Psicopatología Clínica, Legal y Forense, Vol. 13, 2013, pp.105-137.
ISSN: 1576-9941

DIRECTRICES PARA SELECCIONAR TEST PSICOLÓGICOS EN EL ÁMBITO CLÍNICO FORENSE

Jesús Sanz[1]
María Paz García Vera
Universidad Complutense de Madrid

Resumen
El considerable desarrollo de la evaluación psicológica en España ha puesto a disposición de los psicólogos que trabajan en el ámbito clínico forense un número muy grande de test psicológicos y, en consecuencia, una de las tareas a las que deben enfrentarse es decidir qué test en concreto deberían utilizar en una evaluación forense dada. Esta decisión implica, en primer lugar, determinar los objetivos de la evaluación forense y la población a la que pertenece la persona evaluada, y, en segundo lugar, valorar la adecuación a esos objetivos y a esa población de las características del test y de las interpretaciones de las medidas que ese test proporciona. Para hacer esta valoración, Heilbrun (1992) propuso siete directrices que tienen en cuenta criterios psicométricos así como criterios de disponibilidad, documentación, aplicación, interpretación y relevancia para las cuestiones forenses. En este trabajo, se han actualizado, ampliado y especificado esas directrices con el objetivo último de que las mismas puedan servir de ayuda a los psicólogos españoles que trabajan en el ámbito clínico forense. Finalmente, se ejemplifica la utilidad de esas directrices analizando a partir de ellas los test más utilizados para la evaluación de la gravedad de la depresión en la población clínica española.
PALABRAS CLAVE: *evaluación psicológica, evaluación forense, tests psicológicos, buenas prácticas.*

Abstract
Psychological assessment in Spain has undergone considerable development and, therefore, psychologists working in clinical forensic settings have at their disposal a very large armamentarium of psychological tests. Consequently, one of the tasks faced by these psychologists is to decide which specific test should be used in a forensic assessment given. This decision involves, first, determining the objectives of the forensic evaluation and the population to

[1] *Correspondencia:* Jesús Sanz. Departamento de Personalidad, Evaluación y Psicología Clínica. Facultad de Psicología. Universidad Complutense de Madrid. Campus de Somosaguas. 28223 Pozuelo de Alarcón (Madrid). España. Correo electrónico: jsanz@psi.ucm.es
Fecha de recepción del artículo: 11-10-2013.
Fecha de aceptación del artículo: 26-11-2013

which the assessed person belongs to, and, second, to assess the adequacy of test characteristics and interpretations of test measures to those objectives and that population. To make this assessment, Heilbrun (1992) proposed seven guidelines that take into account psychometric criteria as well as criteria of availability, documentation, administration, interpretation and relevance to forensic issues. In this paper, those guidelines have been updated, expanded and specified with the ultimate goal that they can be of assistance to Spanish psychologists working in clinical forensic settings. Finally, in order to illustrate their usefulness, an analysis based on these guidelines is carried out on the most commonly used tests for assessing the severity of depression in Spanish clinical population.
KEYWORDS: *psychological assessment, forensic evaluation, psychological tests, best practices.*

Introducción

Existen muchos datos para considerar que la evaluación psicológica en España goza en la actualidad de un nivel de desarrollo importante. Por un lado, Buela-Casal, Sierra, Carretero-Dios y de los Santos-Roig (2002), tras analizar el número de artículos sobre evaluación psicológica publicados en tres revistas españolas representativas de la producción científica española en psicología, una más centrada en la psicología clínica y de la salud (*Análisis y Modificación de Conducta*) y otras dos de temática más general (*Psicothema* y *Revista de Psicología General y Aplicada*), encontraron que estas revistas tenían, respectivamente, un 60%, 36% y 22% de artículos sobre evaluación. Es decir que, por ejemplo, 1 de cada 3 artículos de *Psicothema* versan sobre evaluación psicológica, lo cual es muy relevante dado no sólo el carácter tan general de esa revista, sino también su prestigio y difusión entre la comunidad científica española en psicología. Entre las revistas de psicología españolas, *Psicothema* ocupa la primera posición según el índice H de impacto o difusión basado en citas bibliográficas de *Google Scholar* (Delgado López-Cózar, Marcos Cartagena, Jiménez Contreras y Ruiz Pérez, 2013), la segunda posición según la opinión del propio conjunto de especialistas españoles en psicología que ofrece la plataforma de indicadores de calidad de revistas científicas españolas RESH (Grupo de Investigación de Evaluación de Publicaciones Científicas y Grupo de Investigación de Evaluación de la Ciencia y de la Comunicación Científica, 2013) y también la segunda posición según el índice IN-RECS de impacto o difusión basado en citas bibliográficas españolas y acumulado para los años 2000-2009 y 2005-2009 (Grupo de Investigación de Evaluación de la Ciencia y de la Comunicación Científica, 2013). Por otro lado, cuando se pregunta a los profesores universitarios españoles del área de conocimiento de personalidad, evaluación y tratamiento psicológico cuál es su campo principal de actividad

investigadora, el mayor porcentaje de ellos, igualado con el campo de psicopatología, contesta que el campo de evaluación psicológica y psicodiagnóstico (Sanz, 2002). Finalmente, los resultados de una encuesta de opinión sobre la práctica de los test realizada en 2009 a 3.126 psicólogos profesionales españoles, todos ellos miembros del Colegio Oficial de Psicólogos, indican que éstos tienen una actitud favorable hacia los test cuando se utilizan combinados con otros datos psicológicos, los utilizan habitualmente en su desempeño profesional y consideran que en la última década el uso de los test en España ha mejorado (Muñiz y Fernández-Hermida, 2010).

Una de las consecuencias deseables de ese notable desarrollo de la evaluación psicológica en España es que, en la actualidad, el número de test psicológicos que tienen a su disposición los psicólogos clínicos forenses españoles es muy numeroso. Por ejemplo, sólo teniendo en cuenta un tipo de test muy concreto, los cuestionarios, escalas e inventarios autoaplicados para evaluar la depresión, Sanz, Izquierdo y García-Vera (2013) realizaron recientemente una búsqueda en las base de datos bibliográficos Psicodoc y PsycINFO sobre su utilización en estudios realizados por psicólogos españoles. Esta búsqueda dio lugar, desde enero de 1990 a diciembre de 2012, a 416 publicaciones en Psicodoc y a 746 en PsycINFO, a partir de las cuales Sanz et al. (2013) pudieron identificar 31 cuestionarios, escalas e inventarios autoaplicados que: (a) habían sido desarrollados con el objetivo específico de evaluar la depresión clínica en la población clínica adulta en general y (b) habían sido creados en España, tenían una adaptación española o contaban con al menos un estudio sobre sus propiedades psicométricas en población clínica española. Es decir, que para evaluar la depresión en adultos, los psicólogos españoles pueden elegir entre un mínimo de 31 cuestionarios, escalas e inventarios autoaplicados, a los cuales habría que añadir los heteroaplicados, los centrados en la evaluación de un único síntoma depresivo, los desarrollados específicamente para ciertas poblaciones adultas como, por ejemplo, ancianos, mujeres en el período posparto o pacientes con trastornos específicos (p. ej., esquizofrenia), así como los desarrollados para evaluar en la población general adulta la tristeza, el estado de ánimo deprimido, el afecto negativo, la depresión rasgo y constructos similares. De hecho, actualmente existe un número tan grande de cuestionarios, escalas e inventarios para medir la depresión en adultos, que el problema al que se enfrentan los profesionales e investigadores en el ámbito de la depresión es, precisamente, seleccionar de ese amplio repertorio cuál es el más apropiado para sus objetivos.

En definitiva, el considerable desarrollo de la evaluación psicológica en España ha traído consigo un enorme aumento del número de test psicológicos adaptados a la población española o creados para la población española, lo cual debería ser motivo de satisfacción, pero, paradójicamente, esta deseable riqueza supone que una de las tareas más complejas que deben llevar a cabo los psicólogos clínicos forenses en España, al igual que ocurre en otros países con

un cierto desarrollo de la psicología, es decidir qué test psicológicos son los más útiles para su actividad profesional, dado que las posibilidades existentes son muy numerosas y variadas. Con el objetivo de ayudar en esta tarea, en este artículo se revisarán las directrices propuestas por Heilbrun (1992) para determinar si un test psicológico dado debería utilizarse en una evaluación forense, se actualizarán, ampliarán y concretarán esas directrices y, finalmente, se pondrá un ejemplo de su aplicación en el caso de los test más utilizados en España para evaluar la gravedad de la depresión.

La selección de un test psicológico para su uso en el ámbito clínico forense

Los objetivos para los cuales se pueden utilizar los test psicológicos en el ámbito clínico en general y en el área de la psicología clínica forense en particular, son muy variados: cribado o despistaje (*screening*), diagnóstico y clasificación, descripción y análisis de síntomas y áreas problemáticas, descripción y análisis de la personalidad, las aptitudes o la inteligencia, descripción de factores psicosociales, formulación del caso y comprobación de hipótesis clínicas, planificación de programas de prevención o tratamiento, predicción de conductas relevantes (p. ej., reincidencia criminal, conductas de suicidio, violencia), evaluación de los resultados de los programas de prevención o tratamiento, etc., y la mayoría de las situaciones clínicas suelen demandar la consecución de varios de esos objetivos. Aunque algunos test pueden ser útiles para conseguir de forma aceptable varios de esos objetivos, es poco probable que un único test sea apropiado para alcanzar todos ellos con unas mínimas garantías. Por tanto, la selección de un test psicológico para la evaluación en el ámbito clínico forense implica, en primer lugar, determinar los objetivos de dicha evaluación y la población a la que pertenece la persona o personas que van a ser evaluadas, y, en segundo lugar, valorar la adecuación a esos objetivos y a esa población de las características del test y de las interpretaciones o inferencias de las medidas que ese test proporciona.

Entre dichas características habría que tener en cuenta, por un lado, qué relación costes-beneficios presenta el test respecto a aspectos prácticos tales como la cantidad de tiempo que demanda del evaluado y del evaluador, su grado de complejidad cognitiva (p. ej., legibilidad, comprensibilidad), su utilidad incremental respecto a la información que ya se ha obtenido por otros test, medios o instrumentos, o su complejidad a la hora de puntuarse e interpretarse. Por otro lado, y de manera fundamental, habría que tener en cuenta cuáles son las propiedades psicométricas de las interpretaciones o inferencias de las medidas o puntuaciones que ofrece el test, es decir, en qué grado las puntuaciones del test miden lo que se quiere medir para conseguir los objetivos

propuestos (validez) y en qué grado tales puntuaciones son consistentes o estables (fiabilidad).

En resumen, y tal y como la Comisión Internacional de Tests (2000) ha resaltado, para optimizar el uso adecuado de los test es importante elegir test técnicamente correctos y adecuados a cada situación, así como estimar la utilidad potencial del test para la situación evaluativa en cuestión.

Además, la utilización de test psicológicos en la evaluación psicológica en el ámbito clínico forense no sólo se debe ajustar a los objetivos que justifican su realización, sino que también debería integrarse con la utilización e información proporcionada por otras técnicas de evaluación, particularmente las entrevistas (clínicas, semiestructuradas o estructuradas), en el contexto de una exploración psicopatológica completa y del acceso a información procedente de terceras personas, aspecto este último muy importante en el ámbito forense (Heilbrun, Warren y Picarello, 2003). Así, por ejemplo, los test psicológicos se deberían seleccionar, aplicar e interpretar tras revisar y tener en cuenta datos documentales tales como historias clínicas, informes de otros profesionales, sentencias, expedientes y protocolos penitenciarios, etc., y, en la medida de lo posible, también se deberían llevar a cabo entrevistas o aplicar test psicológicos a terceras personas relevantes (p. ej., esposos o parejas, familiares, amigos, vecinos, jefes y compañeros de trabajo, personal médico, etc.), todo lo cual debería ayudar a evaluar los aspectos forenses del caso.

Sin embargo, estas recomendaciones, por muy razonables y plausibles que sean, son muy generales y para que los psicólogos clínicos forenses las puedan poner en práctica en su actividad profesional sería necesario un mayor nivel de concreción y de ajuste a las características propias del ámbito clínico forense.

¿Qué test psicológicos son más útiles para la evaluación en Psicología clínica forense?

Como cabría suponer, no existe una respuesta única a esta pregunta puesto que, como ya se ha mencionado antes, la respuesta viene determinada por los objetivos de la evaluación, estos objetivos pueden ser muy diversos y ningún instrumento por sí solo parece cubrir con garantías todos ellos. En consecuencia, la selección de un test psicológico para la evaluación en psicología clínica forense implica, en primer lugar, determinar los objetivos de dicha evaluación y la población a la que pertenece la persona o personas que van a ser evaluadas, y, en segundo lugar, valorar la adecuación a esos objetivos y a esa población de las características del instrumento y de las interpretaciones o inferencias de las medidas que esos instrumentos proporcionan, para lo cual es necesario tener en cuenta simultáneamente varios criterios tanto psicométricos y prácticos como de relevancia para las cuestiones forenses.

Tratando de concretar esos criterios, Heilbrun (1992, pp. 264-267) especificó siete directrices para ayudar a los psicólogos a determinar si un test psicológico debería usarse en el ámbito forense.

1. "El test está disponible comercialmente y documentado de forma adecuada en dos fuentes de referencia. Primero, se acompaña de un manual que describe su desarrollo, propiedades psicométricas y procedimiento de aplicación. Segundo, aparece listado y revisado en el *Mental Measurement Yearbook* o en alguna otra fuente de referencia fácilmente accesible".
2. "Se debería considerar la fiabilidad. El uso de un test con un coeficiente de fiabilidad menor de 0,80 no es aconsejable. La utilización de un test menos fiable requeriría una justificación explícita por parte del psicólogo".
3. "El test debería ser relevante para la cuestión legal o para un constructo psicológico que subyazca tras la cuestión legal. Cuando sea posible, esta relevancia debería estar apoyada en la existencia de investigación de validación publicada en revistas con revisión por pares".
4. "Debería utilizarse una aplicación estandarizada, con unas condiciones de aplicación del test tan cercanas como sea posible al ideal de tranquilidad y ausencia de distracciones".
5. "Tanto la selección de un test como su interpretación deberían guiarse por la aplicabilidad a una población concreta y para un propósito dado. Los resultados de un test (distintos del comportamiento observado durante su administración) no deberían aplicarse a un propósito para el cual el test no fue desarrollado (p. ej., inferir psicopatología a partir de los resultados de un test de inteligencia). La especificidad de la población y de la situación deberían guiar la interpretación. Cuanto mayor sea el "ajuste" entre un individuo dado y la población y situación utilizadas en la investigación de validación, más confianza se puede tener en la aplicabilidad de los resultados".
6. "Los test objetivos y los datos actuariales son preferibles cuando hay datos apropiados de resultado y existe una "fórmula"".
7. "Se debería evaluar explícitamente el estilo de respuesta usando aproximaciones sensibles a la distorsión, y se deberían interpretar los resultados de la aplicación del test dentro del contexto del estilo de respuesta del individuo. Cuando el estilo de respuesta parezca ser de simulación, defensivo o irrelevante en lugar de sincero/fiable, quizás sea necesario minimizar la importancia de los resultados de la aplicación del test o incluso ignorarlos y enfatizar en mayor medida otras fuentes de datos".

Es obvio que las directrices de Heilbrun (1992) son una importante contribución a la hora de concretar unos criterios tanto psicométricos como de disponibilidad, documentación, aplicación, interpretación y relevancia que ayuden en la tarea de seleccionar los test psicológicos más adecuados para una evaluación forense dada. Sin embargo, en algunos aspectos tales directrices podrían concretarse aún un poco más para así facilitar su aplicación. Por ejemplo, la directriz tercera propone, como no podría ser de otra manera, que los test deben ser relevantes bien para la cuestión legal implicada o bien para el constructo psicológico que subyace tras esa cuestión legal, pero, además, señala, muy acertadamente, que esa relevancia se debe evaluar consultando los datos empíricos de validación del test que existan en los estudios publicados en revistas con revisión por pares o expertos, es decir, en la literatura científica que cumple unos mínimos estándares de calidad[2]. Sin embargo, la directriz tercera no concreta qué tipo de datos de validación serían apropiados y cómo se podrían interpretar esos datos. En los siguientes epígrafes se concretará un poco más esa directriz así como algunas otras de las propuestas por Heilbrun (1992) y, puesto que las directrices fueron propuestas hace más de 20 años, se actualizarán algunas referencias que incluyen y tanto su contenido como las referencias se ajustarán al contexto español, confiando en que con todos estos cambios puedan ayudar aún más a los psicólogos clínicos forenses españoles en la selección de los test más adecuados para su práctica profesional o investigadora.

Disponibilidad y documentación del test

1. *"El test está disponible comercialmente y documentado de forma adecuada en dos fuentes de referencia. Primero, se acompaña de un manual que describe su desarrollo, propiedades psicométricas y procedimiento de aplicación. Segundo, aparece listado y revisado en el Mental Measurement Yearbook o en alguna otra fuente de referencia fácilmente accesible"* (Heilbrun, 1992, p. 264).

Entre esas fuentes de referencia fácilmente accesibles, Heilbrun (1992) mencionaba las revisiones de Brodsky y Smitherman (1983) o Grisso (1986), a

[2] En las revistas que cuentan con sistema de revisión por pares o expertos, cada artículo recibido para su publicación es leído y analizado por dos o más evaluadores o revisores que determinan tanto la validez de las ideas, el método y los resultados como su significación para la ciencia y la profesión, y tanto del estudio en sí mismo como de su presentación. Estos revisores los eligen los editores de las revistas entre los investigadores con más prestigio en las diferentes disciplinas y áreas de la misma. Aunque este sistema no está exento de críticas es, sin embargo, el más utilizado ya que no existen alternativas mejores consolidadas (Benos et al., 2007).

las que actualmente se podrían unir otras más actualizadas, tanto generales del ámbito clínico (p. ej., Antony, Orsillo y Roemer, 2001; Corcoran y Fischer, 2013; Nezu, Ronan, Meadows y McClure, 2000; Ronan, Dreer, Maurelli, Ronan y Gerhart, 2014) como específicas del ámbito forense (p. ej., Grisso, 2005). En España, algunas fuentes de referencia generales del ámbito clínico podrían ser Bulbena, Berrios y Fernández de Larrinoa Palacios (2000), Caballo (2005, 2006), G.-Portilla González, Bascarán Fernández, Sáiz Martínez, Parallada Redondo, Bousoño García y Bobes García (2011) y Muñoz, Roa, Pérez Santos, Santos-Olmo y de Vicente (2002).

Por otro lado, quizás pueda resultar extraño que Heilbrun (1992) requiera que los test estén comercializados. Posiblemente, el motivo de incluir este requisito tiene que ver con el hecho de que si un test está comercializado es más probable que su accesibilidad sea mayor para los profesionales y que esté documentado y, además, dada la protección de derechos intelectuales que implica la comercialización, es también más probable que se pueda evitar en mayor medida la proliferación de múltiples versiones de un instrumento sin las garantías de calidad adecuadas. En este sentido, puede servir de ejemplo paradigmático la situación de la Escala de Valoración de la Depresión de Hamilton (HAM-D, HRSD o HDRS; Hamilton, 1960, 1967), la escala heteroaplicada que durante años ha sido el instrumento de evaluación clínica más utilizado para medir la gravedad de la depresión y la referencia más importante para evaluar la eficacia de los tratamientos farmacológicos para la depresión (Nezu, Nezu, Friedman y Lee, 2009; Trajković et al., 2011) y que está en dominio público.

La versión original de la HDRS incluía 17 ítems, aunque en la hoja de registro de las puntuaciones también se incluían cuatro ítems adicionales que según el autor no tenían relevancia para medir la gravedad de la depresión (Hamilton, 1960, pp. 56-57). En un trabajo posterior, Hamilton (1967) publicó solo la versión de 17 ítems con ligeras modificaciones respecto a la previa. Aunque estas versiones de 17 ítems son las más utilizadas, existen más de 20 versiones de la HDRS que difieren, entre otros aspectos, en el número de ítems que la componen (6, 7, 21, 24, 25 y 27 ítems), en el formato (p. ej., autoaplicada de lápiz y papel, autoaplicada por ordenador, con entrevista estructurada) e incluso en el contenido (p. ej., modificaciones en la redacción de los ítems, inclusión de nuevos ítems o de más descripciones de los ítems) (véase una revisión en Williams, 2001). Desafortunadamente, los investigadores y profesionales no siempre informan bien de qué versión han utilizado o están utilizando en su estudios o en su práctica clínica. Así, Zitman, Mennen, Griez y Hooijer (1990) solicitaron a los autores de varios estudios publicados en revistas de prestigio que les enviaran una copia de la HDRS que realmente habían utilizado en dichos estudios, y, para su sorpresa, encontraron que sólo 4 de los

51 autores que respondieron habían utilizado una versión que era igual a la de la publicación que citaban como fuente de la HDRS; además, menos de la mitad citaban la publicación correcta. Es decir, se estaban utilizando diferentes versiones de la HDRS sin tener en cuenta las consecuencias a nivel de sus respectivas propiedades psicométricas, ya que muchas de esas versiones no habían sido objeto de análisis psicométrico alguno.

Lamentablemente, ese es un problema común a otros instrumentos que tienen varias versiones, que no cuentan con una versión comercial del mismo y que gozan de gran popularidad y difusión (p. ej., la primera edición del Inventario de Depresión de Beck o BDI), problema que se agudiza cuando existen además varias traducciones y adaptaciones a otros idiomas. Siguiendo con el ejemplo de la HDRS, en España existen dos adaptaciones de la escala. Por un lado, Conde y Franch (1984) realizaron una traducción al español de la versión de 21 ítems y, posteriormente, dicha traducción, con ligerísimos cambios, fue validada por Bobes et al. (2003). Por otro lado, Ramos-Brieva y Cordero Villafáfila (1986a) realizaron una traducción al español de la versión de 17 ítems de Hamilton (1967), pero en la que introdujeron modificaciones propias (p. ej., exigir que el paciente se despierte al menos dos horas antes de lo habitual durante tres o cuatro días para puntuar 2 en el ítem 6 de insomnio tardío) y modificaciones tomadas de las versiones de otros autores (la de Guy, 1976, la del grupo de Michigan de Feinberg et al., 1985, y la de Rehm y O'Hara, 1985, todos ellos citados por Ramos-Brieva y Cordero Villafáfila, 1986, pp. 326-327), y llevaron a cabo los estudios de adaptación correspondientes (Cordero Villafáfila y Ramos-Brieva, 1986; Ramos-Brieva y Cordero Villafáfila, 1986a, 1986b, 1988; Ramos Brieva, Cordero Villafáfila y Yánez Sáez, 1994). Cuando se comparan ambas versiones españolas, las diferencias entre ellas afectan prácticamente a todos los ítems y son más que notables. Por citar algunas, en el ítem 14 (síntomas genitales), en la versión de Ramos-Brieva y Cordero Villafáfila no se incluyen los trastornos menstruales, sólo la pérdida de la libido, usa una escala de 0 a 2 ("ausente", "ligero" y "pérdida completa de apetito sexual") en lugar de 0 a 3 ("ausente", "débil", "grave" e "incapacitante") y, además, incluye indicaciones adicionales para hacer las valoraciones (p. ej., "1 ligero: descenso de la libido: actividad sexual alterada (inconstante, poco intensa)"), mientras que en el ítem 1 ("estado de ánimo deprimido") las indicaciones para hacer las valoraciones difieren en las dos versiones (p. ej., "4 extremo: llanto muy frecuente (o ganas); frecuente tendencia al aislamiento; contenidos depresivos exclusivos en el pensamiento o la comunicación verbal; pérdida de la capacidad de reacción a estímulos placenteros" en la versión de Ramos-Brieva y Cordero Villafáfila, en lugar de "4. El paciente manifiesta estas sensaciones en su comunicación verbal y no verbal de forma espontánea" en la versión de Conde y Franch). Aunque afortunadamente ambas versiones

españolas están validadas, dadas las numerosas y notables diferencias entre ellas, no se puede asumir que ambas identifiquen los mismos síntomas depresivos y arrojen puntuaciones parecidas de gravedad de la depresión.

Propiedades psicométricas

2. *"Se debería considerar la fiabilidad. El uso de un test con un coeficiente de fiabilidad menor de 0,80 no es aconsejable. La utilización de un test menos fiable requeriría una justificación explícita por parte del psicólogo"* (Heilbrun, 1992, p. 265).

Aunque un coeficiente de fiabilidad de 0,80 parece adecuado (véase Prieto y Muñiz, 2010), lo cierto es que ese criterio no tiene en cuenta los tipos de fiabilidad existentes (p. ej., los coeficientes de fiabilidad test-retest suelen ser menores que los coeficientes de consistencia interna), ni los tipos de test (p. ej., con algunos test psicopatológicos y de personalidad es difícil conseguir coeficientes de consistencia interna superiores a 0,70, mientras que algunos test de inteligencia o de aptitudes llegan a alcanzar coeficientes superiores a 0,90) ni las dificultades de la adaptación de los test a otros idiomas y culturas (p. ej., es habitual que las versiones adaptadas tengan índices de fiabilidad inferiores a las versiones originales). Quizás los criterios propuestos por Prieto y Muñiz (2000) para evaluar la calidad de los test utilizados en España podrían servir como directrices complementarias. Estos criterios especifican que, para los índices de equivalencia (fiabilidad de formas paralelas), serían adecuados coeficientes de correlación iguales o mayores que 0,60 (buenos: $0,70 \leq r < 0,80$, y excelentes: $r \geq 0,80$), para los índices de fiabilidad de consistencia interna (p. ej., coeficiente *alfa* de Cronbach y similares), valores iguales o mayores que 0,70 (buenos: $0,80 \leq alfa < 0,85$, y excelentes: $alfa \geq 0,85$) y, para los índices de estabilidad temporal (fiabilidad test-retest), coeficientes de correlación iguales o mayores que 0,65 (buenos: $0,75 \leq r < 0,80$, y excelentes: $r \geq 0,80$).

Respecto a los índices de estabilidad temporal es importante tener en cuenta el tiempo transcurrido entre la primera (test) y segunda aplicación (retest) del instrumento así como el marco temporal de sus instrucciones, ya que, por ejemplo, muchos síndromes y trastornos psicológicos son episódicos y, por tanto, se espera que fluctúen con el tiempo, máxime si hay por medio algún tipo de tratamiento psicológico o biomédico que pretenda eliminar o reducir su presencia. Por tanto, no hay que tener en cuenta los índices test-retest obtenidos tras la aplicación de un tratamiento psicológico o biomédico y hay que tomar con mucha precaución los índices obtenidos con períodos test-retest muy largos en relación con las instrucciones del instrumento, en especial cuando se han obtenido con muestras clínicas de participantes.

Finalmente, respecto a los instrumentos para los cuales uno de los aspectos más importantes de la fiabilidad es la fiabilidad interjueces (p. ej., escalas, cuestionarios e inventarios heteroaplicados, entrevistas diagnósticas), los criterios más consensuados (véase Cicchetti, 1994) indican que índices de acuerdo interjueces (p. ej., kappa, coeficiente de correlación intraclase) iguales o mayores que 0,40 serían adecuados (buenos: $0{,}60 \leq \kappa < 0{,}75$, y excelentes: $\kappa \geq 0{,}75$).

3. *"El test debería ser relevante para la cuestión legal o para un constructo psicológico que subyazca tras la cuestión legal. Cuando sea posible, esta relevancia debería estar apoyada en la existencia de investigación de validación publicada en revistas con revisión por pares"* (Heilbrun, 1992, p. 265).

Afortunadamente, existe un número cada vez mayor de revistas especializadas en cuestiones prácticas y de investigación relevantes al ámbito de la psicología clínica forense, incluyendo, entre otras, *American Journal of Forensics Psychology*, *Behavioral Sciences and Law*, *Criminal Justice and Behavior*, *Expert Evidence: The International Digest of Human Behaviour, Science and Law*, *Journal of Forensic Psychology Practice*, *Journal of Forensic Psychiatry & Psychology*, *Law and Human Behavior*, *Legal and Criminological Psychology*, *Psychological Injury and Law*, *Psychology, Public Policy, and Law*, y, en España, *Anuario de Psicología Jurídica*, *European Journal of Psychology Applied to Legal Context* y *Psicopatología Clínica, Legal y Forense*, aunque muchas de las revistas punteras en psicología clínica también publican de manera regular artículos sobre cuestiones forenses. De hecho, la mayoría de los estudios que avalan si un instrumento mide de forma válida un constructo psicológico que subyace tras una cuestión legal se encuentran en revistas dedicadas a la evaluación psicológica en general o a la evaluación en psicología clínica.

En todas esas revistas, pues, se debería buscar información sobre la validación de los test potencialmente útiles, pero para concretar la directriz en cuanto a qué aspectos de la validez examinar y cómo interpretar los datos al respecto, la directriz tercera se puede traducir en cinco requisitos: 1) si el instrumento es relevante para evaluar cuestiones legales o forenses o para evaluar un constructo psicológico que subyazca tras cuestiones legales o forenses, y si existen datos que apoyen dicha relevancia en cuanto a 2) la validez de contenido, 3) la validez convergente, 4) la validez factorial y 5) la validez de criterio. En este sentido, para valorar la validez convergente se pueden utilizar los criterios ya mencionados de Prieto y Muñiz (2000) y evaluar los coeficientes de correlación de un test con otros instrumentos que miden el mismo constructo.

En concreto, los criterios de Prieto y Muñiz (2000) especifican que coeficientes de correlación iguales o mayores que 0,40 se podrían considerar índices adecuados de validez convergente (buenos: $0,50 \leq r < 0,60$, y excelentes: $r \geq 0,60$). La valoración de la validez convergente debería realizarse, además, en el contexto de los resultados que indiquen cuál es la validez discriminante de las puntuaciones de un test, es decir, en qué medida dicho test, en población española, muestra: (a) coeficientes de correlación nulos o pequeños con instrumentos que miden otros constructos diferentes (p. ej., $r < 0,30$, ya que este valor es considerado una correlación media según los estándares de Cohen, 1988), (b) correlaciones con instrumentos que miden otros constructos que son más bajas que las correlaciones que presenta con instrumentos que miden el mismo constructo, o (c) soluciones factoriales en las que sus ítems saturan en un factor (o factores) distinto al factor (o factores) en el que lo hacen los ítems de los instrumentos que miden otros constructos diferentes.

Para valorar la validez factorial, la literatura científica sobre la adaptación o validación en España de un determinado test debería indicar que los análisis factoriales demuestran, de manera relativamente consistente, que los ítems o escalas del test forman los factores que la teoría que subyace al test hipotetiza, es decir, que suponía *a priori*. Por ejemplo, el modelo de los cinco factores de la personalidad hipotetiza que la estructura de los rasgos de personalidad se resume en cinco dimensiones de personalidad independientes (neuroticismo, extraversión, apertura a la experiencia, amabilidad y responsabilidad), por lo que un test que mida de manera válida el modelo de los cinco factores debería mostrar una solución factorial de cinco factores ortogonales o sólo ligeramente correlacionados.

Finalmente, para valorar la validez de criterio habría que tener en cuenta los estudios publicados en revistas científicas con revisión por pares que corroboren que, en población española, el test psicológico predice de forma concurrente, predictiva o retrospectiva, criterios relevantes desde el punto de vista clínico, legal o forense (p. ej., conductas violentas, reincidencia criminal, intentos de suicidio, bajas laborales, diagnóstico psicopatológico). En este sentido, Prieto y Muñiz (2000) proponen que correlaciones del test con el criterio mayores o iguales que 0,20 se podrían considerar suficientes como índices de validez de criterio (buenos: $0,35 \leq r < 0,45$; muy buenos: $0,45 \leq r < 0,55$, y excelentes: $r \geq 0,55$). Una forma de examinar la validez de criterio muy habitual en el área clínica es la realización de estudios de validez de grupos contrastados, es decir, examinar en qué medida las puntuaciones de un test permiten diferenciar a grupos de personas que cabría esperar se diferenciaran en cuanto a su nivel en un constructo determinado (p. ej., evaluar si en un test de sintomatología ansiosa las personas con un diagnóstico de trastorno de ansiedad se diferencian de las personas sin ningún diagnóstico de trastorno mental o con

un diagnóstico de otro trastorno mental que no suele estar acompañado de sintomatología ansiosa). Por supuesto, en la valoración de los estudios de validez de criterio y de validez de grupos contrastados es muy importante asegurarse de que los criterios relevantes o los criterios de formación de los grupos son en sí mismos fiables y válidos, puesto que si no cualquier resultado que se obtuviera no indicaría nada sobre las propiedades psicométricas del test evaluado, sino sobre la baja fiabilidad o la baja validez de la propia medida del criterio.

Aplicación

4. *"Debería utilizarse una aplicación estandarizada, con unas condiciones de aplicación del test tan cercanas como sea posible al ideal de tranquilidad y ausencia de distracciones"* (Heilbrun, 1992, p. 266).

La estandarización de las condiciones de aplicación de un test es un factor importante que afecta a la fiabilidad y validez de la información que se obtenga del mismo. Este es un aspecto que hay que tener en muy cuenta en los test psicológicos que están basados en la realización previa de una entrevista con la persona evaluada como, por ejemplo, en las escalas, inventarios y cuestionarios heteroaplicados diseñados para ser aplicados por clínicos o personal entrenado no especializado. Un caso paradigmático al respecto, de nuevo, es la HRDS. Como todos los instrumentos heteroaplicados, la HDRS no la completa el paciente, sino que requiere que otra persona, en este caso, un clínico entrenado, la conteste después de haber mantenido con el paciente una entrevista. En este sentido, Hamilton (1960, p. 56) afirmaba que la HDRS, "se usa para cuantificar los resultados de una entrevista, y su valor depende enteramente de la habilidad del entrevistador de obtener la información necesaria. El entrevistador puede, y debería, usar toda la información disponible que le ayude con su entrevista y con la realización de la evaluación final". Sin embargo, la HDRS no tiene instrucciones estandarizadas sobre el formato, la duración o el contenido de la entrevista que debe preceder al cumplimiento de la escala por parte del clínico, lo cual puede afectar de manera importante a la fiabilidad y validez del instrumento en la práctica clínica forense en función de la variabilidad de esos parámetros. Por esa razón, y principalmente desde los años 80 del siglo pasado, se han propuesto diversas opciones para estandarizar la entrevista previa (véase una revisión en Williams, 2001). Una de esas opciones es la creación de guías de entrevista estructurada (p. ej., Endicott, Cohen, Nee, Fleiss y Sarantakos, 1981; Klerman, Weissman, Rounsaville y Chevron, 1984; Williams, 1988; Williams et al., 2008), algunas con el objetivo adicional de que la estructuración facilite la cumplimentación de la HDRS por entrevistadores entrenados no especializados (p. ej., Potts, Daniels, Burnam y Wells, 1990;

Whisman, Strosahl, Fruzzetti, Schmaling, Jacobson y Miller, 1989). En la misma línea, Miller, Bishop, Norman y Maddever (1985) han creado una versión modificada de la HDRS que, además de incrementar el número de ítems para incluir algunos síntomas cognitivos que permitan superar las limitaciones de contenido de la versión original, también han modificado las descripciones de los ítems para hacerlos más específicos y, por tanto, más fáciles de valorar, y han añadido preguntas concretas a las definiciones de los ítems con el mismo objetivo de facilitar que personal auxiliar debidamente entrenado pueda completar la escala (se puede encontrar una traducción al español de esta versión modificada de la HDRS en Comeche, Díaz y Vallejo, 1995). Por otro lado, algunas versiones de la HDRS establecen el marco temporal de referencia de la entrevista en la situación del paciente durante la última semana (p. ej., Williams, 1988), mientras que otras versiones lo establecen claramente en el momento actual (p. ej., la adaptación española que se recoge en Bobes et al., 2003) y, por lo tanto, el marco temporal de las instrucciones debería ser otros aspecto que debería estandarizarse de la HDRS.

De hecho, la estandarización es un aspecto especialmente importante si a partir de esa escala, inventario o cuestionario heteroaplicado se quiere llegar a un diagnóstico. Existen dos fuentes principales de variabilidad en una entrevista diagnóstica: la varianza del criterio y la varianza de la información. La varianza del criterio puede definirse como "las variaciones entre clínicos al aplicar estándares sobre lo que es clínicamente relevante… y cuando se cumplen los criterios diagnósticos" (Rogers, 2001, p. 5). Los sistemas de clasificación diagnóstica como el DSM-IV y la CIE-10, al proponer criterios diagnósticos explícitos, han reducido de forma importante la varianza del criterio, pero lamentablemente no han eliminado el problema (Rogers, 1995). Por ejemplo, Blashfield (1992, citado en Rogers, 1995) encontró que los clínicos que utilizan entrevistas no estructuradas o clínicas no aplican sistemáticamente criterios diagnósticos, lo que da lugar a errores diagnósticos en el 60% de las ocasiones. Este resultado no debería sorprender dado que, por ejemplo, el DSM-IV incluye 365 categorías diagnósticas y, si de manera conservadora se estima que cada categoría tiene como media 6 criterios diagnósticos (que es la media de criterios, por ejemplo, de los 18 categorías de los trastornos del estado de ánimo), es fácil suponer los problemas que cualquier clínico puede tener para recordar y tener presente de forma fiable durante una entrevista ¡2190 criterios diagnósticos!

La varianza de la información puede definirse como "las variaciones entre clínicos en cuanto a qué preguntas se plantean, que observaciones se hacen y cómo se organiza la información resultante" (Rogers, 2001, p. 5). Por ejemplo, varios estudios han demostrado que los entrevistadores están sujetos a sesgos confirmatorios, es decir, tienden a formularse una hipótesis antes de haber recogido todos los datos relevante y buscan selectivamente información que

confirme dicha hipótesis, ignorando cualquier otra que la refute y pasando por alto síntomas importantes (Rogers, 2001). En la misma línea, los clínicos tienden a parar la entrevista después de haber identificado el primer trastorno mental, de manera que pueden pasar por alto muchos diagnósticos, particularmente si son raros (Rogers, 2001). Todos estos sesgos conducen a que diferentes entrevistadores tengan diferentes tipos y cantidades de información, lo cual, obviamente, da lugar a diferentes diagnósticos.

En definitiva, los test basados en la realización previa de una entrevista con la persona evaluada, tanto si sus objetivos son diagnósticos como si son medir la gravedad de un síndrome o trastorno, deberían contar con una entrevista estructurada o semiestructurada que garantice la estandarización de las condiciones de aplicación del test y, en consecuencia, la fiabilidad y validez del test.

Interpretación

5. *"Tanto la selección de un test como su interpretación deberían guiarse por la aplicabilidad a una población concreta y para un propósito dado. Los resultados de un test (distintos del comportamiento observado durante su administración) no deberían aplicarse a un propósito para el cual el test no fue desarrollado (p. ej., inferir psicopatología a partir de los resultados de un test de inteligencia). La especificidad de la población y de la situación deberían guiar la interpretación. Cuanto mayor sea el "ajuste" entre un individuo dado y la población y situación utilizadas en la investigación de validación, más confianza se puede tener en la aplicabilidad de los resultados"* (Heilbrun, 1992, p. 266).

Aunque hay test que han sido evaluados con muestras de muy distintas poblaciones y en muy distintas situaciones (p. ej., el MMPI, el MCMI), esto no es necesariamente así respecto a sus adaptaciones españolas. Es necesario, pues, analizar en profundidad las muestras de estandarización con las que se desarrolló el test así como las muestras que posteriormente han sido evaluadas con dicho instrumento en la literatura científica para así poder determinar en qué medida la persona evaluada se ajusta a dichas muestras y poder matizar adecuadamente cualquier conclusión sobre los resultados que obtenga dicha persona en ese test en cuestión.

6. *"Los test objetivos y los datos actuariales son preferibles cuando hay datos apropiados de resultado y existe una "fórmula""* (Heilbrun, 1992, p. 267).

A la hora de hacer una predicción sobre un resultado o condición relevante en el ámbito clínico legal y forense (p. ej., la reincidencia en violencia de pareja, abusos sexuales o delitos violentos, la comisión de suicidio, etc.), un psicólogo puede utilizar su experiencia, sensibilidad interpersonal o perspectiva teórica para recordar, sintetizar e interpretar las características y circunstancias de una persona. A esta aproximación a la predicción de resultados basada en la utilización de procesos informales o intuitivos para combinar o integrar los datos de la persona evaluada se le suele denominar predicción clínica. Frente a esta aproximación existe la denominada predicción estadística o actuarial, basada en las relaciones empíricas establecidas entre las características y circunstancias de una muestra de la población relevante y la condición o resultado que se quiere predecir. Esta predicción actuarial o estadística se traduce en la utilización de fórmulas, ecuaciones, tablas actuariales, algoritmos o gráficos de decisión en los que se introduce la información sobre la persona evaluada y se integra con la información empírica de las muestras de la población relevante para llegar a una predicción. Desde el clásico libro de Meehl (1954/1996), los resultados de numerosas revisiones y metaanálisis realizados en los últimos 60 años son coincidentes en señalar la superioridad de los métodos estadísticos o actuariales frente a los métodos clínicos a la hora de hacer predicciones más exactas en una gran variedad de contextos clínicos, legales y forenses (Ægisdóttir et al., 2006; Bonta, Law y Hanson, 1998; Dawes, Faust y Meehl, 1989; Grove y Meehl, 1996; Grove, Zale, Lebow, Snitz, y Nelson, 2000; Hanson y Bussière, 1998; Hilton y Harris, 2005; Hilton, Harris y Rice, 2006). Sin embargo, la utilización de la aproximación actuarial o estadística presupone que ya se han obtenido datos en personas similares a la persona evaluada, que se han medido de forma sistemática los resultados o las condiciones de interés, que se han identificado adecuadamente las variables de predicción, que tales variables han sido ponderadas de manera óptima y que la ecuación, fórmula o algoritmo resultante ha sido validado en nuevas muestras de replicación. Desgraciadamente, la ausencia de fórmulas, ecuaciones o algoritmos de predicción validados en muestras de población española es más la norma que la excepción para la mayoría de los test psicológicos relevantes para el ámbito clínico forense. Como afirmaba Heilbrun (1992, p. 267), parafraseando a Meehl (1957), "si no existe ninguna "fórmula", entonces no tenemos más alternativa que usar nuestras cabezas".

7. *"Se debería evaluar explícitamente el estilo de respuesta usando aproximaciones sensibles a la distorsión, y se deberían interpretar los resultados de la aplicación del test dentro del contexto del estilo de respuesta del individuo. Cuando el estilo de respuesta parezca ser de simulación, defensivo o irrelevante en lugar de sincero/fiable, quizás sea*

> *necesario minimizar la importancia de los resultados de la aplicación del test o incluso ignorarlos y enfatizar en mayor medida otras fuentes de datos"* (Heilbrun, 1992, p. 267).

Es obvio que los estilos de respuesta pueden afectar a la validez de las interpretaciones que se pueden extraer de las puntuaciones de un test, que en el contexto legal y forense la presencia de estilos de simulación ("producción intencionada de síntomas físicos o psicológicos desproporcionados o falsos, motivados por incentivos externos", APA, 1994/1995, p. 698) defensivos (negación deliberada o minimización flagrante de síntomas físicos o psicológicos) e irrelevantes (la persona evaluada no llega a estar psicológicamente comprometida con el proceso de evaluación y, por ejemplo, no se esfuerza en responder con exactitud a las preguntas que se le plantean) son habituales y que todos los test de autoinforme (p. ej., escalas, cuestionarios e inventarios autoaplicados) son especialmente susceptibles a los efectos de dichos estilos de respuesta (Heilbrun, Bennett, White y Kelly, 1990; Otto, 2008; Rogers, 2008). Por tanto, es más que evidente la necesidad de evaluar el estilo de respuesta de la persona evaluada y de validar la información que nos ofrecen los test de autoinforme con la información proporcionada por terceras personas (p. ej., parejas, familiares, amigos, vecinos, compañeros de trabajo, personal médico) o por fuentes documentales (p. ej., historias clínicas, informes de otros profesionales, sentencias, expedientes penitenciarios). En el campo de la evaluación clínica forense, muchos test incluyen escalas adicionales para evaluar el estilo de respuesta de la persona evaluada (p. ej., las escalas de validez, en sus diferentes versiones, del MMPI, del MCMI o del PAI), pero la mayoría no, por lo que, en el contexto legal y forense, sería necesario utilizar de manera complementaria algún otro instrumento que cubra específicamente dicha evaluación como, por ejemplo, en España, la escala de Deseabilidad Social de Marlowe y Crowne (MCSDS; Crowne y Marlowe,1960), que permite evaluar los estilos defensivo y de simulación y de la que existen dos adaptaciones españolas (Ávila y Tomé, 1989; Ferrando y Chico, 2000), y el Inventario Estructurado de Simulación de Síntomas (SIMS; Widows y Smith, 2005, 2009; véase también González-Ordi, Santamaría-Fernández y Fernández-Martín, 2010), el cual incluye, además de una escala global, cinco escalas denominadas Psicosis, Deterioro neurológico, Trastornos amnésicos, Baja inteligencia y Trastornos afectivos desarrollada específicamente para evaluar el grado en que una persona informa de síntomas atípicos de tales trastornos y, por tanto, los puede estar simulando. A la hora de detectar la simulación también puede ser útil la adaptación española del Test de Simulación de Problemas de Memoria (TOMM; Tombaugh, 1996, 2011), ya que, no sólo es útil para evaluar el engaño o la exageración de los problemas de memoria en los trastornos neurológicos o

mentales con base orgánica, sino que también puede ser útil en los trastornos mentales funcionales. Por ejemplo, diversos estudios han encontrado que los pacientes depresivos, incluso los más graves, presentan en el TOMM un rendimiento similar al de las personas sin trastornos psicológicos (p. ej., Ashendorf, Constantinou y McCaffrey, 2004; Rees, Tombaugh y Boulay, 2001; Yanez, Fremouw, Tennant, Strunk y Coker, 2006). Por tanto, un rendimiento bajo en el TOMM (p. ej., por debajo del punto de corte de 45) en una persona que "presenta" sintomatología depresiva constituiría un indicio de simulación de un trastorno depresivo (Rees et al., 2001). Esta ausencia de diferencias entre pacientes depresivos y personas sin depresión en un test de memoria parecería que contradice los datos que demuestran problemas de rendimiento cognitivo en algunos pacientes con depresión, especialmente con depresiones graves; sin embargo, es importante recordar que tales problemas afectan sobre todo a las tareas que implican procesos cognitivos controlados, es decir, procesos que requieren gran cantidad de atención, esfuerzo y recursos de procesamiento, como, por ejemplo, tareas de recuerdo libre, mientras que apenas afectan a las tareas que implican procesos cognitivos automáticos (véase la revisión de Hartlage, Alloy, Vázquez y Dyckman, 1993), y precisamente el TOMM es una tarea de memoria de reconocimiento visual que evalúa un procesamiento más bien automático.

Por otro lado, existen varias estrategias para tratar de reducir o anular los efectos del estilo de respuesta. Por ejemplo, una de las más eficaces consiste en confeccionar baremos para los instrumentos a partir de las puntuaciones obtenidas en contextos iguales a aquellos en los que se sospecha puede haber un interés directo en simular síntomas depresivos o en negarlos o minimizarlos (Salgado, 2005). Así, se podrían desarrollar unos baremos específicos con muestras de personas que presentan demandas por daño personal para conseguir compensaciones económicas (contexto en el que cabría esperar simulación de síntomas psicopatológicos) y otros baremos distintos con muestras de personas que litigan para conseguir la guardia y custodia de sus hijos (contexto en el que cabría esperar una negación o minimización de síntomas, máxime cuando del otorgamiento de la guarda y custodia depende la atribución de la vivienda familiar y la designación del deudor de la pensión de alimentos). Los baremos así construidos incluirían ya una parte de la puntuación normativa que corresponde a la simulación o al estilo defensivo y que es común a todas las personas evaluadas en esa situación específica, y, por lo tanto, los efectos del estilo de respuesta habrían sido ya parcialmente neutralizados cuando se utilizaran para interpretar las puntuaciones de una persona evaluada en esas mismas situaciones.

Otra estrategia para reducir los efectos del estilo de respuesta puede ser modificar el formato de los instrumentos. Por ejemplo, respecto al BDI, en el

que las alternativas de respuesta para cada síntoma están ordenadas de menor a mayor gravedad y, además, están acompañadas del número de 0 a 3 que indica su puntuación, algunos estudios sugieren que la presentación aleatoria de las distintas afirmaciones de gravedad dentro de cada ítem y la eliminación del número que indica su puntuación puede tener la ventaja de romper sesgos de respuestas tendentes a escoger la primera afirmación o la última (Dahlstrom, Brooks y Peterson, 1990) y, por tanto, o bien asegura que las personas evaluadas prestan atención a todas las afirmaciones que componen cada ítem del BDI, lo que permite obtener un rango mejor de puntuaciones, o bien dificulta los intentos de simulación o minimización de síntomas, ya que la gravedad de algunas alternativas no es tan obvia si no aparece ordenada y con su puntuación correspondiente. Así, como sugieren Echeburúa, Amor y Corral (2003), si en el ítem 10 sobre llanto del BDI-IA, la alternativa de mayor gravedad ("Antes era capaz de llorar, pero ahora no puedo incluso aunque quiera") no se sitúa en último lugar acompañada de la máxima puntuación ("3"), sino en otro lugar y sin dicha puntuación, es muy probable que muchas personas que tratan de simular depresión no la elijan y, en cambio, escojan "Lloro continuamente" que, en realidad, es menos grave y se puntúa con un 2 en lugar de 3.

Aplicación de las directrices: requisitos de un test psicológico para evaluar la gravedad de la sintomatología depresiva

A partir de las directrices de Heilbrun (1992) ampliadas y especificadas en la sección anterior, es posible destilar y analizar una serie de requisitos que debería cumplir un instrumento para la evaluación psicológica en el ámbito clínico forense y analizar en qué medida los cumplen un conjunto de test concreto que sea inicialmente relevante para los objetivos de una evaluación forense dada.

Como ejemplo, en la Tabla 1 se presenta el resultado de un análisis realizado a partir de esas directrices y requisitos sobre los test más utilizados para la evaluación de la presencia y gravedad de la sintomatología depresiva en su aplicación a la población clínica de España.

Los test analizados son 11 de las escalas, inventarios y cuestionarios autoaplicados más utilizados en España para ese fin según la revisión bibliográfica de Sanz et al. (2013), y a los que se ha unido la escala heteroaplicada HDRS dada su amplia utilización para evaluar la gravedad de la depresión y los cambios producidos por los tratamientos antidepresivos, especialmente los farmacológicos (Nezu et al., 2009; Trajković et al., 2011). Para el análisis de todos estos test, se ha tenido en cuenta tanto la información que se recoge en los manuales de sus respectivas adaptaciones españolas o, en el

caso de que no existan tales manuales, en los artículos de revistas científicas en las que presentan dichas adaptaciones, como toda la literatura científica localizada sobre sus propiedades psicométricas en población adulta clínica española y, secundariamente, en la población general adulta española, todo lo cual se resume en la Tabla 1.

Tabla 1. Referencias bibliográficas consultadas para obtener los datos de fiabilidad y validez de los test más utilizados para evaluar la presencia y gravedad de sintomatología depresiva en su adaptación a la población española

Test	**Referencias bibliográficas**
Inventario de Depresión de Beck, versión corregida o de 1978 (BDI-IA)	Alonso Suárez y Florit Robles (2002); Vázquez y Sanz (1997, 1999)
Inventario de Depresión de Beck, segunda edición (BDI-II)	Beck et al. (2011); Ibáñez et al. (2011); Sanz y García-Vera (2013); Sanz et al. (2005); Sanz et al. (en prensa); Sanz et al. (2003)
Escala Autoaplicada de Depresión de Zung (SDS)	Aragonés et al. (2001); Conde et al. (1970a,b); Conde López y Esteban Chamorro (1973, 1974, 1975a,b); Ramos Brieva (1986); Ramos Brieva et al. (1991); Romera et al. (2008)
Escala de Depresión del Centro de Estudios Epidemiológicos (CES-D)	Losada et al. (2012); Ros et al. (2011); Soler et al. (1997); Vázquez et al. (2007)
Escalas de Depresión Mayor (CC) y Distimia (D) del Inventario Clínico Multiaxial de Millon, segunda edición (MCMI-II)	Borda Más et al. (2008); Millon (1998, 1999, 2002); Pedrero Pérez et al. (2012); Pedrero Pérez et al. (2005)
Escalas de Depresión Mayor (CC) y Distimia (D) del Inventario Clínico Multiaxial de Millon, tercera edición (MCMI-III)	Millon et al. (2007); Ortiz-Tallo et al. (2011)
Escala Clínica de Depresión (D) del Inventario Multifásico de Personalidad de Minnesota (MMPI)	Baillés et al. (2004); Grassot Esteba y Llinàs Reglà (1997); Hathaway y McKinley (1975); Seisdedos Cubero (1980)
Escalas Clínica de Depresión (D) y de Depresión basada en el contenido (DEP) del Inventario Multifásico de Personalidad de Minnesota, segunda edición (MMPI-2)	Hathaway y McKinley (1999); Rodríguez Pulido et al. (2008)
Escala de Valoración de la Depresión de Hamilton (HDRS)	Baca García et al. (1998); Ballesteros et al. (2007); Cordero Villafáfila y Ramos-Brieva (1986); Guillén et al. (2012); Ramos-Brieva y Cordero Villafáfila (1986a, 1986b, 1988); Ramos Brieva et al. (1994)

Los resultados de dicho análisis se presentan en la Tabla 2. Por ejemplo, en este caso, todos los test parecen relevantes para evaluar un constructo psicológico, la depresión clínica, que subyace tras muchas cuestiones legales y forenses de interés (p. ej., la predicción del riesgo de suicidio, el esclarecimiento de si una muerte ha sido accidental o ha sido un suicidio, la valoración de los eximentes de trastorno mental), y en lo que difieren, por ejemplo, es en la existencia de un mayor o menor número de estudios sobre su validez para medir la depresión clínica en población española y publicados en revistas científicas con revisión por pares, y en el mayor o menor apoyo empírico a esa validez que se obtiene de dichos estudios.

Como se ha comentado antes, el análisis está basado fundamentalmente en las referencias que se recogen en la Tabla 1, pero en algunos casos se han utilizado los resultados y conclusiones de trabajos que a su vez analizaban y comparaban los test de la Tabla 2. En concreto, para valorar el uso frecuente y la validez contenido se han tenido en cuenta los resultados del estudio de Sanz et al. (2013). En este estudio se realizó una búsqueda bibliográfica de los cuestionarios, escalas e inventarios autoaplicados más populares en España para evaluar la depresión clínica en adultos y, posteriormente, se analizaron dichos test desde una doble perspectiva: (a) respecto a si el contenido de sus ítems era apropiado para medir los síntomas de las definiciones de episodio depresivo y distimia del DSM-IV y la CIE-10 (relevancia) y (b) respecto a si sus ítems eran proporcionales a los síntomas de tales definiciones (representatividad). Los resultados de ese estudio se utilizaron para realizar las valoraciones que se presentan en la Tabla 2 sobre el criterio de validez de contenido y, en el caso de la HDRS, que dado que es un test heteroaplicado no fue analizado en el estudio de Sanz et al. (2013), se analizó su validez de contenido en función de los mismos criterios y perspectivas que se utilizaron en el estudio de Sanz et al. (2013)[3].

[3] En función de los criterios y parámetros analizados por Sanz et al. (2013) y fijándose en el síntoma principal que pretende medir cada ítem de la HDRS, ya que muchos sus ítems evalúan varios síntomas a la vez, la HDRS evalúa, teniendo en cuenta los criterios diagnósticos del DSM-IV, el 88,9% de los criterios sintomáticos de la depresión mayor, el 57,1% del trastorno distímico y el 42,8% de los síntomas atípicos o especificaciones de los trastornos depresivos, mientras que, teniendo en cuenta los criterios diagnósticos de la CIE-10, la HDRS evalúa el 80% de los criterios sintomáticos de la depresión mayor, el 50% de la distimia y el 85,7% de los síntomas atípicos o especificaciones de los trastornos depresivos. Además, de los 17 ítems de la HDRS, el 82,3% de esos ítems miden síntomas depresivos según los criterios diagnósticos del DSM-IV o la CIE-10. Finalmente, en lo que respecta a la representatividad del constructo de depresión clínica, el 5,9% de los 17 ítems de la HDRS evalúan síntomas anímicos, el 11,8% síntomas motivacionales, el 35,3% síntomas físicos, el 11,8% síntomas motores, el 11,8% síntomas cognitivos, el 5,9% síntomas atípicos o de especificaciones y el 17,6% otros síntomas no depresivos (véase Sanz et al., 2013, para los porcentajes de criterios diagnósticos DSM-IV o CIE-10 que especifican los distintos tipos de síntomas depresivos).

Tabla 2. Comparación en cuanto al cumplimiento de unos requisitos mínimos para su utilización en el ámbito clínico forense de los test psicológicos más utilizados para evaluar la gravedad de la depresión en su adaptación a la población española

Requisitos mínimos	BDI-IA	BDI-II	SDS	CES-D	MCMI-II/ CC y D	MCMI-III/ CC y D	MMPI/D	MMPI-2/D y DEP	HDRS
Comercializado	No	Si	No	No	Si	Si	Si	Si	No
Manual técnico	No*	Si	No*	No*	Si	Si	Si	Si	No*
Uso frecuente	+++	+++	++	++	+++	+++	+++	+++	+++
Fiable (consistencia interna)	+++	+++	++	+++	+++	+++	S.d.	S.d.	+
Fiable (test-retest)	+	S.d.	++	S.d.	S.d.	S.d.	S.d.	S.d.	+++
Fiable (acuerdo interjueces)	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	+++
Relevancia legal/forense	Si	Si	Si	Si	Si	Si	Si	Si	Si
Válido (contenido)	+++	+++	+++	+++	++	++	+	+	++
Válido (convergente)	+++	+++	+++	+++	+++	+++	++	++	+++
Válido (factorial)	++	+++	+	++	SD	SD	SD	SD	+
Valido (criterio)	++	++	++	++	+	++	SD	SD	++
Aplicación estandarizada	+++	+++	+++	+++	+++	+++	+++	+++	+
Corrección estandarizada	+++	+++	+++	+++	+++	+++	+++	+++	+
Baremos adecuados	+	++	+	–	+	+	–	+	++
Predicción actuarial	S.d.	S.d.	S.d.	S.d.	S.d.	S.d.	S.d.	S.d.	S.d.
Estilos de respuesta	No	No	No	No	++	++	+	+++	No

Nota. +++ = Excelente. ++ = Bueno. + = Adecuado. – = Pobre o con carencias. S.d. = sin datos en España. N.a. = No aplicable.

*Aunque hay publicados en revistas científicas con revisión por pares varios artículos sobre sus propiedades psicométricas en población española.

BDI-IA: Inventario de Depresión de Beck, versión corregida o de 1978 (Vázquez y Sanz, 1997). BDI-II: Inventario de Depresión de Beck, segunda edición (Beck et al., 2011). SDS = Escala Autoaplicada de Depresión de Zung (Conde et al., 1970). CES-D: Escala de Depresión del Centro de Estudios Epidemiológicos (Soler et al., 1997; Vázquez et al., 2007). MCMI-II/CC y D: escalas de Depresión Mayor (CC) y Distimia (D) del Inventario Clínico Multiaxial de Millon, segunda edición (Millon, 1998, 1999, 2002). MCMI-III/CC y D: escalas de Depresión Mayor (CC) y Distimia (D) del Inventario Clínico Multiaxial de Millon, tercera edición (Millon et al., 2007). MMPI/D: Escala Clínica de Depresión del Inventario Multifásico de Personalidad de Minnesota (Hathaway y McKinley, 1975). MMPI-2/D y DEP: Escalas Clínica de Depresión (D) y de Depresión basada en el contenido (DEP) del Inventario Multifásico de Personalidad de Minnesota, segunda edición (Hathaway y McKinley, 1999). HDRS: Escala de Valoración de la Depresión de Hamilton (Ramos-Brieva y Cordero Villafáfila, 1986a).

En este sentido, entre los test de la Tabla 2 destacan, en cuanto a su validez de contenido, las diferentes versiones del BDI (BDI-IA y BDI-II), la SDS y la CES-D, ya que sus ítems presentan un mayor grado de relevancia y representatividad del contenido de las definiciones sintomáticas de la depresión clínica del DSM-IV y la CIE-10, mientras que, por el contrario, se aprecian deficiencias y limitaciones, desde la perspectiva de la validez de contenido, en la escala de Distimia (D) del MCMI-III, la escala de Depresión del SCL-90-R y las escalas clínicas de Depresión (D) del MMPI y MMPI-2 (Sanz et al., 2013).

Conclusiones

El considerable desarrollo de la evaluación psicológica en España ha traído consigo un enorme aumento del número de test psicológicos de los que puede disponer un psicólogo para realizar una evaluación clínica forense, lo que implica que una de las primeras preguntas que se debe plantear dicho psicólogo en su actividad profesional e investigadora es qué test utilizar. Como cabría suponer, no existe una respuesta única a esta pregunta puesto que la respuesta viene determinada por los objetivos de la evaluación, estos objetivos pueden ser muy diversos y ningún test por sí solo parece cubrir con garantías todos ellos, aunque algunos test pueden ser útiles para conseguir de forma aceptable uno o varios.

En consecuencia, la selección de un test psicológico para la evaluación psicológica en el ámbito clínico forense implica, en primer lugar, determinar los objetivos de dicha evaluación y la población a la que pertenece la persona o personas que van a ser evaluadas, y, en segundo lugar, valorar la adecuación a esos objetivos y a esa población de las características del test y de las interpretaciones o inferencias de las medidas que ese test proporciona. Para hacer estas valoraciones es necesario tener en cuenta simultáneamente varios criterios tanto psicométricos y prácticos como de relevancia para las cuestiones forenses y, en este sentido, Heilbrun (1992) ha propuesto siete directrices sobre la disponibilidad y documentación del test, sus propiedades psicométricas y su interpretación. En este trabajo, se han actualizado, ampliado, especificado y ejemplificado esas directrices con el objetivo último de que las mismas puedan ser de ayuda a los psicólogos españoles que trabajan en el ámbito clínico forense para determinar si un test psicológico dado debería usarse en una evaluación forense en concreto.

Dada la complejidad y responsabilidad de la actividad que desempeñan los psicólogos forenses, en los últimos años diversas instituciones en España

como el Colegio Oficial de Psicólogos de Madrid o la Consejería de Justicia e Interior de la Comunidad de Madrid han promovido la creación de guías de buenas prácticas profesionales bien generales o bien centradas en los diversos ámbitos de actuación del psicólogo forense para así: (a) establecer criterios de calidad que puedan servir de referente no sólo a los profesionales, sino también a los usuarios, las administraciones públicas y la sociedad en general, y (b) orientar al profesional en una práctica basada en los mejores datos y conocimientos científicos disponibles en cada momento y en la aplicación de las normas éticas y legales exigibles en cada caso (Bartolomé Tutor et al., 2013; Chacón Fuertes et al., 2009; Consejería de Justicia e Interior de la Comunidad de Madrid, 2012; Gómez Hermoso et al., 2012). Por ejemplo, el Colegio Oficial de Psicólogos de Madrid ha publicado hasta la fecha la "Guía de buenas prácticas para la elaboración de informes psicológicos periciales sobre custodia y régimen de visitas de menores" (Chacón Fuertes et al., 2009), la "Guía de buenas prácticas para la evaluación psicológica forense del riesgo de violencia contra la mujer en las relaciones de pareja (VCMP)" (Gómez Hermoso et al., 2012) y la "Guía de buenas prácticas para la elaboración de informes psicológicos periciales sobre custodia y régimen de visitas de menores adaptada a casos de violencia de género" (Bartolomé Tutor et al., 2013). En este contexto hay que enmarcar también la necesidad y utilidad del presente trabajo, en la medida en que pueda contribuir a la discusión y desarrollo de guías de buena práctica profesional para el uso de los test psicológicos en el ámbito clínico forense. De hecho, las directrices aquí presentadas podrían ayudar a los profesionales que utilicen las guías de buenas prácticas antes mencionadas a seleccionar los test psicológicos que en concreto les puedan ayudar en cada una de las tareas que se mencionan en dichas guías, es decir, la realización de informes sobre custodia y régimen de visita de menores y la evaluación del riesgo de violencia contra la mujer en las relaciones de pareja, ya que en esas guías se ofrece, acertadamente, un listado de instrumentos orientativos comúnmente utilizados para conseguir esos objetivos, pero, en cambio, no se ofrece ninguna indicación o directriz sobre cuál o cuáles son los más apropiados o cómo se podría valorar la adecuación de unos u otros instrumentos, y, claramente, dichas indicaciones o directrices parecen necesarias cuando en tales guías, por ejemplo, para valorar la personalidad se listan entre 7 y 8 test y para valorar la psicopatología o los rasgos clínicos entre 13 y 15 test (Bartolomé Tutor et al., 2013; Chacón Fuertes et al., 2009; Gómez Hermoso et al., 2012).

Referencias

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*, 341-382.

Alonso Suárez, M., y Florit Robles, A. C. (2002). Características psicométricas de la versión española del Inventario de Depresión de Beck en pacientes crónicos con esquizofrenia. *Anales de Psiquiatría, 18*, 154-160.

American Psychiatric Association (1995). *DSM-IV. Manual diagnóstico y estadístico de los trastornos mentales*. Barcelona: Masson. (Orig. 1994).

Antony, M. M., Orsillo, S. M., y Roemer, L. (Eds.). (2001). *Practitioner's guide to empirically based measures of anxiety*. Nueva York: Plenum Press.

Aragonés Benaiges, E., Masdéu Montalà, R. M., Cando Guasch, G., y Coll Borrás, G. (2001). Validez diagnóstica de la Self-Rating Depression Scale de Zung en pacientes de atención primaria. *Actas Españolas de Psiquiatría, 29*, 310-316.

Ashendorf, L., Constantinou, M., y McCaffrey, R. J. (2004). The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology, 19*, 125-130.

Ávila, A., y Tomé, M.C. (1989). Evaluación de la deseabilidad social y correlatos defensivos y emocionales. Adaptación castellana de la Escala de Crowne y Marlowe. En A. Echevarría y D. Páez (Eds.), *Emociones: Perspectivas psicosociales* (pp. 505-514). Madrid: Fundamentos.

Baca-García, E., Díaz-Sastre, C., Rico, F., y Sáiz Ruiz, J. (1998). Valoración de la fiabilidad de la evaluación clínica entre los investigadores de un ensayo clínico multicéntrico. *Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 26*, 358-362.

Baillés, E., Pintor, L., Fernandez-Egea, E., Torres, X., Matrai, S., de Pablo, J., et al. (2004). Psychiatric disorders, trauma, and MMPI profile in a Spanish sample of nonepileptic seizure patients. *General Hospital Psychiatry, 26*, 310-315.

Ballesteros, J., Bobes, J., Bulbena, A., Luque, A., Dal-Ré, R., Ibarra, N., et al. (2007). Sensitivity to change, discriminative performance, and cutoff criteria to define remission for embedded short scales of the Hamilton depression rating scale (HAMD). *Journal of Affective Disorders, 102*, 93-99.

Bartolomé Tutor, A., Chacón Fuertes, F., García Gumiel, J. F., García Moreno, A., Gómez Hermoso, M. R., Gómez Martín, R., y Vázquez Mezquita, B. (2013). *Guía de buenas prácticas para la elaboración de informes psicológicos periciales sobre custodia y régimen de visitas de menores adaptada a casos de violencia de género*. Madrid: Colegio Oficial de Psicólogos de Madrid. Disponible en: http://www.copmadrid.es./webcopm/recursos/guiadebuenaspracticascustodiamenores.pdf

Beck, A. T., Steer, R. A., y Brown, G. K. (2011a). *Manual. BDI-II. Inventario de Depresión de Beck-II* (Adaptación española: Sanz, J., y Vázquez, C.). Madrid: Pearson Educación.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., et al. (2007). The ups and downs of peer review. *Advances in Physiology Education, 31*, 145-152.

Bobes, J., Bulbuena, A., Luque, A., Dal-Ré, R., Ballesteros, J., e Ibarra, N. (2003). Evaluación psicométrica comparativa de las versiones en español de 6, 17, 21 ítems de la Escala de valoración de Hamilton para la evaluación de la depresión. *Medicina Clínica, 120*, 693-670.

Bonta, J., Law, M., y Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: a meta-analysis. *Psychological Bulletin, 123*, 123-142.

Borda Más, M., Torres Pérez, I., y del Río Sánchez, C. (2008). Distimia en anorexia nerviosa y bulimia nerviosa. *International Journal of Clinical and Health Psychology, 8*, 65-75.

Brodsky, S. L., y Smitherman, H. O. (1983). *Handbook of scales for research in crime and delinquency*. Nueva York: Plenum.

Buela-Casal, G., Sierra, J.C., Carretero-Dios, H., y de los Santos-Roig, M. (2002). Situación actual de la evaluación psicológica en lengua castellana. *Papeles del Psicólogo, 83*, 27-33.

Bulbena, A., Berrios, G., y Fernández de Larrinoa Palacios, P. (Eds.). (2000). *Medición clínica en psiquiatría y psicología*. Barcelona: Masson.

Caballo, V. E. (Ed.). (2005). *Manual para la evaluación clínica de los trastornos psicológicos. Estrategias de evaluación, problemas infantiles y trastornos de ansiedad*. Madrid: Pirámide.

Caballo, V. E. (Ed.). (2006). *Manual para la evaluación clínica de los trastornos psicológicos. Trastornos de la edad adulta e informes psicológicos*. Madrid: Pirámide.

Chacón Fuertes, F., García Gumiel, J. F., García Moreno, A., Gómez Hermoso, R., y Vázquez Mezquita, B. (2009). *Guía de buenas prácticas para la elaboración de informes psicológicos periciales sobre custodia y régimen de visitas de menores*. Madrid: Colegio Oficial de Psicólogos de Madrid. Disponible en: http://www.copmadrid.es./webcopm/recursos/guiadebuenaspracticasA4.pmd.pdf

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2ª ed*. Hillsdale, NJ: LEA.

Comeche, M. I., Díaz, M. I., y Vallejo, M. A. (1995). *Cuestionarios, inventarios y escalas. Ansiedad, depresión y habilidades sociales*. Madrid: Fundación Universidad-Empresa.

Comisión Internacional de Tests (ITC). (2000). Directrices internacionales para el uso de los tests. [Traducción de la Comisión de Tests del Colegio Oficial de Psicólogos de España]. *INFOCOP, 77*, 21-32. Consultado el 30 de abril de 2013 en: http://www.cop.es/infocop/vernumeroCOP.asp?id=1000

Conde, V., Escriba, J. A., e Izquierdo, J.A. (1970a). Evaluación estadística y adaptación castellana de la Escala Autoaplicada para la Depresión de Zung. *Archivos de Neurobiología, 33*, 185-206.

Conde, V., Escriba, J. A., e Izquierdo, J.A. (1970b). Evaluación estadística y adaptación castellana de la Escala Autoaplicada para la Depresión de Zung. *Archivos de Neurobiología, 33*, 281-303.

Conde, V., Esteban, T., y Useros, E. (1976). Revisión crítica de la adaptación castellana del Cuestionario de Beck. *Revista de Psicología General y Aplicada, 31*, 469-497.

Conde López, V., y Esteban Chamorro, T. (1973). Revisión crítica de dos adaptaciones castellanas de la "Self Rating Depresión Scale" (SDS) de Zung. *Archivos de Neurobiología, 36*, 375-392.

Conde López, V., y Esteban Chamorro, T. (1974). Contribución al estudio de la S.D.S. (Self-Rating Depression Scale) de Zung, en una muestra estratificada de población normal. *Revista de Psicología General y Aplicada, 29*, 515-554.

Conde López, V., y Esteban Chamorro, T. (1975a). Fiabilidad de la S.D.S. (Self-Rating Depression Scale) de Zung. *Revista de Psicología General y Aplicada, 30*, 903-914.

Conde López, V., y Esteban Chamorro, T. (1975b). Validez de la S.D.S. (Self-Rating Depression Scale) de Zung. Archivos de Neurobiología, 38, 225-246.

Conde, V., y Franch, J. I. (1984). *Escalas de evaluación comportamental para la cuantificación de la sintomatología psicopatológica en los trastornos angustiosos y depresivos*. Madrid: Upjohn Farmaquímica.

Consejería de Justicia e Interior de la Comunidad de Madrid (2007). *Guía orientativa de buenas prácticas de psicólogos forenses*. Madrid: Comunidad de Madrid.

Corcoran, K., y Fischer, J. (2013). *Measures for clinical practice and research. A sourcebook, 5ª ed. Volume 2: adults*. Nueva York: Oxford University Press.

Cordero Villafáfila, A., y Ramos-Brieva, J. (1986). Estructura factorial de la versión castellana de la Escala de Hamilton para la Depresión. *Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 14*, 339-342.

Crowne, D. P., y Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Dahlstrom, W. G., Brooks, J. D., y Peterson, C. D. (1990). The Beck Depression Inventory: item order and the impact of response set. *Journal of Personality Assessment, 55*, 224-233

Dawes, R. M., Faust, D., y Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Delgado López-Cózar, E., Marcos Cartagena, D., Jiménez Contreras, E., y Ruiz Pérez, R. (2013). *Índice H de las revistas españolas de Ciencias Sociales y Jurídicas según Google Scholar (2002-2011)*. EC3 (Grupo de Investigación de Evaluación de la Ciencia y de la Comunicación Científica). Informes, 4: 29 de mayo de 2013. Universidad de Granada.

Echeburúa, E., Amor, P. J., y Corral, P. de (2003). Autoinformes y entrevistas en el ámbito de la psicología clínica forense: limitaciones y nuevas perspectivas. *Análisis y Modificación de Conducta, 29*, 503-522.

Endicott, J., Cohen, J., Nee, J., Fleiss, J., y Sarantakos, S. (1981). Hamilton Depression Rating Scale: extracted from regular and change versions of the Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 38*, 98-103.

Ferrando, P. J., y Chico, E. (2000). Adaptación y análisis psicométrico de la escala de deseabilidad social de Marlowe y Crowne. *Psicothema, 12*, 383-389.

García-Portilla González, M. P., Bascarán Fernández, M. T., Sáiz Martínez, P. A., Parallada Redondo, M., Bousoño García, M., y Bobes García, J. (2011). *Banco de instrumentos básicos para la práctica de la psiquiatría clínica, 6ª ed*. Majadahonda, Madrid: Comunicación y Ediciones Sanitarias.

Gómez Hermoso, M. R., Muñoz Vicente, J. M., Vázquez Mezquita, B., Gómez Martín, R., y Mateos de la Calle, N. (2012). *Guía de buenas prácticas para la evaluación psicológica forense del riesgo de violencia contra la mujer en las relaciones de pareja (VCMP)*. Madrid: Colegio Oficial de Psicólogos de Madrid. Disponible en: http://www.copmadrid.es./webcopm/recursos/guiaviolenciacontralamujer.pdf

González-Ordi, H., Santamaría-Fernández, P., y Fernández-Martín, P. (2010). Precisión predictiva del Inventario de Simulación de Síntomas–SIMS en el contexto médico-legal. *Edupsykhé, 9*, 3-22.

Grassot Esteba, G., y Llinàs Reglà, J. (1997). Comparación de la psicopatología medida con el SCL-90-R y otros instrumentos psicométricos. *Psiquis, 18*, 43-50.

Grisso, T. (1986). *Evaluating competencies: forensic assessments and instruments*. Nueva York: Plenum Press.

Grisso, T. (2003). *Evaluating competencies: forensic assessments and instruments, 2ª ed*. Nueva York: Kluwer Academic/Plenum Press.

Grove, W. M., y Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., y Nelson, C. (2000). Clinical vs. mechanical prediction: a meta-analysis. *Psychological Assessment, 12*, 19-30.

Grupo de Investigación de Evaluación de la Ciencia y de la Comunicación Científica (2013). *IN-RECS (Índice de impacto de las Revistas Españolas de Ciencias Sociales). Psicología*. Universidad de Granada. Consultado el 11 de octubre de 2013 en: http://ec3.ugr.es/in-recs/

Grupo de Investigación de Evaluación de Publicaciones Científicas y Grupo de Investigación de Evaluación de la Ciencia y de la Comunicación Científica (2013). *RESH (Revistas Españolas de Ciencias Sociales y Humanidades). Indicadores. Psicología*. Centro de Ciencias Humanas y Sociales (CCHS)/Consejo Superior de Investigaciones Científicas y Universidad de Granada. Consultado el 11 de octubre de 2013 en: http://epuc.cchs.csic.es/resh/

Guillén, V., Santos, B., Yllá, L., Bulbena, A., Bilbao, J., Fernández, E., et al. (2012). Depressive dimensions and item response analysis of the Hamilton Depression Rating Scale–17 in eating disorders. *Comprehensive Psychiatry, 53*, 396-402.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23*, 56-62.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6*, 278-296.

Hanson, R. K., y Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*, 348-362.

Hartlage, S., Alloy, L., Vázquez, C., y Dyckman, B. (1993). Automatic and effortful processing in depression. *Psychological Bulletin, 113*, 247-278.

Hathaway, S. R., y McKinley, J. C. (1975). *Cuestionario de personalidad MMPI. Manual* (Adaptación española: Departamento de Psicología de TEA). Madrid: TEA.

Hathaway, S. R., y McKinley, J. C. (1999). *MMPI-2. Inventario Multifásico de Personalidad de Minnesota-2. Manual* (Adaptación española: Ávila-Espada, A., y Jiménez-Gómez, F.). Madrid: TEA Ediciones.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257-272.

Heilbrun, K., Bennett, W. S., White, A. J., y Kelly, J. (1990). An MMPI-based empirical model of malingering and deception. *Behavioral Sciences & the Law, 8*, 45-53.

Heilbrun, K., Warren, J., y Picarello, K. (2003). Third party information in forensic assessment. En Alan M. Goldstein (Ed.), *Handbook of psychology, vol. 11. Forensic psychology* (pp. 67-86). Hoboken, NJ: John Wiley & Sons.

Hilton, N. Z., y Harris, G. T. (2005). Predicting wife assault: a critical review and implications for policy and practice. *Trauma, Violence, and Abuse, 6*, 3-23.

Hilton, N. Z., Harris, G. T., y Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist, 34*, 400-409.

Ibáñez, I., del Pino, A., Olmedo, E., y Gaos, M. T. (2010). Fiabilidad y validez de una versión Española del Inventario de Depresión de Beck-II en una muestra de la población general Canaria. *Behavioral Psychology/Psicología Conductual, 18*, 35-56.

Klerman, G. L., Weissman, M. M., Rounsaville, B. J., y Chevron, E. S. (1984). Interview format for the Hamilton Rating Scale for Depression (Appendices, pp. 223-233). En G. L. Klerman, M. M. Weissman, B. J. Rounsaville y E. S. Chevron (Eds.), *Interpersonal psychotherapy of depression*. Nueva York: Basic Books.

Losada, A., de los Ángeles Villareal, M., Nuevo, R., Márquez-González, M., Salazar, B. C., Romero-Moreno, R., Carrillo, A. L., y Fernández-Fernández, V. (2012). Cross-cultural confirmatory factor analysis of the CES-D in Spanish and Mexican dementia caregivers. *The Spanish Journal of Psychology, 15*, 783-792.

Meehl, P. E. (1954). *Clinical vs. statistical prediction. A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. (Reimpresión de 1996; Nueva York: Jason Aronson Inc.).

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*, 268-273.

Miller, I. W., Bishop, S., Norman, W. H., y Maddever, H. (1985). The modified Hamilton Rating Scale for Depression: reliability and validity. *Psychiatry Research, 14*, 131-142.

Millon, T. (1998). *MCMI-II. Inventario Clínico Multiaxial de Millon II. Manual* (Adaptación española: Ávila-Espada, A.). Madrid: TEA.

Millon, T. (1999). *MCMI-II. Inventario Clínico Multiaxial de Millon II. Manual* (2ª ed.) (Adaptación española: Ávila-Espada, A.). Madrid: TEA.

Millon, T. (2002). *MCMI-II. Inventario Clínico Multiaxial de Millon II. Manual* (3ª ed.) (Adaptación española: Ávila-Espada, A.). Madrid: TEA.

Millon, T., Davis, R. D., y Millon, C. (2007). *MCMI-III. Inventario Clínico Multiaxial de Millon-III. Manual* (Adaptación española: Cardenal, V., y Sánchez, M. P.). Madrid: TEA Ediciones.

Muñiz, J., y Fernández-Hermida, J. R. (2010). La opinión de los psicólogos españoles sobre el uso de los test. *Papeles del Psicólogo, 31*, 108-121.

Nezu, A. M., Nezu, C. M., Friedman, J., y Lee, M. (2009). Assessment of depression. En I. H. Gotlib y C. L. Hammer (Eds.), *Handbook of depression (2ª ed.)* (pp. 44-68). Nueva York: Guilford Press.

Nezu, A. M., Ronan, G. F., Meadows, E. A., y McClure, K. S. (Eds.). (2000). *Practitioner's guide to empirically based measures of depression*. Nueva York: Kluwer Academic/Plenum Press.

Ortiz-Tallo, M., Cardenal, V., Ferragut, M., y Cerezo, M. V. (2011). Personalidad y síndromes clínicos: Un estudio con el MCMI-III basado en una muestra española. *Revista de Psicopatología y Psicología Clínica, 16*, 49-59.

Otto, R. K. (2008). Challenges and advances in assessment of response style in forensic examination contexts. En R. Rogers (Ed.), *Clinical assessment of malingering and deception, 3ª ed*. (pp. 365-375). Nueva York: Guilford Press.

Pedrero Pérez, E. J., y López-Durán, A. (2005). Autoinformes de sintomatología depresiva en drogodependientes: nivel de coincidencia del BDI, SCL-90-R y MCMI-II. ¿Depresión o malestar inespecífico? *Adicciones, 17*, 215-230.

Pedrero Pérez, E. J., López Durán, A., y Fernández del Río, E. (2012). Dimensiones factoriales del cuestionario de Millon (MCMI-II) en adictos a sustancias. *Psicothema, 24*, 661-667.

Potts, M. K., Daniels, M., Burnam, M. A., y Wells, K. B. (1990). A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *Journal of Psychiatry Research, 24*, 335-350.

Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.

Ramos-Brieva, J. A. (1986). La Escala de Zung-Conde para la Depresión: su validez predictiva. *Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 14*, 123-127.

Ramos-Brieva, J. A., y Cordero Villafáfila, A. (1986a). Validación de la versión castellana de la escala de Hamilton para la depresión. *Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 14*, 324-334.

Ramos-Brieva, J. A., y Cordero Villafáfila, A. (1986b). Relación entre validez y seguridad de la versión castellana de la escala de Hamilton para la depresión.

Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 14, 335-338.

Ramos-Brieva, J., y Cordero-Villafáfila, A. (1988). A new validation of the Hamilton Rating Scale for Depression. *Journal of Psychiatry Research, 22*, 21-28.

Ramos Brieva, J., Cordero Villafáfila, A., y Yáñez Sáez, R. (1994). Nuevos datos sobre la validez y fiabilidad de la versión castellana de la Escala de Hamilton para la Depresión. *Anales de Psiquiatría, 10*, 146-151.

Ramos Brieva, J. A., Lafuente López, R., Montejo Iglesias, M. L., Moreno Sarmiento, A., Ponce de León Hernández, C., Méndez Barroso, R., y Cordero Villafáfila, A. (1991). Validez predictiva de la Escala de Zung en deprimidos ancianos. *Actas Luso-Españolas de Neurología, Psiquiatría y Ciencias Afines, 19*, 122-126.

Rees, L. M., Tombaugh, T. N., y Boulay, L. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 16*, 501-506.

Rodríguez Pulido, F., Méndez Abad, M., González Dávila, E., y Rodríguez García, A. (2008). Estudio comunitario sobre prevalencia de morbilidad psiquiátrica en personas con parasuicidios previos. *Anales de Psiquiatría, 24*, 211-215.

Rogers, R. (1995). *Diagnostic and structured interviewing: A handbook for psychologists*. Odessa, FL: Psychological Assessment Resources.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York: Guilford.

Rogers, R. (2008). An introduction to response styles. En R. Rogers (Ed.), *Clinical assessment of malingering and deception, 3ª ed*. (pp. 3-13). New York: Guilford Press.

Romera, I., Delgado-Cohen, H., Perez, T., Caballero, L., y Gilaberte, I. (2008). Factor analysis of the Zung Self-Rating Depression Scale in a large sample of patients with major depressive disorder in primary care. *BMC Psychiatry, 8:4*.

Ronan, G. F., Dreer, L., Maurelli, K., Ronan, D., y Gerhart, J. (2014). *Practitioner's guide to empirically supported measures of anger, aggression, and violence*. Nueva York: Springer Publishing.

Ros, L., Latorre, J. M., Aguilar, M. J., Serrano, J. P., Navarro, B., y Ricarte, J. J. (2011). Factor structure and psychometric properties of the Center for Epidemiologic Studies Depression Scale (CES-D) in older populations with and without cognitive impairment. *The International Journal of Aging & Human Development, 72*, 83-110.

Salgado, J. F. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y las organizaciones. *Papeles del Psicólogo, 92*, 115-128.

Sanz, J. (2002). La década de 1989-1998 en la Psicología española: análisis de la investigación en personalidad, evaluación y tratamiento psicológico (psicología clínica y de la salud). *Papeles del Psicólogo, 81*, 54-87.

Sanz, J., García-Vera, M. P., Espinosa, R., Fortún, M., y Vázquez, C. (2005). Adaptación española del Inventario para la Depresión de Beck-II (BDI-II): 3. Propiedades psicométricas en pacientes con trastornos psicológicos. *Clínica y Salud, 16*(2), 121-142.

Sanz, J., Gutiérrez, S., Gesteira, C., y García-Vera, M. P. (en prensa). Criterios y baremos para interpretar el "Inventario de depresión de Beck-II" (BDI-II). *Behavioral Psychology-Psicología Conductual*.

Sanz, J., Izquierdo, A., y García-Vera, M. P. (2013). Una revisión desde la perspectiva de la validez de contenido de los cuestionarios, escalas e inventarios autoaplicados más utilizados en España para evaluar la depresión clínica en adultos. *Psicopatología Clínica, Legal y Forense*.

Sanz, J., Perdigón, L. A., y Vázquez, C. (2003). Adaptación española del Inventario para la Depresión de Beck-II (BDI-II): 2. Propiedades psicométricas en población general. *Clínica y Salud, 14*, 249-280.

Seisdedos Cubero, N. (1980). *MMPI: suplemento técnico*. Madrid: TEA.

Soler, J., Pérez-Sola, V., Puigdemont, D., Pérez-Blanco, J., Figueres, M., y Álvarez, E. (1997). Estudio de validación del Center for Epidemiologic Studies-Depression (CES-D) en una población española de pacientes con trastornos afectivos. *Actas Españolas de Psiquiatría, 25*, 243-249.

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM)*. North Tonawonda, NY: Multi-Health Systems.

Tombaugh, T. N. (2011). *Test de Simulación de Problemas de Memoria (TOMM). Manual* (Adaptación española: Vilar-López, R., Pérez, M., y Puente, A. E.). Madrid: TEA Ediciones.

Trajković, G., Starčević, V., Latas, M., Leštarević, M., Ille, T., Bukumirić, Z., y Marinković, J. (2011). Reliability of the Hamilton Rating Scale for depression: a meta-analysis over a period of 49 years. *Psychiatry Research, 189*, 1-9.

Vázquez, C., y Sanz, J. (1997). Fiabilidad y valores normativos de la versión española del Inventario para la Depresión de Beck de 1978. *Clínica y Salud, 8*, 403-422.

Vázquez, C., y Sanz, J. (1999). Fiabilidad y validez de la versión española del Inventario para la Depresión de Beck de 1978 en pacientes con trastornos psicológicos. *Clínica y Salud, 10*, 59-81.

Vázquez, F. L., Blanco, V., y López, M. (2007). An adaptation of the Center for Epidemiologic Studies Depression Scale for use in non-psychiatric Spanish populations. *Psychiatry Research, 149*, 247-252.

Whisman, M. A., Strosahl, K., Fruzzetti, A. E., Schmaling, K. B., Jacobson, N. S., y Miller, D. M. (1989). A structured interview version of the Hamilton Rating Scale for Depression. *Psychological Assessment, 1*, 238-241.

Widows, M. R., y Smith, G. P. (2005). *SIMS: Structured Inventory of Malingered Symptomatology. Professional manual*. Lutz, FL: Psychological Assessment Resources.

Widows, M. R., y Smith, G. P. (2009). *SIMS. Inventario Estructurado de Simulación de Síntomas* (Adaptación española: González Ordi, H., y Santamaría, P.). Madrid: TEA ediciones.

Williams, J. B. W. (1988). A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry, 45*, 742-747.

Williams, J. B. W. (2001). Standardizing the Hamilton Depression Rating Scale: past, present, and future. *European Archives of Psychiatry and Clinical Neurosciences, 251*(Suppl. 2), II/6-II/12.

Williams, J. B., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., et al. (2008). The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *International Clinical of Psychopharmacology, 23*, 120-129.

Yanez, Y. T., Fremouw, W., Tennant, J., Strunk, J., y Coker, K. (2006). Effects of severe depression on TOMM performance among disability-seeking outpatients. *Archives of Clinical Neuropsychology, 21*, 161-165.

Zitman, F. G., Mennen, M. F. G., Griez, E., y Hooijer, C. (1990). The different versions of the Hamilton Depression Rating Scale. En P. Bech y A. Coppen (Eds.), *The Hamilton Scales (Psychopharmacology Series 9)* (pp. 28-34). Heidelberg, Berlin: Springer-Verlag.