés	Primera	493	9	0
		Segunda	16	2	2
		Tercera	28	2	0
		Cuarta	142	5	5
EBSCO	Inglés	Primera	156	6	0
		Segunda	13	4	0
		Tercera	0	0	0
		Cuarta	10	1	0
	Español	Primera	51	4	0
		Segunda	0	0	0
		Tercera	0	0	0
		Cuarta	0	0	0

*Primera: Malingering AND Assessment, segunda: Malingering AND Psychometrics, tercera: Malingering AND Incremental validity, cuarta: Malingering AND Clinical utility.

Continuación…

Redalyc	Inglés	Primera	0	0	0
		Segunda	0	0	0
		Tercera	0	0	0
		Cuarta	0	0	0
	Español	Primera	18	0	0
		Segunda	1	0	0
		Tercera	0	0	0
		Cuarta	0	0	0
Dialnet	Inglés	Primera	14	1	0
		Segunda	0	0	0
		Tercera	0	0	0
		Cuarta	0	0	0
	Español	Primera	373	1	0
		Segunda	6	0	0
		Tercera	0	0	0
		Cuarta	30	0	0
ScienceDirect	Inglés	Primera	379	9	0
		Segunda	8	0	0
		Tercera	15	1	1
		Cuarta	69	0	0
	Español	Primera	584	4	2
		Segunda	8	1	1
		Tercera	14	0	0
		Cuarta	171	3	3

Como puede desprenderse del cuadro anterior la fuente que arrojó más publicaciones asociadas a los buscadores elegidos fue ScienceDirect (1248 en total), seguida de la base de datos ProQuest (679 artículos) a pesar de solo contar con publicaciones en idioma inglés, luego de ellas Dialnet con 423 y EBSCO con 230 publicaciones, mientras que Scielo y Redalyc aportaron 35 y 19 coincidencias respectivamente.

No obstante llama la atención que tras la aplicación de los criterios de filtro del estudio la base de datos más productiva fue EBSCO al identificarse once artículos relevantes para los fines de la revisión, seguida por ProQuest con nueve artículos, ScieceDirect con cinco y Dialnet con dos, valga la aclaración de que varios de estos

artículos estaban presente en más de una base de datos. Ni las búsquedas de Scielo ni las de Redalyc llevaron a la obtención de artículos científicos que cumplieran con los criterios de inclusión del estudio.

Con respecto a los buscadores o palabras clave elegidas para la Revisión Sistemática la primera combinación (Malingering AND Assessment) fue por mucho la más fructífera en cuanto a cantidad de coincidencias de búsqueda ya que de las 2634 referencias encontradas en total para el estudio 2099 fueron producto de esta combinación de buscadores. Seguida de esta combinación estuvo la asociación de los buscadores Malingering AND Clinical utility con 422 resultados de coincidencia, la combinación Malingering AND Incremental validity con 57 y Malingering AND Psychometrics con 56 publicaciones.

Igualmente la combinación que generó más artículos relevantes tras el filtro de los criterios de inclusión fue Malingering AND Assessment con 34 publicaciones, seguida de la combinación con Clinical utility que obtuvo nueve resultados pertinentes a la búsqueda, la asociación Malingering AND Psychometrics con siete y la restante que fue la asociada al buscador Incremental validity logró seis resultados útiles para la RS.

SOBRE LAS VARIABLES EXTRÍNSECAS

Un elemento notorio producto del análisis de las variables extrínsecas de los artículos seleccionados fue que solo tres de ellos estaban publicados en idioma español. Igualmente llama la atención que en tres estudios de la Revisión Sistemática participó como uno de los autores Ramón Arce de la Universidad de Santiago de Compostela, dos de estos en co-autoría con Francisca Fariña de la Universidad de Vigo, convirtiéndose con ello en los autores con más publicaciones consideradas en esta investigación.

Respecto al año de publicación es importante recordar que el presente estudio contempló artículos empíricos publicados en el último quinquenio (periodo 2012-2016). Ocho de los 19 artículos considerados fueron publicados en el año 2013, siendo éste el año del periodo que contó con más publicaciones relevantes para esta

RS, a este le sigue el año 2016 con cuatro publicaciones, el 2012 con tres y los años 2014 y 2015 con dos cada uno.

En cuanto a la revistas en las que fueron publicados los estudios considerados resalta que solo una era en idioma español. En cuanto al nivel deducción en el tema cinco de ellas contaron con dos publicaciones que cumplieron con los criterios de inclusión planteados, estas son: BMC Psychiatry, The European Journal of Psychology Applied to Legal Context, PloS ONE y Clínica y Salud, Archives of Clinical Neuropsychology.

Nueve revistas más contaron con una publicación pertinente para este estudio: Military Medicine, Journal of Criminal Psychology, Alcoholism and Psychiatry Research, International Journal of Clinical and Health Psychology, Applied Neuropsychology: Adult, Brain Injury, International Journal of Law and Psychiatry, Personality and Individual Differences y Psychosocial Intervention.

SOBRE LAS VARIABLES DE LOS PARTICIPANTES

A continuación se describen los diferentes estudios revisados en cuanto a las variables asociadas a los participantes (véase tabla 4).

Primeramente es importante aclarar que no todos los estudios revisados contaron con datos suficientes para obtener el promedio general de edad de las muestras utilizadas, ya que si bien la mayoría de ellas contaban con datos de rango de edad e incluso promedio, éstos datos estaban divididos por submuestra y no para el total de participantes.

Aún así se logro obtener el dato para siete de los diecisiete estudios, siendo el promedio de edad general para ellos de 31,95 años con una desviación estándar de 12,71 años. Resalta en este punto el estudio de Novo, Fariña, y Arce (2013) con población adolescente, todos las demás muestras fueron de población adulta.

En cuanto a la distribución por género también se enfrentó la limitación de los datos publicados, no obstante para la mayoría estudios, doce de los diecisiete, se logró extraer el porcentaje de hombres y mujeres. Al respecto es importante hacer notar que la mayoría de estudios contaron con población de ambos sexos aunque en

proporciones disimiles la mayoría de ellos, solo en las investigaciones de Ahmadi, Lashabi, Afzali, Tavalaie y Mirza (2013) y de Vilariño, Arce y Fariña (2013) el 100% de la muestra era masculina y femenina respectivamente.

Tabla 4. Estudios revisados según variables de los participantes

#	Estudio	Edad promedio	% masculino	Procedencia de la muestra
1	Ahmadi, et al., (2013)	-	100	Irán
2	Blasco y Pallardó (2013)	-	24	España
3	Denning (2014)	51	91	EEUU
4	Fuermaier, et al., (2016)	-	-	Europa
5	Jones (2016)	31	88	EEUU
6	Kopf, Galic y Matesic (2016)	-	0	Croacia
7	Liu, et al., (2013)	-	-	China
8	Novo, et al., (2013)	15	50	España
9	Ortega, Wagenmakers, Lee, Markowitsch, y Piefke (2012)	-	-	Alemania
10	Osuna, López-Martínez, Arce y Vásquez (2015)	-	-	España
11	Peace y Richards (2014)	20	27,5	Canadá
12	Sánchez, Jiménez, Ampudia y Merino (2012)	-	-	España
13	Strutt, Scott, Lozano, Tieu y Perry (2012)	42	90	Latinoamérica
14	Vilar, Aparicio, Gómez y Pérez (2013)	-	38	España
15	Vilariño, et al., (2013)	-	0	España
16	Woods, Wyma, Herron y Yund (2015)	26	52	EEUU
17	Young, Jacobson, Einzig, Gray y Gudjonsson (2016)	39	73	Reino Unido

Con respecto a la procedencia de la muestra llama la atención un 35,3% de los estudios revisados fueron con población española y en total con población europea un 59%, mientras que un 23,5% fue con muestra norteamericana. Merece una mención especial el estudio de Strutt, et al., (2012) que recolectó muestra latinoamericana.

SOBRE LAS VARIABLES METODOLÓGICAS

Como un primer dato metodológico relevante para las diferentes investigaciones revisadas es los tipos diseños metodológico utilizado, al respecto se encontró que los estudios de tipo cuasi-experimental fueron los que imperaron. Llama la atención de que varios de los estudios de este tipo contemplaron experimentos que incluían a la vez estudio de fingimiento y de grupos naturales comparados.

Once investigaciones fueron estudios de caso y controles para un 64,7% de la muestra, mientras que cinco estudios más fueron transversales (29,4%) y solo uno de ellos se trató de un ensayo controlado aleatorio, el de Peace y Richards (2014). Muy asociado a este dato, la mayoría de los estudios que fueron de caso y control utilizaron muestras intencionales, en total de las diecisiete investigaciones catorce contaron con muestras de este tipo (82,4%), mientras que las tres restantes fueron muestras a conveniencia para un 17,6%. (véase tabla 5).

Tabla 5. Estudios revisados según variables de metodológicas

#	Estudio	Tipo de muestra	Tamaño muestral	Selección muestral
1	Ahmadi, et al., (2013)	Militar	120	Intencional
2	Blasco y Pallardó (2013)	Forense	26	Intencional
3	Denning (2014)	Militar	151	Intencional
4	Fuermaier, et al., (2016)	Mixta	399	Intencional
5	Jones (2016)	Militar	462	Intencional
6	Kopf, et al., (2016)	Mixta	94	Intencional
7	Liu, et al., (2013)	Mixta	220	Intencional
8	Novo, et al., (2013)	Estudiantil	110	Conveniencia
9	Ortega, et al., (2012)	Mixta	36	Intencional
10	Osuna, et al., (2015)	Forense	202	Intencional
11	Peace y Richards (2014)	Universitaria	298	Conveniencia
12	Sánchez, et al., (2012)	Mixta	156	Intencional
13	Strutt, et al., (2012)	Clínica	20	Intencional
14	Vilar, et al., (2013)	Mixta	84	Intencional
15	Vilariño, et al., (2013)	Mixta	105	Intencional
16	Woods, et al., (2015)	General	50	Conveniencia
17	Young, et al., (2016)	Clínica	63	Intencional

Como puede apreciarse en la tabla 5 respecto al tamaño de muestra la más pequeña fue de veinte participantes en el estudio de Strutt, et al. (2012) y la más grande fue la de Jones (2016) que incluyó 462 participantes, siendo el tamaño muestral promedio de 152,71 con una desviación estándar de 128,36 casos, en total siete de los estudios tuvieron muestras menores a las cien personas.

Es comprensible que los estudios con tamaños muestrales más amplios generalmente incluían entre sus participantes a universitarios junto a otros grupos. En total 41,2% de los estudios revisados contaron con muestras mixtas de este tipo, mientras que se contó además con tres muestras militares (17,6%), dos de ellas de veteranos, además se contó con dos muestras específicamente clínicas, dos forenses, una de universitarios, otra de estudiantes de secundaria y una muestra general de adultos.

SOBRE LAS VARIABLES DE LOS INSTRUMENTOS

En esta sección es importante aclarar que los instrumentos reportados en la tabla 6 no necesariamente son todos los utilizados en los estudios revisados, sino que solo se mencionan para efectos de esta investigación aquellos instrumentos que reportaron información sobre su utilidad clínica y propiedades psicométricas relevantes para la evaluación de la simulación.

Habiendo aclarado esto la herramienta más utilizada en las investigaciones contempladas es el Test de Simulación de la Memoria (TOMM) con un total de cinco apariciones en los diferentes estudios lo que representa un 29,41% del total de la muestra, seguido de ella está el Inventario Multifásico de la Personalidad MMPI en sus diferentes versiones (MMPI-2, MMPI-2-RF y MMPI-A), todos los demás instrumentos son mencionados en un solo estudio cada uno, para un total de dieciocho instrumentos reportados en relación al tema.

Dichos instrumentos se podrían clasificar en las categorías de Instrumentos de validez de síntomas (cuatro instrumentos), Instrumentos clínicos tradicionales (seis test), Instrumentos neuropsicológicos o cognitivos tradicionales (seis test) y Guías estructuradas (tres herramientas) (véase tabla 6).

Tabla 6. Estudios revisados según variables de los instrumentos

#	Estudio	Instrumentos utilizados	Categoría
1	Ahmadi, et al., (2013)	M-FAST[a]	Validez de Síntomas
2	Blasco y Pallardó (2013)	SIMS[b] / MMPI-2-RF[c]	Validez de Síntomas / Test clínico
3	Denning (2014)	TOMM[d]	Validez de Síntomas
4	Fuermaier, et al., (2016)	EFT[e]	Test neuropsicológico
5	Jones (2016)	EI[f]	Test neuropsicológico
6	Kopf, et al., (2016)	MMPI-2[g]	Test clínico
7	Liu, et al., (2013)	SIRS-2[h] / MMPI-2[g]	Validez de Síntomas / Validez de Síntomas
8	Novo, et al., (2013)	MMPI-A[i]	Test clínico
9	Ortega, et al., (2012)	Técnica bayesiana	Guía estructurada
10	Osuna, et al., (2015)	MMPI-2[g]	Test clínico
11	Peace y Richards (2014)	IES-R[j] / PCL[k] / TSI[l]	Test clínicos
12	Sánchez, et al., (2012)	TOMM[d] / VP A.Rec[m]	Validez de Síntomas / Test neuropsicológico
13	Strutt, et al., (2012)	TOMM[d]	Validez de Síntomas
14	Vilar, et al., (2013)	TAVEC[n] / TOMM[d]	Test neuropsicológico / Validez de Síntomas
15	Vilariño, et al., (2013)	Entrevista Clínico-forense	Guía estructurada
16	Woods, et al., (2015)	C-TMT[o]	Test neuropsicológico
17	Young, et al., (2016)	TOMM[d] / RSPM[p]	Validez de Síntomas

Notas: [a]Miller Forensic Assessment of Symptoms Test, [b]Inventario Estructurado de Simulación de Síntomas, [c]Inventario Multifásico de la Personalidad-2-Formula Restructurada, [d]Test of Memory Malingering, [e]Embedded Figures Test, [f]Effort Index del RBANS, [g]Inventario Multifásico de la Personalidad-2, [h]Structured Interview of Reported Symptoms-2, [i]Inventario Multifásico de la Personalidad-Adolescentes, [j] Impact of Events Scale-Revised, [k]Post-traumatic Checklist, [l]Trauma Symptom Inventory, [m]Recognition of Verbal Paired Associates del WMS-III, [n]Test de Aprendizaje Verbal España Complutense, [o] Trail Making Test Computarizado, [p]Raven's Standard Progressive Matrices.

Los autores de las diferentes investigaciones no ahondaron en la mayoría de los casos en el reporte de propiedades psicométricas de los instrumentos utilizados para las muestras a las que se les aplicó, esto no estaba en el alcance de los estudios en la mayoría de ocasiones.

Es así como solo siete del total de estudios (41,2%) indicaron explícitamente el uso de versiones del test adaptadas a la población nacional de la muestra investigada o bien el contar con normas específicas para ese grupo poblacional. Es por ello que el criterio de adecuación de normas fue imposible de definir para la mayoría de estudios excepto en el caso de los estudios de Liu, et al. (2013) y Vilariño, et al., (2013) quienes reportaron datos que permitieron inferir que las normas utilizadas fueron buena y adecuada respectivamente.

Un caso similar fue el de la variable de confiabilidad para la que la mayoría de estudios no aportaron datos originales, pese a que algunos de ellos sí reportaron datos de confiabilidad, principalmente consistencia interna, del manual del instrumento. Las excepciones a la regla en este caso fueron los mismos estudios que brindaron datos sobre la baremación en el párrafo anterior.

Por su parte Vilariño, et al., (2013) también aportó datos originales sobre la consistencia inter-codificador de la guía de entrevista clínica y forense que revisaron, lamentablemente estos datos no lograron una clasificación superior a la de "menos que adecuados" según el criterio de la escala de valoración de instrumentos de Hunsley y Mash (véase anexo 3). Por su parte el estudio de Woods, et al. (2015) fue el único que reportó datos sobre la consistencia test-retest los cuales fueron valorados como adecuados según el criterio mencionado.

En relación a la validez de contenido no se logró obtener datos de ninguno de los estudios contemplados en esta investigación, vale la pena aclara que este tipo de indicador generalmente se reporta en estudios que reportan la construcción o validación de instrumentos que son diferentes a los revisados acá.

Por su parte la validez de constructo fue la propiedad psicométrica para la que los autores generaron más datos, muy amparado al tipo de estudios propuestos, específicamente los autores se centraron en el reporte de datos sobre validez

predictiva aunque algunos pocos reportaron datos de validez convergente y divergente, en los siguientes apartados se desarrolla el detalle de los hallazgos.

En resumen ocho estudios contaron con una validez de constructo adecuada (47%) según la escala de valoración de este estudio, cuatro estudios más contaron con evidencia buena y excelente para un 23,52% del total de estudios y dos investigaciones aportaron información de validez de constructo clasificada como menos que adecuada, estos son los de Osuna, et al., (2015) y de Strutt et al., (2012).

Sobre el criterio de la generalización de la validez lo más común fue más bien el contar con evidencia de constructo menos que adecuada conformando el 47% de los casos, mientras que cinco se reconocieron como adecuada desde el criterio de la escala de valoración de Husley y Masch. Un estudio contó con evidencia buena (Woods, et al., 2015) y otro excelente (Liu, et al. 2013).

En cuanto a la utilidad clínica seis estudios se ubicaron en el rango de adecuada (35,29%), y otros seis más en el de buena, mientras que tres ocuparon la etiqueta de excelente, estos fueron los estudios de Kopf, et al. (2016), el de Liu, et al. (2013) y el de Sánchez, et al., (2012).

A continuación se resumen los principales hallazgos psicométricos concretos de los instrumentos reportados por los estudios revisados:

SOBRE LOS INSTRUMENTOS DE VALIDEZ DE SÍNTOMAS

En esta sección se incluyen los instrumentos especializados para la detección de simulación excepto por el TOMM, el cual por su especificidad neuropsicológica se incluirá en la siguiente sección junto a los instrumentos neuropsicológicos generales.

Con respecto al SIRS-2 Liu, et al., (2013) refieren que en su estudio las escalas de Síntomas evidentes (BL), Síntomas sutiles (SU), Síntomas indiscriminados (SEL) y Síntomas severos (SEV) lograron identificar a los simuladores respecto a pacientes genuinos más efectivamente que el resto de escalas y subescalas y que en general todas las escalas primarias y subescalas del SIRS-2 lograron discriminar bien entre estudiantes simuladores, estudiantes honestos y pacientes reales excepto las

subescalas de Evaluación Directa de la Honestidad (DA), Síntomas defensivos (DS) y Inconsistencia de síntomas (OS).

Como hallazgo interesante Liu, et al., (2013) encontraron que la propiedad de sensibilidad del SIRS-2 para detectar a los simuladores entre estudiantes entrenados para simular fue menor (60%) que para detectar a simuladores de la muestra forense del estudio (85%) siendo esta última superior incluso a la reportada por el manual del instrumento. Así mismo el instrumento mostró correlaciones significativas y en el sentido esperable pero bajas con respecto a los indicadores de validez de respuesta del MMPI-2 (escalas F, L y K) para la muestra universitaria estudiada.

En relación al M-FAST en el contexto de detección de simulación de Síndrome de Estrés Postraumático en veteranos de guerra Ahmadi, et al. (2013) ratifican el puntaje de 6 propuesto por la autora original del instrumento como el punto de corte más equilibrado para garantizar sensibilidad (92%) y especificidad (87%) psicométricas, atribuyendo el mantenimiento de este punto de corte a la ausencia de diferencias significativas a nivel cultural entre el grupo de referencia original del test y la muestra del estudio.

Como parte del estudio de Blasco y Pallardó (2013) se encontró que el Inventario Estructurado de Simulación de Síntomas (SIMS) logró al igual que el MMPI-2-RF una buena capacidad de discriminación entre dos grupos, uno de pacientes con trastorno mixto ansioso-depresivo y adaptativo respecto a otro con sospecha de simulación en contexto medicolegal. El mejor indicador fue la puntuación total del SIMS que logró una correcta clasificación de los sujetos de ambos grupos en un 92,3% a partir del punto de corte de 16 propuesto por el manual del instrumento. En el estudio se evidencia una alta correlación positiva y estadísticamente significativa entre la puntuación total del SIMS y las escalas de infrecuencia del MMPI-2-RF F-r, Fp-r y Fs, lo que justificaría según los autores su uso conjunto como parte de un protocolo de simulación.

SOBRE LOS TEST NEUROPSICOLÓGICOS

En relación al TOMM Strutt, et al., (2012) encontraron que un 18,8% de pacientes hispanoahablantes con trauma craneoencefálico de su muestra fueron erróneamente clasificados como fingidores a partir del punto de corte de 45 propuesto por el Manual del Instrumento. Los autores atribuyen a un bajo nivel educativo de estos participantes el desempeño diferencial por lo proponen un punto de corte de 34 como el óptimo en poblaciones con mínimo nivel educativo para población hispanohablante con mínimo nivel educativo.

Por su parte Denning (2014) encontró para una muestra de veteranos de guerra sanos que dos medidas derivadas del TOMM como lo son la prueba 1 (TOMM1) y los Errores en los primeros 10 ítems (TOMMe10), asociadas cada una con la respuesta de señalar y nombrar en las primeas 10 láminas del TOMM mantienen una especificidad del 90% y aumentan la sensibilidad para la detección de simulación hasta en un 7% para obtener 84% en la combinación de TOMM1 con la conducta y 89% en la combinación de TOMMe10 con la conducta de señalar y nombrar.

Sánchez, et al. (2012) concluyen a partir de sus datos que la combinación del TOMM-R con un punto de corte ≤ 48 y la subprueba de reconocimiento de pares asociados del WMS-III con un punto de corte de ≤ 23 "exhibe la mayor precisión y provee un tamizaje rápido, válido y confiable para la detección de la simulación" (p. 152). Ambos instrumentos obtuvieron rangos de sensibilidad y especificidad superiores al 91% para discriminar entre sujetos simuladores y normales y al 82% entre simuladores y pacientes con Deterioro Cognitivo Moderado, superando los obtenidos por otras cuatro subpruebas del WMS-III en ambos casos.

Por su parte Young, et al. (2016) clasificaron a 63 personas con alegatos de compensación civil en cuanto a la posibilidad de simulación de problemas de memoria a través del test TOMM, donde 36,5% cumplieron el criterio de simulación según el manual, ubicando solo al 3,2% como probables simuladores según el punto de corte de 17 puntos del test y el otro 33,3% como posibles simuladores según el criterios de 45 ítems, lo que según los autores se corresponde con lo esperable para esta población según estudios previos.

Así mismo se les aplicó el test Raven (RSPM) donde el 6,3% cumplió con criterio de simulación de problemas cognitivos, siendo la correlación entre el ambos test para el criterio de simulación probable de solo .30, lo cual los autores justifican a raíz de que la simulación de problemas de memoria no implica la simulación de problemas cognitivos en general. También se correlacionó la puntuación de estos instrumentos con un diagnóstico psiquiátrico a partir de juicio clínico para 40 de los 63 participantes donde la correlación fue nula con el juicio clínico, lo que para los autores fundamenta la necesidad de utilizar este tipo de mediciones psicométricas comúnmente para este ámbito de evaluación.

Siempre en el ámbito de simulación de problemas de memoria Vilar et al., (2013) exploraron los indicadores de sensibilidad y especificidad del Test de Aprendizaje Verbal España Complutense (TAVEC) a través de la comparación de grupos con traumatismo craneoencefálico leve y universitarios simuladores, obteniendo una especificidad entre el 90,9% y el 97,7% para los 5 indicadores del test, mientras que una sensibilidad que rondó entre el 60% y el 40% en los diferentes indicadores, lo que reproduce los hallazgos de estudios previos.

Fuermaier, et al., (2016) revisaron la utilidad del Test de Figuras Incrustadas (EFT) para la detección de simuladores adultos de Trastorno de Déficit Atencional con Hiperactividad (TDAH) encontrando que "los pacientes con TDAH tuvieron tiempos de respuesta más largos y menos errores comparados con participantes instruidos para fingir TDAH" (p. 14-15) e incluso . Así mismo lograron obtener indicadores favorables de sensibilidad y especificidad (88.1% y 90.2% respectivamente) para el punto de corte de -0.25 en el EFT, siendo la capacidad del test para discriminar entre pacientes reales y simuladores tan buena que ni siquiera simuladores expertos en el tema lograron brindar perfiles de respuesta similares a los de pacientes reales.

Sobre la medida de validez llamada Índice de Esfuerzo (EI) de la Batería Replicable de Evaluación del Status Neuropsicológico (RBANS) Jones (2016) examinó diferentes puntos de corte para diferenciar entre grupos con algún grado de sospecha de simulación de traumatismo cerebral en contexto militar identificados a través de otras herramientas de validez. El autor encontró indicadores de especificidad de 97% para los puntos de corte ≥ 1 y 2 y del 100% para el umbral ≥ 3 para todos los grupos

de simuladores, mientras que de 89% para los mismos puntos de corte de 1 y 2 y de 80% para los de 3, esto específicamente para el grupo de simuladores más evidentes, siendo mucho más baja para el resto de grupos.

Woods, et al. (2015) desde un diseño experimental exploraron el fingimiento en la versión computarizada del Trail Making Test (TMT) donde comparó a un grupo de 50 voluntarios fingidores consigo mismos en una aplicación previa y con un grupo control con tiempos demorados. En este estudio se encontró que el tiempo de ejecución de la primera parte de la prueba (C-TMT-A) aumentó en 58% respecto a una primera aplicación que se realizó al grupo y que a través de un punto de corte específico se logró identificar al 50% de los fingidores y al 82% de participantes del grupo control, mientras que en la segunda parte del test (C-TMT-B) las anormalidades en la tendencia del grupo de simuladores permitió clasificar a 83% de los fingidores y al 100% de los controles a partir de un punto de corte específico. Esta tendencia marcada hacia un aumento del tiempo requerido en la primera parte en sujetos simuladores está documentada en investigaciones previas.

SOBRE LOS TEST CLÍNICOS

En cuanto a los instrumentos de evaluación de estrés postraumático en su estudio con estudiantes universitarios simuladores de abuso infantil en contexto de reclamo penal y civil con incentivos positivos, negativos y sin incentivos Peace y Richards (2014) plantean que las escalas total y de evitación del Impact of Events Scale-Revised (IES-R) obtuvieron puntuaciones estadísticamente significativas más elevadas en el rol de contexto penal que en el civil.

Por su parte en el Trauma Symptom Inventory (TSI) se encontró una tendencia menos defensiva en las escalas de validez del test para el contexto civil en el grupo de estudiantes simuladores sin incentivo, mientras que lo contrario sucedió en el grupo con un incentivo positivo. Por su parte en cuanto a las escalas clínicas el grupo de incentivos negativos se asoció a puntuaciones más bajas en varias de las escalas.

Respecto al Post-traumatic Checklist (PCL) se visualizó una tendencia a elevar todas sus escalas en la condición de incentivo positivo, no obstante no se encontró que dichas evidencias fueran significativas.

Sobre el MMPI-2 Kopf, et al., (2016) evidenció la validez incremental de las escalas de infrecuencia F, Fb, Fp y el índice F-K sobre el resto de escalas del instrumento para discriminar entre mujeres con trastorno mixto ansioso-depresivo y estudiantes universitarias simuladoras de dicho trastorno, logrando obtener niveles de sensibilidad y especificidad de 97,9% en ambos casos usando el conjunto de escalas. Por su parte el aporte para la diferenciación de ambos grupos de ocho de las diez escalas clínicas básicas (Hs, D, Hy, Pd, Pa, Pt, Ma y Si) y todas las escalas de contenido, principalmente LSE, SOD y TRT, fue también alto lográndose una sensibilidad del 91,5% y una especificidad del 97,9%.

Siempre sobre el mismo instrumento Osuna, et al., (2015) encontró datos contrastantes sobre la utilidad del MMPI-2 para la detección de la simulación ya que a través de las diferentes escalas de validez del instrumento no logró obtener datos adecuados de sensibilidad para la detección de población clínica forense, identificando como falsos positivos de simulación a pacientes con diagnósticos de enfermedad mental establecidos. Diferente fue el caso de las escalas clínicas básicas de Paranoia (Pa) y Esquizofrenia (Es), las cuales contaron con validez convergente respecto al diagnóstico previo de la muestra, validando así los diagnósticos principales del grupo.

En un estudio similar al de Kopf, et al., (2016), Blasco y Pallardó (2013) encontraron que las escalas de infrecuencia (Fs, FBS-r, F-r y Fp-r) y algunas clínicas y psicopatológicas (RC1, HPC, NUC, COG) de la versión reestructurada del MMPI-2 (MMPI-2-RF) obtuvieron diferencias significativas para un grupo de pacientes con trastorno mixto ansioso-depresivo y adaptativo respecto a otro con sospecha de simulación en contexto medicolegal, siendo el mejor indicador de todo el test la escala Fs con una precisión del 92,3% en la clasificación del total de casos.

En su investigación Novo et al., (2013) lograron establecer indicadores de sensibilidad y especificidad del MMPI-A para el ámbito de daño psicológico por acoso

escolar, mediante una metodología en la que aplicó la herramienta a un grupo amplio de estudiantes honestos y posteriormente con la instrucción de simulación.

Los autores encontraron que en comparación con la aplicación que tuvo la respuesta honesta la respuesta de simulación obtuvo puntuaciones en las escalas básicas en rango de significancia clínica con una probabilidad superior al 50% e incluso del 90% en algunas escalas específicas, así mismo se encontró que las escalas de infrecuencia (F, F1, F2) también fueron sensibles a la instrucción, así por ejemplo F1 con el punto de corte de 90 logró identificar al 66,4% de los protocolos simulados, F2 el 55,1% sin ningún falso negativo, al igual que F con un 69,2% de sensibilidad y una especificidad perfecta, mientras que F-K > 12 logró sensibilidad del 81,3%, así mismo con la escala K con punto de corte de < 40 se logró una sensibilidad de 57,9%.

Los autores advierten que si bien el instrumento encontró rangos de sensibilidad similares a los obtenidos en otros estudios del MMPI, el porcentaje de falsos negativos encontrados para las escalas del instrumento no permiten su utilidad para su uso en ámbito forense como evidencia única en casos de daño psíquico por acoso escolar.

SOBRE LAS GUÍAS ESTRUCTURADAS

Si bien tanto el SIR-2 como el M-FAST son guías estructuradas los hallazgos encontrados sobre estas herramientas se reportan en la sección de test de validez de síntomas, ya que tradicionalmente han sido clasificados en dicha categoría en la literatura científica.

Arce y Fariña (2001) citados por Vilariño, et al., (2013) proponen una entrevista clínico-forense para la valoración del daño psicológico y su diagnóstico diferencial de simulación, la cual Vilariño et al., (2013) contrastaron con los resultados del MMPI-2 en una muestra de mujeres víctimas de violencia de género y otro de simuladoras. La entrevista propuesta obtuvo un coeficiente Kappa de acuerdo inter-codificador de .61 y logró demostrar una adecuada consistencia interna con un alpha de Cronbach de .74 para el grupo de simulación, mostró además una validez convergente moderada entre .48 y .57 con las escalas de simulación del MMPI-2 (F,

Fb, Fp, F-K, S-O, FBS y Ds), demostrando una buena sensibilidad aunque una menor especificidad de alrededor del 50%.

Por su parte Ortega et al., (2012) encontraron a través de dos experimentos que una novedosa técnica de método bayesiano (un cálculo de la probabilidad de un diagnóstico X a través de un test a la luz de la observación directa del caso) logró diferenciar correctamente entre grupos de sujetos simuladores respecto a sujetos normales y con traumatismo cerebral, frente a un test tradicional de validez de síntomas (SVT) con el cual se lograron adecuados niveles de especificidad pero no de sensibilidad. A partir de estos datos los autores rescatan la utilidad potencial de esta metodología para la detección de simuladores en el contexto de lesión cerebral debido a que resulta ser un método simple e intuitivo que logra niveles de precisión adecuados de la decisión clínica.

DISCUSIÓN

Existe una cantidad importante de producción científica alrededor del tema de simulación de enfermedad mental, no obstante la revisión de la literatura lleva al reconocimiento de que mucha de la que aporta información realmente relevante y teóricamente bien orientada fue producida desde la década de los 90`s, ejemplo de ello es la cantidad de veces que aún se cita el texto: Clinical Assessment of Malingering and Deception (Rogers, 1997) en la literatura más reciente así como la cantidad de citas de dicha década en el apartado de referencias de los artículos revisados.

En esta misma línea sobre las características metodológicas de los estudios contemplados es notable la influencia de los diseños cuasi-experimentales clásicos de décadas anteriores que son retomados en este periodo, incluso con algunas herramientas igualmente antiguas en sus versiones más actuales pero con poblaciones diferentes. La clásica advertencia de Rogers sobre las ventajas y desventajas de los diseños de simulación y grupos naturales sobre la validez interna y externa parece ser orientadora en la producción investigativa actual en este campo llevando a un grupo importante de los estudios a combinar ambos experimentos.

Sobre los tipos de instrumentos propuestos para la evaluación de la simulación se corrobora la relevancia de las principales herramientas de validez de síntomas como lo son el SIRS-2, el M-FAST y el TOMM, así como las bondades de sus propiedades de validez predictiva generalmente superiores a las de otros instrumentos tradicionales, una excepción a la regla parecer ser las escalas de infrecuencia del MMPI-2 y su versión reestructurada, que han demostrado buenos indicadores de sensibilidad y especificidad, un caso diferente es el del MMPI-A a nivel de especificidad.

Como se mencionará más adelante, una tendencia sobre todo en el ámbito neuropsicológico aunque Peace y Richards (2014) dan un ejemplo para el ámbito de la evaluación del trauma, es el uso de test no especializados para la detección de la simulación. La evidencia parece respaldar la utilidad clínica en términos de eficiencia

de algunas subpruebas de baterías más amplias como el WMS-III y el RBANS, convirtiéndose en un campo amplio por explorar.

En cuanto a las propiedades psicométricas reportadas para los instrumentos de los estudios revisados debe reconocerse que fueron limitadas, al tratarse la mayoría de investigaciones contempladas de diseño cuasi-experimental y muy posiblemente por el tamaño de muestra reducido que caracterizó a varios de dichos artículos no se reportaron propiedades psicométricas más allá de la validez predictiva, excepciones a esta regla fueron los estudios de Liu, et al., (2013) y el de Vilarino et al., (2013) que hicieron referencia indicadores de confiabilidad de sus guías estructuradas, siendo los únicos indicadores psicométricos originales reportados en toda la revisión sobre la confiabilidad de los instrumentos.

Respecto a la generalización de la validez que versa sobre la posibilidad de extrapolación de los resultados para otras poblaciones y/o en otros contextos esta fue limitada, a siete de los estudios se les asignó una etiqueta de menos que adecuada en este criterio de calidad psicométrica. Esta limitación está muy asociada al tamaño de las muestras y la existencia de grupos normativos del test cercanos a la población.

En cuanto a las normas de los instrumentos revisados en esta investigación es importante aclarar si bien solo en una minoría de los estudios se reportó si se utilizaban normas cercanas o adaptaciones del test para la población estudiada, lo cual es una buena práctica de consenso (AERA, APA, NCME, 1999), esto pasaba a un segundo plano en varios de ellos, pues era parte de los objetivos la elaboración de puntos de corte propios de simulación a partir del contraste de la tendencia de respuesta de muestras comparadas.

Resulta importante no obstante advertir que la mayoría de estos estudios no contaban con tamaños de muestras suficientemente amplios, incluso varios eran menores a 100 sujetos y no eran representativos como lo fue el caso de las muestras militares y de un solo género. Desde un punto de vista técnico esto no permite la generalización de dichos puntos de corte para poblaciones más amplias u otros contextos evaluativos (Carretero-Dios y Pérez, 2005; Muñiz, et al., 2011). El reconocimiento de esta limitación es claro en algunas de las publicaciones revisadas

(Jones, 2016; Kopf, 2016) mientras que en otras esto no aparece explícitamente planteado en el documento y queda en manos del lector concluirlo.

Otra limitación para la generalización de los resultados de estos estudios fue la popularidad de los llamados diseños de simulación entre los artículos revisados (Ortega, et al., 2012; Liu, et al. 2013; Peace y Richards, 2014). Como lo advierte Liu, et al. (2013) este tipo de diseño favorece la validez interna por su control experimental más fuerte pero muestra una reducida validez externa al no permitir la generalización de los datos por darse en condiciones artificiales y con instrucción entre otras cosas.

Los indicadores de validez de constructo incluyeron datos de sensibilidad y especificidad y en menor grado de validez convergente y divergente entre instrumentos. Estos apenas lograron una clasificación de adecuada en ocho de los diecisiete estudios, donde el criterio era simplemente aportar alguna evidencia independiente de validez de constructo. Otros casos menos favorables fueron los de estudios como el de Strutt, et al., (2012) y Osuna, et al., (2015) quienes pese a brindar datos de validez predictiva estos no fueron favorables de acuerdo a los estándares internacionales para estos indicadores.

Como ya se mencionó el principal indicador psicométrico generado en los estudios contemplados para esta investigación fueron los de sensibilidad y especificidad así como sus variaciones de poder predictivo positivo y negativo. Autores como Sánchez, et al., (2012) reconocen a tales indicadores de validez predictiva como claves en los test diagnósticos debido a la relevancia que reviste la capacidad de un instrumento para clasificar individuos de forma precisa en diferentes categorías relevantes clínicamente.

Y es que al ser la exageración la principal estrategia de quien simula contar con evidencia del punto de corte óptimo para poder confirmar o descartar por probabilidad su condición o no de simulador es crítico para la toma de decisiones fundamentadas. Por ejemplo así lo reconocen Novo et al. (2013) para el campo forense:

> En el campo forense, el criterio de decisión ha de ser estricto: los falsos negativos no son admisibles, esto es, el forense no puede informar que un

protocolo simulado es honesto por las implicaciones que tiene para la condena del encausado (e.g., in dubio pro- reo, principio de duda razonable). (p.39).

Con respecto a la utilidad clínica reportada por los estudios esta se mencionó muy subjetivamente en la mayoría de los casos y en otros simplemente debió inferirse ya que como se ha mencionado la investigación psicométrica aún no tiene como estándar un enfoque de Evaluación Basada en Evidencia que permita dar cuenta más sistemáticamente de las variables asociadas a la utilidad práctica de los instrumentos y sea un aporte a la toma de decisiones clínicas (Hunsley y Mash, 2007).

Dentro de esta línea diferentes estudios buscaron dar cuenta de la capacidad de sus instrumentos para diferenciar entre varios tipos de simuladores según su grado de entrenamiento o instrucción para simular, sus motivaciones hacia la simulación e incluso el grado en el que se simulaba (Fuermaier, et al., 2016; Jones, 2016). Los resultados encontrados no parecen ser concluyentes en la mayoría de ellos, pues generalmente no se encontraban diferencias significativas en los promedios de respuesta de los diferentes grupos de simuladores o bien si se encontraban no resultaban ser tan amplias o estadísticamente significativas.

Por otro lado, fue común encontrar diferencias amplias y estadísticamente significativas entre los perfiles de respuesta de pacientes reales y simuladores en varios de los estudios. Como lo menciona Vilar, et al., (2013) esto parece ser "una tendencia entre los simuladores a obtener puntuaciones inferiores a los pacientes en las variables neuropsicológicas" (p. 173).

Siempre relacionado con la utilidad clínica en cuanto a la estrategia de evaluación algunos estudios (Sánchez, et al. 2012; Blasco y Pallardó, 2013; Vilar et al., 2013; Denning, 2014) recomiendan la combinación de diferentes instrumentos, incluso tradicionales o no especializados en simulación, para lograr un mayor poder predictivo de la simulación, lo cual parece una tendencia más marcada aún en el ámbito neuropsicológico. Al respecto Schutte y Axelrod (2013) citados por Denning (2014) reconocen las ventajas de utilizar combinaciones de test como lo son un más eficiente uso del tiempo evaluativo, el no requerir la inclusión de más instrumentos

para dar cuenta de la validez y el menor riesgo de manipulación de la evaluación al combinarse diferentes indicadores de diferentes test.

Weiner (2003) habla de la relevancia de la validez incremental como criterio para la elección de una batería de pruebas, así mismo menciona las propiedades complementarias, aditivas y confirmatorias que distintos instrumentos deben cumplir respecto a los demás para aportar a dicha validez incremental. No obstante la existencia de correlaciones moderadas o altas o bien niveles similares de especificidad y sensibilidad no justifican por sí mismos un aporte a este criterio, sino a la redundancia en la medición, esta es una consideración a la luz de la cual deberían revisarse las recomendaciones de los autores mencionados.

IMPLICACIONES PARA LA PRÁCTICA PROFESIONAL

Como Peace y Richards (2014) mencionan las motivaciones e incentivos para fingir una condición o enfermedad mental en diferentes contextos clínicos y forenses pueden ser múltiples y cada una tiene el potencial de modificar de manera diferente la tendencia de respuesta de los sujetos a las pruebas psicométricas. Es importante para el profesional el conocer de dichas motivaciones o incentivos para implementar prácticas evaluativas que permitan controlar o al menos cuantificar su impacto en el autoreporte del evaluado.

No ajeno a ello diferentes autores (Ortega, et al. 2012; Liu, et al., 2013) advierten la mayor utilidad clínica de estudios basados en diseños como los de comparación de grupos conocidos frente a otros como los de diseños de simulación que como se mencionó más arriba son los más populares en este campo. Según Ortega et al., 2012 los primeros cuentan con mayor validez externa y con ello se aumenta la probabilidad de generalización de los resultados a otros contextos de evaluación en la vida real, por lo que orientar la elección de instrumentos en la práctica clínica que sean respaldados por investigación que denote una buena validez externa es una recomendación bastante razonable.

Otra buena práctica evaluativa en el campo de la simulación en la que diferentes autores de los revisados redundan es el apoyarse en múltiples métodos y

técnicas de evaluación para un mismo caso y con ello trascender el sesgo inherente a herramientas y metodologías específicas (González, et al., 2012; Sánchez, et al., 2012; Osuna, et al. 2015; Vilar, et al., 2013; Vilariño, et al. 2013).

Esto reviste más relevancia aún en contextos evaluativos donde la toma de decisiones respecto al caso exigen niveles de certeza y precisión mayores por la importancia de las mismas y la mayor exposición profesional y técnica al cuestionamiento de su proceder como lo es el ámbito forense. El contar con modelos de recolección de datos y de toma de decisiones desde un enfoque multimétodo-multifuente debe ser una aspiración del profesional en su práctica clínica, un ejemplo de ello es la propuesta de Slick, et al., (1999) citados por Strutt et al., (2012) para la valoración del llamado síndrome neurocognitivo fingido.

Es importante además el contar con herramientas sensibles a la realidad de la población meta, esto implica la adaptación y baremación de los instrumentos para la detección de la simulación que utilizan nuestros profesionales. Por ejmplo Strutt, et al. (2012) advierten sobre la posibilidad de que niveles educativos bajos podría reducir el desempeño de las personas en pruebas como el TOMM haciendo creer que la persona es simuladora cuando en realidad no lo es, es decir un falso positivo.

Es así como el considerar los indicadores de validez predictiva en las pruebas utilizadas para la evaluación de la simulación resulta un criterio más a considerar para selección de instrumentos de evaluación y como se ha visto en este documento, esto no necesariamente implica la adquisición de pruebas especializadas de validez de síntomas.

LIMITACIONES DEL ESTUDIO

Con respecto a la elección de los buscadores o palabras clave para este estudio se siguió la advertencia de Inda, et al., (2005) sobre la diferencia entre las acepciones en idioma inglés del concepto de simulación “malingering” y “faking” por lo que se optó en esta Revisión Sistemática por la elección de la primera por su mayor precisión conceptual para los fines del estudio. No obstante ya en la labor de recolección de datos se corroboró la alta popularidad del segundo buscador y sus derivaciones así

como su uso a modo de sinónimo del término malingering, lo que siembra la sana inquietud de contemplar este buscador para futuras revisiones sobre el tema.

Sobre los criterios de inclusión del estudio es importante reconocer que no se consideró la llamada literatura "gris" u "opaca", lo que excluyó información potencialmente relevante para el tema de estudio, disponible en formatos diferentes a los artículos científicos y tesis, como lo son informes técnicos, memorias de congresos, consultas a expertos, resúmenes y estudios no publicados. Según diferentes autores (Beltrán, 2005; Petticrew y Roberts, 2006; Sánchez-Meca y Botella, 2010) el no incluir dichas fuentes de información favorece el sesgo, incluso McAuley et al., (2000), citado por Beltrán (2005) advierte que se puede dar un efecto de sobreestimación de la literatura que sí fue revisada cuantificado en alrededor de un 33%.

Otro elemento a considerar en relación a los criterios de inclusión fue el no considerar literatura en terceros idiomas, lo que redujo el aporte de publicaciones de algunas de las bases de datos consideradas en el estudio como lo fue en particular el caso de Scielo que cuenta con un acervo considerable de artículos en idioma portugués que no fueron contemplados.

A nivel de codificación el no contar con un sistema de doble ciego como lo recomiendan diferentes autores (Sánchez-Meca y Botella, 2010; Perestelo-Pérez, 2013) que permitiera además definir el grado de acuerdo entre condificadores pudo afectar la confiabilidad de este proceso favoreciendo el sesgo del investigador.

Siempre a nivel de la codificación si bien esta labor contempló la adaptación de un protocolo de revisión previamente publicado para la codificación de uno de los grupos de variables contempladas, no todas las variables contaron con protocolos previamente publicados que orientaran la codificación, sino que fueron de elaboración propia lo cual podría potencialmente generar un riesgo de sesgo del autor. Perestelo-Pérez (2013) advierte sobre este riesgo y menciona algunas recomendaciones de escalas potencialmente útiles para futuros estudios.

En cuanto al análisis de los datos al tratarse de una investigación de Revisión Sistemática y no de un Meta-análisis no fue posible la realización de un análisis estadístico de los datos obtenidos más complejo que incluyera entre otras cosas el tamaño del efecto de los diferentes estudios revisados, lo que a su vez no permitió dar

cuenta del grado de homogeneidad o heterogeneidad de las investigaciones, ni la generación de hipótesis sobre dichas diferencias como lo recomienda Sánchez-Meca y Botella (2010).

LÍNEAS DE INVESTIGACIÓN FUTURA

Es importante aspirar a niveles de evidencia más altos en la investigación de la evaluación de la simulación, para ello deben trascenderse las metodologías de estudios de casos y controles que como demuestra este estudio resultan ser las más populares para este tema de investigación. Propuestas metodológicas como las de Peace y Richards (2014) de ensayos controlados aleatorios o la de Lande y Blanks, (2013) de estudios de cohorte pueden considerarse ejemplos de buena práctica en el campo de estudio y como se mencionó anteriormente en este documento estas metodologías tienen la ventaja de responder potencialmente a más preguntas de investigación que otras (Petticrew y Roberts, 2006).

Siguiendo a los autores mencionados promover la Revisión Sistemática como metodología privilegiada para orientar la toma de decisiones clínicas a partir de la producción investigativa seria un paso deseable. Las iniciativas en esta línea podrían orientarse de manera más específica al estudio de instrumentos concretos y a la detección de la simulación de enfermedades y condiciones específicas como depresión, psicosis, estrés postraumático, problemas de memoria y deterioro cognitivo entre las principales.

De igual manera y en concordancia con los planteamientos de Hunsley y Mash (2011) es necesario trascender las iniciativas de Revisión Sistemática desde el enfoque de Instrumentos Basados en Evidencia para incursionar en la revisión de las practicas de integración de datos y la toma de decisiones clínicas a partir de dichos instrumentos, esto desde un enfoque más ambicioso de Evaluación Basada en Evidencia siendo las preguntas de investigación que buscan dar cuenta de la validez incremental y utilidad clínica las más relevantes.

Es importante recalcar que la producción científica en torno a la simulación de enfermedad mental no se ve representada a nivel latinoamericano de una manera

significativa, en este estudio por ejemplo fueron varios los artículos que trabajaron con muestras de habla hispana, no obstante de todas ellas solo un estudio reportó datos de población latinoamericana (Strutt, et al. 2012) el cual no obstante era de autoría estadounidense. Una mayor promoción de esta línea de investigación en los países de la región aportaría datos más sensibles y contextualizados a nuestra propia realidad, ya que como lo evidencian Strutt, et al. (2012) variaciones educativas y socioculturales pueden influir significativamente en la utilidad clínica de las herramientas para medir simulación.

En relación a la detección de la simulación de estrés postraumático como tema de estudio resulta relevante continuar explorando la utilidad potencial de más herramientas de medición del trauma y la simulación para este objetivo, ya que como advierte Gray (2010) citado por Peace y Richards (2014) aún no se han establecido estándares de oro para la evaluación de la simulación del trauma, entre otras razones por la misma complejidad y subjetividad en la expresión del constructo. Incluso esto deja el camino abierto para la elaboración de instrumentos especializados para la detección de la simulación de Trastorno de Estrés Postraumático.

Un segundo tópico de relevancia en este mismo tema evaluativo es la promoción de mayor investigación de simulación del estrés postraumático con etiologías diferentes a las experiencias militares traumagénicas de las cuales hay múltiples ejemplos de producción científica (Ahmadi, et al., 2013; Lande y Blanks, 2013). Como es sabido en nuestro país esta problemática específica es reducida en comparación con países involucrados en conflictos armados, pero en general no son ajenos a nuestro sistema de salud y a nuestra realidad social los casos de trauma asociados a abuso físico y sexual y a violencia doméstica entre otros.

En el área de evaluación de la simulación de problemas neuropsicológicos es importante trascender la medición psicométrica de los problemas de memoria y la amnesia para los que como se ha logrado evidenciar en este documento ya se cuenta con herramientas con demostrada validez predictiva como el TOMM.

Un modelo prometedor parecer ser el propuesto Ortega, et al. (2012) que utiliza técnicas inspiradas en el método bayesiano para la detección de simuladores con resultados que parecen superar a los obtenidos a través de test de validez de

síntomas (SVT) para detectar el esfuerzo subnormal asociad a la simulación en ésta área de evaluación, no obstante mayor investigación de esta metodología sería deseable.

Como Sánchez, et al. (2012) advierten una limitación para el avance el campo de estudio de la simulación de déficits cognitivos es la escases de pacientes reales con problemas a este nivel o su simulación, ellos proponen el uso de grupos experimentales con participantes entrenados para fingir estas condiciones como un primer paso.

CONCLUSIONES

Existe una alta producción de literatura científica asociada al fenómeno de la simulación, no obstante son pocos los estudios empíricos relevantes a nivel de evidencia científica que para el periodo 2012-2016 han sido publicados estando realmente accesibles.

Dichas investigaciones se orientan principalmente a una metodología de estudios de caso y control, existiendo vacíos de evidencia a partir de diseños más potentes como lo son los estudios de cohorte, estudios aleatorios controlados, revisiones sistemáticas y meta-análisis.

Los diseños de estudios de simulación y de grupos naturales parecen ser los más comunes en la revisión realizada, una combinación de ambas metodologías experimentales parece ser el mejor balance para garantizar la validez interna y externa de sus resultados.

Dentro de los estudios de simulación la tendencia parece ser diferenciar entre distintos grados de instrucción para los grupos, que van desde niveles de simulación ingenua hasta niveles expertos. La evidencia muestra diferencias en la tendencia de respuesta a los test de simulación según el grado de instrucción, no obstante los instrumentos no siempre logran diferenciar a estos subgrupos entre sí sino solo con respecto a pacientes reales y población normal no simuladora.

Las poblaciones más utilizadas para los estudios de simulación suelen ser estudiantes universitarios, lo que limita posibilidad de generalización de los hallazgos de estos estudios. Una alternativa que permite subsanar esta limitación es la realización de estudios con grupos naturales, idealmente en diferentes contextos de simulación y con diferentes motivaciones potenciales para el fingimiento, lo cual es un reto desde el punto de vista práctico.

Los artículos destacan el liderazgo europeo y norteamericano en esta línea de investigación. No obstante algún grado de evidencia muestra la necesidad de contar con investigación propiamente latinoamericana en cuanto a este tema, ya que

variables socioculturales y económicas podrían distorsionar la validez predictiva de dichos instrumentos con estas poblaciones.

Así mismo la investigación parece haberse centrado en el uso de los llamados test de validez de síntomas (VST), es decir aquellos especializados en detectar simulación como el SIRS-II, el TOMM y el M-Fast, no obstante algún grado de evidencia reflejan la utilidad limitada de estos como herramientas para la detección de la simulación.

Por mucho las principales propiedades psicométricas reportadas por los estudios revisados fueron las de sensibilidad (% de falsos positivos) y especificidad (% de falsos negativos) a nivel de validez predictiva y se asumieron como adecuadas otras reportadas por los manuales de los instrumentos elegidos. No obstante es necesario advertir que las propiedades psicométricas de los instrumentos de evaluación son específicas de las muestras a las que se aplican y no generales del instrumento, lo que justifica una revisión más a fondo de los test explorados.

Una serie de estudios han venido aportando evidencia sobre la capacidad de test de tipo clínico más tradicionales como el WMS-III y mediciones alternativas como las de método bayesiano para discriminar entre pacientes reales y simuladores, con la ventaja de que han demostrado ser metodologías eficientes entre otras bondades.

La combinación de diferentes instrumentos parece haber demostrado validez incremental a la detección de la simulación, esto fundamenta la necesidad de apegarse a la buena práctica internacional de apoyarse en estrategias evaluativas multimétodo-multifuente. Un reto en esta misma línea sería generar modelos de integración de la información y toma de decisiones clínicas.

REFERENCIAS

Adetungi, B., Basil, B., Mathews, M., Williams, A., Osinowo, T. y Oladinni, O. (2006) Detection and management of malingering in clinical settings. *Primary Psychiatry, 13*(1): 68-75.

Ahmadi, K., Lashabi, Z., Afzali, M. H., Tavalaie, S. A. y Mirzaee, J. (2013). Malingering and PTSD: Detecting malingering and war related PTSD by Miller Forensic Assessment of Symptoms Test (M-FAST). *BMC Psychiatry, 13*:154-158.

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (1999). *Standars for eduational and psychological testing*. Washington; Autor.

Bagby, R. M., Marshall, M. B. y Bacchiochi, J. R. (2005). The validity and clinical utility of the MMPI-2 Malingering Depression scale. *Journal of Personality Assessment, 85*(3): 304-11.

Beltrán, O. (2005). Revisiones sistemáticas de la literatura. *Revista Colombiana de Gastroenterología, 20*(1): 60-69.

Blasco, J. L. y Pallardó, L. (2013). Detección de exageración de síntomas mediante el SIMS y el MMPI-2-RF en pacientes diagnosticados de trastorno mixto ansioso-depresivo y adaptativo en el contexto medicolegal: un estudio preliminar. *Clínica y Salud 24*:177-183.

Carretero-Dios, H. y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clínical and Health Psychology, 5*(3): 521-551.

Chica-Urzola, L. L., Escobar-Córdoba, F. y Folino, J. O. (2005). La entrevista psiquiátrica del sujeto simulador. *Revista Colombiana de Psiquiatría, XXXIV*(1): 60-72.

Colegio Profesional de Psicólogos de Costa Rica. (2011a). *Código de ética y deontológico del Colegio Profesional de Psicólogos/as de Costa Rica*. Publicado en: La Gaceta No. 57 del martes 22 de marzo del 2011. San José: Costa Rica.

Comité Consultivo del Colegio Profesional de Psicólogos de Costa Rica. (2014). *Pautas para la Elaboración de Informes Psicológicos*. San José: CPPCR.

Conroy, M. A. y Kwartner, P. P. (2006). Malingering. *Applied Psychology in Criminal Justice, 2*(3): 29-51.

Consejo Federal de Psicología. (2008). *Sistema de avaliação dos testes psicológicos – SATEPSI*. São Paulo: Autor.

Denning, J. H. (2014). Combining the Test of Memory Malingering Trial 1 With Behavioral Responses Improves the Detection of Effort Test Failure. *Applied Neuropsychology: Adult, 21*: 269–277.

Fiscalía del Colegio Profesional de Psicólogos de Costa Rica. (2011). *Recomendaciones de la Fiscalía 01-2011. Algunas recomendaciones y consideraciones en la aplicación de Pruebas Psicológicas*. San José: CPPCR.

Fuermaier, A. B. M., Tucha, O., Koerts, J., Grabski, M., Lange, K. W., Weisbrod, M.,…Tucha, L. (2016). The Development of an Embedded Figures Test for the Detection of Feigned Attention Deficit Hyperactivity Disorder in Adulthood. *PLoS ONE, 11*(10): 25.

Gervais, R. O., Rohling, M. L., Green, P. y Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of clinical neuropsychology, 19*(4): 475-87.

González, H., Santamaría, P. y Capilla, P. (2012). *Estrategias de detección de la simulación. Un manual clínico multidisciplinar*. Madrid, España: TEA Ediciones.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G. y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 37*:192-197

Hunsley, J. y Mash, E. J. (2007). Evidence-Based Assessment. *Anual Rev. Clin. Psychol, 3*: 29-51.

Hunsley, J. y Mash, E. J. (2011). Evidence-Based Assessment [Chapter 5]. En: D. H. Barlow. *The Oxford Handbook of Clinical Psychology*. New York: Oxford University Press.

Inda, M., Lemos, S., López, A. M. y Alonso, J. L. (2005). La simulación de enfermedad física o trastorno mental. *Papeles del Psicólogo, 26*: 99-108.

International Test Commission. (2000). *Directrices Internacionales para el Uso de los Tests de la Comisión Internacional de Tests (ITC)*.[Traducción de la Comisión de Tests del Colegio Oficial de Psicólogos]. Madrid: Autor.

International Organization for Standarization. (2011). *Procedures and methods to assess people in work and organizational settings* (parte 1 y 2). Genova: Autor.

Jones, A. (2016). Repeatable Battery for the Assessment of Neuropsychological Status: Effort Index Cutoff Scores for Psychometrically Defined Malingering Groups in a Military Sample. *Archives of Clinical Neuropsychology, 31*: 273–283.

Killgore, W. D. y DellaPietra, L. (2000). Using the WMS-III to detect malingering: empirical validation of the rarely missed index (RMI). Journal of clinical and experimental neuropsychology, 22(6): 761-71.

King, J. y Sullivan, K. A. (2009). Deterring malingered psychopathology: The effect of warning simulating malingerers. *Behavioral sciences & the law, 27*(1): 35-49.

Kopf, T., Galic, S. y Matesic, K. (2016). The efficiency of MMPI-2 validity scales in detecting malingering of mixed anxiety-depressive disorder. *Alcoholism and Psychiatry Research, 52*: 33-50.

Lande, G. y Banks, L. (2013). Prevalence and Characteristics of Military Malingering. *Military medicine, 178*(1):50-55.

Liu, C., Liu, Z., Chiu, H. F. K., Carl, T. W-C., Zhang, H., Wang, P.,…Xue, Z. (2013). Detection of malingering: psychometric evaluation of the Chinese version of the structured interview of reported symptoms-2. *BMC Psychiatry, 13*: 254-270.

Masip, J. (2005). ¿Se pilla antes a un mentiroso que a un cojo? Sabiduría popular frente a conocimiento científico sobre la detección no verbal del engaño. *Papeles del Psicólogo. 26*: 78-91.

Meyer, G. J., Finn, S. E., Eyde, S. E., Kay, G. G., Moreland, K. L., Dies, R. R.,…Reed, G. M. (2001). Psychological testing and Psychological Assessment. A review of evidence and Issues. *American Psychologist, 56*(2); pp. 128-165.

Mora, C. (2013). *Simulación de síntomas psicóticos y su evaluación psicológica: algunos elementos psicojurídicos y su relación con aspectos clínicos y forenses en el*

contexto costarricense. Tesis para optar por el grado de Especialista en Psicología Clínica. San José; UCR/CCSS.

Muñiz, J. y Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á. y Peña-Suárez, E. (2011). *Evaluación de tests editados en España. Papeles del Psicólogo, 32*(2); pp. 113-128.

Novo, M., Fariña, F. y Arce, R. (2013). Eficacia del MMPI-A en casos forenses de acoso escolar; Simulación y daño psicológico. *Psychosocial Intervention, 22*: 33-40.

Ortega, A., Wagenmakers, E. J., Lee, M. D., Markowitsch, H. J. y Piefke, M. (2012). A Bayesian Latent Group Analysis for Detecting Poor Effort in the Assessment of Malingering. *Archives of Clinical Neuropsychology, 27*: 453–465.

Osuna, E., López-Martínez, M., Arce, R. y Vásquez, M. J. (2015). Analysis of response patterns on the MMPI-2 in psychiatric prison inmates. *International Journal of Clinical and Health Psychology, 15,* 29-36.

Peace, K. A. y Richards, V. E. S. (2014). *Faking it: incentives and malingered PTSD. Journal of Criminal Psychology, 4*(1): 19-32.

Perestelo-Pérez, L. (2013). Standards on how to develop and report systematic reviews in psychology and health. *International Journal of Clinical and Health Psychology, 13*(1), 49-57

Petticrew, M. y Roberts, H. (2006). *Systematic Reviews in the Social Sciences. A practical Guide*. Oxford, UK: Blackwell Publishing.

Rogers, R. (1997). *Clinical Assessment of Malingering and Deception* (2da edición). New York, EEUU: The Guilford Press.

Sánchez-Meca, J. y Botella, J. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional. *Papeles del Psicólogo, 31*(1): 7-17.

Sánchez, G., Jiménez, F., Ampudia, A. y Merino, V. (2012). In search of a fast screening method for detecting the malingering of cognitive imp airment. *The European Journal of Psychology Applied to Legal Context, 4*(2): 135-158.

Scott, C. (2006). *Assessment of malingerers and psychopaths: Fakes or Snakes?* Recuperado el 3 de noviembre del 2015 de: www.forensicmentalhealthassociation.org

Strutt, A. M., Scott, B. M., Lozano, V. J., Tieu, P. G. y Perry, S. (2012). Assessing sub-optimal performance with the Test of Memory Malingering in Spanish speaking patients with TBI. *Brain Injury, 26*(6): 853–863.

Sullivan, K. y Richer, C. (2002). Malingering on subjective complaint tasks: an exploration of the deterrent effects of warning. *Archives of clinical neuropsychology, 17*(7): 691-708.

Sullivan, L. E. (2000). *Malingering of head injury on neuropsychological instruments: A meta-analytic review*. Tesis de Doctorado en Filosofía de la Universidad Simon Fraser.

Urbina, S. (2007). *Claves para la evaluación con tests psicológicos*. Madrid: TEA Ediciones.

Vilar, R., Aparicio, M., Gómez, M. y Pérez, M. (2013). Utilidad de los índices de memoria verbal para detectar simulación en población española. *Clínica y Salud, 24*: 169-176.

Vilariño, M., Arce, R. y Fariña, F. (2013). Forensic-clinical interview: reliability and validity for the evaluation of psychological injury. *The European Journal of Psychology Applied to Legal Context, 5*(1): 1-21

Walters, G., Berry, D., Duncan, C., Payne, J., Rogers, R., Miller, H.,....Granacher, R. (2008). Malingering as a Categorical or Dimensional Construct: The Latent Structure of Feigned Psychopathology as Measured by the SIRS and MMPI–2. *Psychological Assessment. 20*(3): 238–247.

Weiner, I. B. (2003). The assessment process. En J.R. Graham y J.A. Naglieri (Ed.), *Handbook of Psychology. Assessment Psychology*. [Volume 10]. New Jersey: John Wiley and Sons, Inc.

Woods, D. L., Wyma, J. M., Herron, T. J. y Yund, E. W. (2015) The Effects of Aging, Malingering, and Traumatic Brain Injury on Computerized Trail-Making Test Performance. *PLoS ONE, 10*(6): 30.

Young, S., Jacobson, R., Einzig, S., Gray, K. y Gudjonsson, G. H. (2016). Can we recognise malingerers? The association between malingering, personality traits and clinical impression among complainants in civil compensation cases. *Personality and Individual Differences, 98*: 235–238.

ANEXOS

Anexo 1. Protocolo de registro de variables de la publicación para artículos elegidos

VARIABLES EXTRÍNSECAS		
1. Nombre del artículo		
2. Año de la publicación		
3. Nombre de los autores		
4. Nombre de la revista		
VARIABLES DE LOS PARTICIPANTES		
5. Edad promedio de los participantes		
6. Composición por género		
Masculino		
Femenino		
7. País de procedencia de la muestra		
VARIABLES METODOLÓGICAS		
8. Tipo de muestra		
Población general	Población clínica	
Población forense	Otra	Mixta
9. Tamaño muestral		
10. Selección de la muestra		
11. Instrumentos utilizados		
12. Método de evaluación del instrumento		
Observación estructurada	Test en base a normas general	
Entrevista estructurada	Test en base a normas específico	
Test proyectivo	Medición fisiológica	

Anexo 2. Protocolo de registro de variables del método para instrumentos elegidos

13. Normas		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
14. Consistencia interna		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
15. Confiabilidad inter-evaluador		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
16. Confiabilidad test-retest		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
17. Validez de contenido		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
18. Validez de constructo		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
19. Generalización de la validez		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente
20. Utilidad clínica		
Inviable	No aplica	Menos que adecuada
Adecuada	Buena	Excelente

Anexo 3. Traducción de escala de valoración de instrumentos de Hunsley y Mash

NORMAS
Adecuada: Están disponibles medidas de tendencia central y distribución para la puntuación total (y sus subescalas si es relevante) basadas en una amplia y relevante muestra clínica.
Buena: Están disponibles medidas de tendencia central y distribución para la puntuación total (y sus subescalas si es relevante) basadas en varias muestras amplias y relevantes (la mayoría incluye datos de muestras clínicas y no clínicas)
Excelente: Están disponibles medidas de tendencia central y distribución para la puntuación total (y sus subescalas si es relevante) basadas en varias muestras representativas (la mayoría incluye datos de muestras clínicas y no clínicas)
CONSISTENCIA INTERNA
Adecuada: Evidencia preponderante indica valores alpha de .70-.79
Buena: Evidencia preponderante indica valores alpha de .80-.89
Excelente: Evidencia preponderante indica valores alpha de ≥.90
CONFIABILIDAD INTEREVALUADOR
Adecuada: Evidencia preponderante indica valores *k* de .60-.74; la evidencia preponderante indica valores de correlación o correlación intraclase de .70-.79
Buena: Evidencia preponderante indica valores *k* de .75-.84; la evidencia preponderante indica valores de correlación o correlación intraclase de .80-.89
Excelente: Evidencia preponderante indica valores *k* ≥.85; la evidencia preponderante indica valores de correlación o correlación intraclase . ≥.90
CONFIABIILIDAD TEST-RETEST
Adecuada: Evidencia preponderante indica correlaciones test-retest de al menos .70 para un periodo de varios días a varias semanas
Buena: Evidencia preponderante indica correlaciones test-retest de al menos .70 para un periodo de varios meses
Excelente: Evidencia preponderante indica correlaciones test-retest de al menos .70 para un periodo de un año o más
VALIDEZ DE CONTENIDO
Adecuada: Los desarrolladores definieron claramente el dominio del constructo siendo evaluado y asegurándose que eligieron los ítems que representan la totalidad de facetas incluidas.
Buena: Además del criterio usado para la clasificación de adecuada, todos los elementos del instrumento (ej. instrucciones, items) fueron evaluados por jueces (ej. por expertos, por participantes del estudio piloto)
Excelente: Además del criterio usado para la clasificación de buena, se usaron múltiples grupos de jueces y usaron una clasificación cuantitativa.

Continuación…

VALIDEZ DE CONSTRUCTO
Adecuada: Hay alguna evidencia de validez de constructo replicada independientemente (ej. validez predictiva, concurrente, convergente y discriminante)
Buena: evidencia de validez de constructo replicada independientemente preponderante a través de múltiples tipos de validez (ej. validez predictiva, concurrente, convergente y discriminante), es indicativo de validez de constructo.
Excelente: Además del criterio para clasificación buena, hay evidencia de validez incremental con respecto a otros datos clínicos.
GENERALIZACIÓN DE LA VALIDEZ
Adecuada: Alguna evidencia soporta el uso de este instrumento para más de un grupo específico (basado en características sociodemográficas como edad, género y etnicidad) o en múltiples contextos (ej. hogar, escuela, contextos de cuidado primario, ámbitos de internamiento)
Buena: Evidencia preponderante soporta el uso de este instrumento para más de un grupo específico (basado en características sociodemográficas como edad, género y etnicidad) o en múltiples contextos (ej. hogar, escuela, contextos de cuidado primario, ámbitos de internamiento)
Excelente: Evidencia preponderante soporta el uso de este instrumento para más de un grupo específico (basado en características sociodemográficas como edad, género y etnicidad) y a través de múltiples contextos (ej. hogar, escuela, contextos de cuidado primario, ámbitos de internamiento)
UTILIDAD CLÍNICA
Adecuada: Tomando en cuenta consideraciones prácticas (ej. costo, facilidad de aplicación, disponibilidad de instrucciones para la aplicación y la calificación, duración de la evaluación, disponibilidad de puntos de corte relevantes, aceptación de pacientes), los datos resultantes de la evaluación son probablemente útiles clínicamente.
Buena: Además del criterio usado para clasificación adecuada, hay alguna evidencia publicada de que el uso de los datos resultantes de la evaluación confiere un beneficio clínico demostrable (ej. mejor resultado del tratamiento, menor tasa de abandono del tratamiento, mayor satisfacción de los pacientes con los servicios).
Excelente: Además del criterio usado para clasificación adecuada, hay evidencia replicada de forma independiente publicada de que el uso de los datos resultadantes de la evaluación confiere un beneficio clínico demostrable.

Tomado de: Hunsley y Mash (2011, p. 86 y 87)

Anexo 4. Estadísticos descriptivos de las variables moduladoras del estudio

# de Variable	Código de Variable	Variable	Codificación	Tipo	Categoría	Tipo de Análisis
1	Articulo	Nombre del artículo	No aplica	Cualitativa	Nominal	Cualitativo
2	Año_pub	Año de publicación	2012-2016	Cuantitativa	Continua	Frecuencia, rango
3	Autores	Nombres de los autores	No aplica	Cualitativa	Nominal	Cualitativo
4	Revista	Nombre de la revista	No aplica	Cualitativa	Nominal	Frecuencia
5	Prom_edad	Promedio de edad de la muestra	99 No disponible	Cuantitativa	Continua	Cualitativo
6	Porc_masc	Porcentaje de la muestra sexo masculino	No aplica	Cuantitativa	Continua	Cualitativo
7	Porc_fem	Porcentaje de la muestra sexo femenino	No aplica	Cuantitativa	Continua	Cualitativo
8	Nac_muest	Nacionalidad de la muestra	No aplica	Cualitativa	Nominal	Conteo simple, frecuencia
9	Tipo_muest	Tipo de muestra	1 Población general 2 Población clínica 3 Población forense 4 Otra 5 Varias muestras 99 No disponible	Cualitativa	Nominal	Conteo, frecuencia, porcentaje
10	Tam_muest	Tamaño muestral	No aplica	Cuantitativa	Continua	Cualiatativo
11	Sele_muest	Criterio de selección de la muestra	1 Conveniencia 2 Intencional 3 Aleatoria 4 Estratificada 5 Conglomerados 99 No disponible	Cualitativa	Nominal	Frecuencia, porcentaje

Continuación…

12	Instr_1	Instrumento 1	No aplica	Cualitativa	Nominal	Frecuencia
13	Instr_2	Instrumento 2	No aplica	Cualitativa	Nominal	Frecuencia
14	Instr_3	Instrumento 3	No aplica	Cualitativa	Nominal	Frecuencia
15	Método	Método de evaluación del instrumento	1 Observación estructurada 2 Entrevista estructurada 3 Test proyectivo 4 Test psicométrico general 5 Test psicométrico específico 6 Medición fisiológica 99 No disponible	Cualitativa	Nominal	Frecuencia, porcentaje
16	Normas	Normas	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
17	Con_intern	Consistencia interna	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
18	Con_eval	Consistencia inter-evaluador	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
19	Con_retest	Consistencia test-retest	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia

Continuación…

20	Val_cont	Validez de contenido	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
21	Val_constr	Validez de constructo	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
22	Gen_val	Generalización de la validez	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia
23	Utilidad	Utilidad Clínica	1 Inviable 2 No aplica 3 Menos que adecuada 4 Adecuada 5 Buena 6 Excelente	Cualitativa	Nominal	Frecuencia