Cuadernos de RES PUBLICA en derecho y criminología
Nº 07 / 2026 ISSN: 2990-0697 DOI: 10.46661/respublica.12511

Metodología de investigación con tests psicológicos en criminología

Research methodology with psychological tests in criminology

Ana María Ruiz-Ruano García
Universidad de Granada. Granada. (España)
amruano@ugr.es
ORCID:0000-0002-7260-0588

Jorge López Puga
Universidad de Granada. Granada. (España)
jlpuga@ugr.es
ORCID:0000-0003-0693-0092

Resumen

En este trabajo se presenta un somero recorrido por los principales aspectos a tener en cuenta frente a la investigación con tests, dada la utilidad que tiene esta tecnología, en el ámbito de la criminología. El trabajo se enfoca desde el punto de vista del desarrollo de tests abordando la fase de diseño y pilotaje del instrumento de medida. También se abordan recomendaciones sobre el procesamiento estadístico de los datos prestando atención a los conceptos de fiabilidad y validez. Por último, se presenta una reflexión sobre el uso ético y responsable de la información recogida con tests. Para optimizar la utilidad práctica de este trabajo, presentamos un ejemplo de construcción de test para medir el riesgo de agresión a la pareja.

Palabras clave: desarrollo de tests; estadística; fiabilidad; validez.

Abstract

This work presents a brief overview of the most important aspect to keep in mind when researching with tests in the criminology field of expertise. Tests are useful tools into this area of knowledge and this paper is aimed to provide some general guidelines. The paper is organized from the perspective of tests development perspective. Recommendations are provided from the statistical analysis point of view and addressing the psychometrics concepts of reliability and validity. Finally, a discussion on ethical commitment and social responsibility is also introduced to stress the relevance of these aspects when using psychological tests. To optimize the practical utility of this work, it is presented a construction test example to assess the partner aggression risk.

Keywords: test development; statistics, reliability; validity.

Cómo citar este trabajo: Ruiz-Ruano García, Ana María y López Puga, Jorge. (2025). Metodología de investigación con tests psicológicos en criminología. *Cuadernos de RES PUBLICA en derecho y criminología, (7),* 01–17. https://doi.org/10.46661/respublica.12511

Recepción: 08.08.2025 **Aceptación:** 23.09.2025 **Publicación:** 29.09.2025

1. Introducción

La recolección de información con cuestionario pre-codificado tiene sus inicios en la primera mitad del siglo XX. Esta actividad estuvo vinculada inicialmente al ámbito de la publicidad y del estudio de las actitudes y progresivamente se fue expandiendo a otras áreas de trabajo (Fernández-Abellán y Gómez, 2002). Hoy en día, la recogida de información mediante encuestas podría considerarse como una pieza arquimédica del estado del bienestar en ámbitos tan dispares como la salud, la educación, el trabajo, la política o la economía. Así, por ejemplo, en el ámbito laboral, la Encuesta de Población Activa que realiza anualmente el Instituto Nacional de Estadística (www.ine.es) es considerada como un indicador global del empleo en sus diferentes categorías y es utilizada para planificar u optimizar contingencias socio-políticas que regulen el mercado de trabajo.

No obstante, pese a la presencia reiterada de los tests en un amplio espectro de situaciones de nuestra sociedad contemporánea, hay ocasiones en las que estas tecnologías son producidas y utilizadas bajo dudosos estándares de calidad (Muñiz, 1998). Como señalan Muñiz y Fernández (2000), «los tests son susceptibles de usarse adecuada o inadecuadamente» (p. 41).

Esta situación de uso de tests indiscriminado e insensibilizado es particularmente llamativa en nuestro país (Rojas, 2002). Así, mientras en otros países (como en los Estados Unidos de América) la legislación regula estrictamente el uso de tests y las consecuencias que se derivan de su uso, en España todo lo relacionado con la tecnología de medición psicológica queda desamparada en este sentido.

En este trabajo trataremos de presentar los aspectos más generales que se deberían tener en cuenta frente a la utilización o a la investigación mediante tests en el ámbito criminológico dado que el uso de tests para la investigación en este ámbito es frecuente y se espera que tienda a aumentar (p.e., García, 2000; Lara y Molina, 2014). Para abordar esta descripción de los elementos que requieren atención frente a la investigación con tests, enfocaremos esta discusión considerando los pasos que se deberían tomar en el proceso de desarrollo de este tipo de instrumentos de medida. Aunque hay que advertir que el proceso de creación de tests es largo, complejo y que está condicionado por los modelos psicométricos teóricos que auspician la creación del test, nos basaremos en la propuesta introducida por Muñiz y Fonseca-Pedrero (2008) en la que identifican diez etapas en la construcción de tests:

1. marco general del instrumento de medida,
2. definición operativa de las variables de medida,
3. especificaciones del instrumento de medida,
4. construcción de los ítems,
5. elaboración de normas de puntuación,
6. estudio piloto cualitativo y cuantitativo,
7. selección de otros instrumentos de medida convergentes,
8. estudio de campo,
9. estimación de las propiedades psicométricas, y
10. versión definitiva del instrumento.

No obstante, este trabajo se enfocará atendiendo a la síntesis que de estos pasos introdujo López Puga (2013) con ánimo de aportar claridad expositiva al asunto. Para presentar esta exposición, vamos a basarnos en un ejemplo ficticio en el que un equipo de investigación estuviese interesado en desarrollar y utilizar una escala destinada a valorar o predecir la violencia en alguna de sus formas.

La predicción de la violencia mediante el uso de tests es sólo un ejemplo de los fenómenos psicológicos que pueden ser tratados en la investigación criminológica (Pérez, Sáiz y Sáiz,

2005). Como bien es sabido, la agresión es posiblemente un tipo de comportamiento consustancial a la naturaleza humana, aunque la violencia tiene un significado conceptual muy diferente desde el punto de vista criminológico (Bartol y Bartol, 2017).

Supongamos que un equipo de investigación estuviese interesado en desarrollar un test destinado a evaluar y predecir la violencia en la pareja de manera parecida a como lo hace el SARA o *Spousal Assault Risk Assessment Guide* (Kropp y Hart, 2000).

2. Diseño

La planificación de cualquier actividad humana es, en muchas ocasiones, una de las claves que permite explicar el éxito de la misma a medio-largo plazo. En la medida en que planificamos lo más eficientemente posible una actividad menores problemas encontraremos cuando tratemos de llevarla a cabo. O, es más, cuando encontremos algún escollo en el camino podremos sortearlo con más facilidad.

Por su parte, las consecuencias indeseables de la carencia de planificación podrían apreciarse en momentos en los que, una vez puesta en marcha cierta actividad, descubrimos que hemos olvidado incluir algo importante en nuestros planes. En estos casos perderemos un valiosísimo tiempo solventando este problema antes de poder proseguir exitosamente con la tarea.

Uno de los puntos más importantes que hemos de tener en cuenta a la hora de planificar nuestro instrumento de medida es el objetivo del mismo. En el ámbito de la psicología se suelen diseñar instrumentos de medida destinados, entre otros menesteres, a: evaluar, seleccionar, ordenar/clasificar, orientar o a diagnosticar. Por ejemplo, el instrumento de medida podría no ser el mismo si quisiésemos evaluar el tipo de personalidad que manifiesta un conjunto de personas, que si quisiésemos clasificar a un conjunto de personas como potencialmente susceptibles de ser consideradas delincuentes en función de su nivel de neuroticismo entendido como una dimensión de la personalidad.

Una vez definido claramente el objetivo de nuestro test deberíamos definir el constructo o constructos que aspiramos a medir (Martín, 2004). Definir un constructo implicaría expresar verbalmente el significado del mismo (definición semántica u operacional) así como explicitar las relaciones teóricas que se establecen entre el constructo y otros constructos relacionados o no relacionados (definición sintáctica o relacional). Por ejemplo, pongamos el caso de estar interesados en medir un constructo llamado *optimismo*.

El optimismo, tal y como lo definen semánticamente Scheier y Carver (1985), se entendería como *la expectativa que tiene una persona sobre su futuro entendiendo que le sucederán más cosas positivas que negativas*. Obsérvese que esta somera definición podría servirnos para diseñar ítems o preguntas destinadas a valorar cuán optimista es una persona. Por ejemplo, podríamos pedirle que responda *sí* o *no* a la siguiente pregunta: *¿crees que es más probable que te ocurra algo bueno que algo malo la semana que viene?* En caso de que la persona respondiese *sí*, la persona estaría ubicándose en una situación optimista tal y como se ha definido desde el punto de vista semántico.

Sin embargo, la definición semántica no es suficiente para garantizar una buena definición de constructo ya que podríamos caer en un fenómeno de circularidad conceptual al tratar de definir el propio constructo.

Algo parecido a lo que observó John B. Watson cuando trataba de entender cómo se explicaban ciertos aspectos del comportamiento humano (Hothershall, 1997). La definición sintáctica consiste en especificar un conjunto de relaciones lógico-matemáticas que relacionen nuestro constructo con otros relacionados, no relacionados y neutrales. Por ejemplo, por seguir con el ejemplo del optimismo, podríamos decir que el optimismo se

relaciona positivamente con la felicidad, negativamente con la ansiedad y neutralmente con el autoritarismo. Dicho de otra manera, que una persona optimista tenderá a ser más feliz y menos ansiosa que una persona pesimista. Por su parte, el ser optimista no tendría relación con ser una persona más o menos autoritaria. Las definiciones semánticas y sintácticas cobran un papel muy relevante cuando se lleva a cabo la validación de las escalas de medida psicológicas. Es por ello que esta primera fase de diseño es tan importante en fases posteriores del desarrollo y utilización de tests.

Una vez que ha sido definido el constructo, o constructos, con el que vamos a trabajar es el momento de diseñar los ítems. Existen diferentes técnicas destinadas a crear ítems y su abordaje escapa de los propósitos de este trabajo. En una primera fase de la construcción de ítems puede acudirse a técnicas como la «tormenta de ideas» (León, 1994) aunque todo el proceso ha de monitorizarse con el objetivo de que se cumplan los máximos estándares de calidad.

En el trabajo de Suen y McClelland (2003) se pueden encontrar algunas recomendaciones y directrices destinadas a orientar el proceso de construcción de ítems. En el caso de que los ítems que queramos incluir en nuestro test provengan de un test ya creado en, por ejemplo, otro idioma o cultura; se deberá realizar una adaptación de los mismos atendiendo a las directrices correspondientes (véase, por ejemplo, Muñiz, Elosua y Hambleton, 2013; Muñiz y Hambleton, 1996).

En cuanto a los tipos de ítems, podríamos clasificarlos en dos grandes grupos siguiendo a Hambleton (1996) cuando hace lo propio para referirse a los tests de evaluación educativa:

1. ítems de respuesta seleccionada,
2. ítems de respuesta construida.

Dentro de los ítems de respuesta seleccionada, o de respuesta pre-codificada, cobran gran protagonismo los ítems de elección múltiple. Es decir, son ítems o preguntas que presentan un conjunto de alternativas de las que el participante del estudio ha de seleccionar solo una de ellas. Un ejemplo de este tipo de ítems lo tendríamos al preguntar por el estado civil de una persona y donde se facilitan las siguientes alternativas: soltero/a, casado/a, divorciado/a y viudo/a.

Los ítems tipo Likert son una variante de los ítems de elección múltiple donde las posibles respuestas están graduadas y con los que la persona evaluada indica cierta cantidad de algo (por ejemplo, satisfacción o acuerdo con lo expresado por el ítem). Lo más común es encontrar este tipo de ítems en escalas de tres, cuatro o cinco alternativas, aunque también es posible encontrar ítems con siete o nueve posibles respuestas.

Existe una variante de este tipo de ítems en los que se pueden seleccionar más de una posible alternativa. Por ejemplo, imagínese que se quisiese conocer el patrón turístico de una persona en el último año. Se le podría plantear un ítem similar a este: indique si ha realizado al menos uno de los siguientes tipos de viaje en el último año. Si las alternativas fuesen: local, regional, nacional e internacional; sería sensato pensar que una persona podría haber tomado parte en más de un tipo de viaje de los presentados (por ejemplo, haber realizado viajes locales y nacionales) en el último año.

Desde el punto de vista estadístico este tipo de ítems se equipara a lo que es denominado como variables de tipo lógico, o a lo que se denomina como variables dicotómicas, ya que la selección de una alternativa del ítem implica asumir que lo que expresa la alternativa es correcto mientras que si la alternativa no es elegida se mantiene, como se pensaba por defecto, en estado incorrecto.

Las preguntas o ítems cuyas posibles respuestas son el «sí» y el «no», así como las preguntas de verdadero-falso también son consideradas variables lógicas. Por último, las preguntas de emparejamiento o relación también son consideradas como ítems de respuesta elegida aunque no son tan comunes como las anteriores.

Mientras que los ítems de respuesta elegida son deseables para muchas situaciones dado que permiten recoger la información de manera estandarizada, también es cierto que dejan poco lugar a la expresión abierta de la persona. Por ello, las preguntas de respuesta construida pueden ser de utilidad cuando se quiere recoger información cualitativa valiosa de la persona que es evaluada. Los dos tipos principales de respuesta construida serían el ítem de respuesta elaborada breve y el de respuesta elaborada larga (o ensayo).

Al utilizar este tipo de preguntas en un test hay que tener disponible un sistema de codificación del texto producido para que podamos utilizar las respuestas a las mismas de manera eficiente en consonancia con nuestro objetivo del test. Pese a que producen una gran cantidad de información rica en detalles, la utilización de ítems de respuesta construida puede ser problemática en numerosos contextos aplicados sobre todo si, como se ha indicado anteriormente, no se cuenta con un sistema de tipificación de las respuestas emitidas por los participantes.

Otro aspecto importante a decidir en la fase de diseño del cuestionario consistirá en especificar el procedimiento que se seguirá para combinar las puntuaciones de los ítems para aquellos casos en los que formen parte de una escala. Es decir, si existen varios ítems que en su conjunto están destinados a valorar un constructo, habrá que indicar cómo se van a combinar para producir una puntuación total del constructo medido.

Uno de los métodos más empleados consiste en sumar la puntuación de cada uno de los ítems para cada una de las personas, aunque también se puede obtener la media o la proporción de respuestas positivas (en el caso de ítems dicotómicos), entre otras.

Por último, también sería conveniente, llegados a este punto, especificar si la puntuación de las personas en los constructos se valorará absoluta o relativamente. Hay diferentes tipos de transformaciones (por ejemplo, tipificación de la puntuación total o transformación percentil de la misma) que pueden ser de utilidad en este ámbito y un resumen de ellas puede encontrarse en el trabajo de López Puga (2013).

Volvamos al ejemplo ficticio que se propuso inicialmente en la primera sección de este trabajo. Como se planteó en su momento, el interés por desarrollar un test para medir la violencia en la pareja podría estar justificado por la necesidad que plantean ciertas situaciones periciales a nivel judicial. Supongamos que el interés del equipo de investigación es diseñar un test que permita identificar el riesgo que existe de que un miembro de una pareja pueda desplegar comportamientos violentos contra el otro miembro de la misma. Una herramienta de este tipo podría ser deseable dado que sería útil para prevenir maltratos o agresiones altamente lesivas en la víctima.

Tras fijar este objetivo legítimo convendría definir operacionalmente el constructo a estudiar. Supongamos que se decide denominar al constructo *riesgo de agresión a la pareja* (RAP). Como se ha indicado anteriormente, el RAP debería definirse tanto semánticamente como sintácticamente buscando apoyo en la literatura científica disponible y relacionada con la idea de constructo que se pretende desarrollar. Por ejemplo, el RAP podría definirse semánticamente como "*la probabilidad subjetiva que estima una persona de que ella misma podría perpetrar acciones (verbales o no-verbales) violentas contra su propia pareja*". Sin embargo, esta definición sería limitada sin disponer de una definición sintáctica de la misma. La definición sintáctica implicaría poner este constructo en relación con otros teóricamente afines. Por ejemplo, podría considerarse que una alta puntuación en la escala que midiese el RAP se relacionaría con altas puntuaciones en otras escalas de violencia, con altas puntuaciones de psicopatía, celotipia o, incluso, con escalas de autoritarismo.

También podrían proponerse relaciones estadísticas en sentido antagónico con la puntuación en el RAP. Por ejemplo, podría

proponerse que altas puntuaciones en esta escala estarían relacionadas con bajos niveles de empatía, con paupérrimas habilidades de resolución de problemas o con bajos niveles de tolerancia a la frustración.

La fase de elaboración de ítems, como ha sido comentado, implicaría numerosas tomas de decisiones. Para la escala ficticia de valoración del RAP podría ser viable diseñar ítems tipo Liker de respuesta graduada. Así, podría solicitarse a la persona evaluada que indicase el grado de acuerdo/desacuerdo con cada uno de los ítems presentados utilizando una escala de, por ejemplo, once puntos cada uno de los cuales correspondería a cada uno de los números enteros que van desde el uno hasta el diez incluyendo al cero en la parte más baja de la escala.

De este modo, se instruiría a la persona evaluada a que seleccionase un punto de esa escala para indicar su acuerdo con lo expresado por el ítem teniendo en cuenta que el cero significa absoluto desacuerdo con el ítem y que diez significaría acuerdo absoluto con el ítem. Como se indicó más arriba, el proceso de generación de ítems puede tomar muchas formas y puede atravesar varias fases, pero finalmente se generarían un conjunto de ítems que podrían ser parecidos a este: "Creo que sería capaz de agredir físicamente a mi pareja".

Obsérvese que este ítem aludiría explícitamente a la simple e hipotética definición de RAP que ha sido introducida anteriormente ya que apela a la creencia subjetiva de tendencia violenta hacia la pareja en la persona que lo responde.

Además de contar con una muestra preliminar de ítems para medir el constructo RAP, en esta fase se debería decidir cómo corregir el test y obtener la puntuación final del mismo. Para el ejemplo que estamos desarrollando quizá sería buena idea expresar la puntuación del test en una escala de cero a 100 independientemente del número final de ítems que éste contenga.

En este sentido, podría definirse algebraicamente la puntuación total en el test de RAP utilizando la siguiente ecuación:

$$RAP = \frac{\sum_{i=1}^{k} x_i}{n_k} \times 100,$$

donde x_i se refiere a la puntuación obtenida en el ítem *i* y n_k es el número de ítems del test.

3. Pilotaje

Tras el diseño del cuestionario es preciso que sea sometido a prueba en un pequeño subconjunto de personas análogas a las que va destinado el test. Por ejemplo, si el test está destinado a valorar ciertos aspectos de las personas que han sufrido robos en sus casas, sería conveniente que se administrase una versión preliminar del instrumento a una pequeña muestra de personas que cumplan esa característica.

El objetivo de llevar a cabo un pilotaje del instrumento en estos términos es tratar de identificar debilidades o aspectos del instrumento que no funcionan como cabría esperar.

El pilotaje del test podría hacerse tanto cualitativa como cuantitativamente. En las siguientes líneas describiremos sucintamente a qué nos referimos con ello.

En el análisis cualitativo de ítems se lleva a cabo un proceso de revisión cualitativa sistemática de cada uno de los ítems que componen el cuestionario con el objetivo de encontrar anomalías o debilidades en el test. Una práctica común bajo la rúbrica del análisis cualitativo de ítems suele consistir en enviar el instrumento de medida a un conjunto de expertos sobre la materia que pretende evaluar el test para que emitan un juicio sobre la idoneidad técnica del instrumento.

Los expertos pueden valorar diferentes aspectos de los ítems. Entre estos aspectos podríamos destacar: la redacción de los ítems, la pertinencia de los mismos o la pertenencia de cada ítem a una u otra dimensión de estudio. En esta fase, es importante que los expertos reciban, junto al test, una

descripción pormenorizada del objeto de estudio y unas instrucciones sobre cómo valorar los ítems. Una vez que los expertos han revisado el test remiten sus conclusiones a los investigadores y éstos, a su vez, realizan una síntesis de las opiniones vertidas por los expertos sobre su instrumento.

Otra forma complementaria de valorar el funcionamiento preliminar del test pasaría por realizar un análisis cuantitativo de los ítems del test. En este caso, se ha de hacer uso de técnicas de análisis de datos estadísticos para corroborar que los ítems del instrumento de medida funcionan debidamente en el ámbito de estudio en que se aplicarán.

Este análisis cuantitativo de ítems depende, como indica López Puga (2013), del modelo teórico-psicométrico que sustenta el desarrollo del test. Para una revisión en mayor detalle se invita al lector a profundizar en ello por medio de bibliografía especializada (véase, por ejemplo, Crocker y Algina, 1986; Muñiz, 1992, 1997, 2010). Por ejemplo, un análisis estadístico descriptivo de las respuestas de la muestra de pilotaje podría servir para identificar funcionamientos anómalos o indeseables de ciertos ítems. Supongamos que cuando calculamos las frecuencias absolutas y relativas de respuestas a las alternativas de un ítem.

Si observásemos que una de las alternativas no es elegida por una muestra piloto representativa de la población a la que va destinado el test, quizá sería conveniente mejorar el ítem eliminando o modificando esa alternativa concreta.

Otro aspecto importante del análisis cuantitativo de los ítems está relacionado con la fiabilidad y la validez de las puntuaciones que genera el test. Este aspecto se abordará en la siguiente sección del trabajo.

Siguiendo con el ejemplo que se introdujo anteriormente sobre la medición del riesgo de agresión a la pareja o RAP, podríamos pilotar la muestra de ítems preliminares pidiendo su opinión al respecto por parte de varios expertos en el tema. Podría contarse con personas expertas en violencia doméstica o en terapia de pareja. Forenses con especialidad en psicología u otros perfiles de profesionales ligados al campo legal también podrían formar parte del panel de expertos que valorarían los ítems preliminares. Entre otras cosas, podríamos pedir a los expertos que valorasen el grado en que los ítems preliminares podrían servir para valorar el RAP. Podrían darnos su opinión sobre el formato de respuesta propuesta o, en caso de que nuestra escala estuviese integrada por varias dimensiones, valorar el grado en que cada uno de los ítems es representativo de las dimensiones consideradas.

El panel de expertos también podría detectar errores gramaticales en los ítems, la pertinencia de los mismos o su significado para valorar el constructo bajo estudio.

Este análisis cualitativo de ítems es crítico frente a los estudios de validez de contenido porque el panel de expertos puede aportar información clave sobre este aspecto psicométrico.

Supongamos que administrásemos nuestros ítems preliminares a una muestra de personas que se considerasen representativas de la población objetivo a la que iría destinado el test. Por ejemplo, supongamos que tenemos la oportunidad de administrar nuestros ítems a un conjunto de 100 personas que han sido procesadas legalmente por presuntos delitos de violencia contra sus parejas.

Con los datos generados por esas respuestas podríamos llevar a cabo análisis cuantitativos de estos ítems y podríamos valorar el grado de adecuación a los supuestos de medida del modelo psicométrico que hayamos considerado en la fase inicial de planificación del test.

En la Figura 1 aparecen tres ejemplos de distribuciones de frecuencias ficticias para tres posibles ítems.

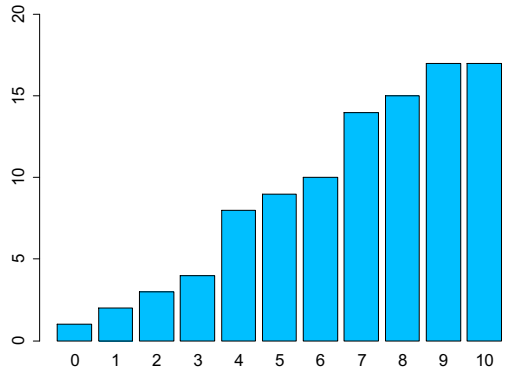
Estas representaciones gráficas podrían ser de utilidad en esta fase del desarrollo del test para detectar ítems que se comportan anómalamente atendiendo a la situación de

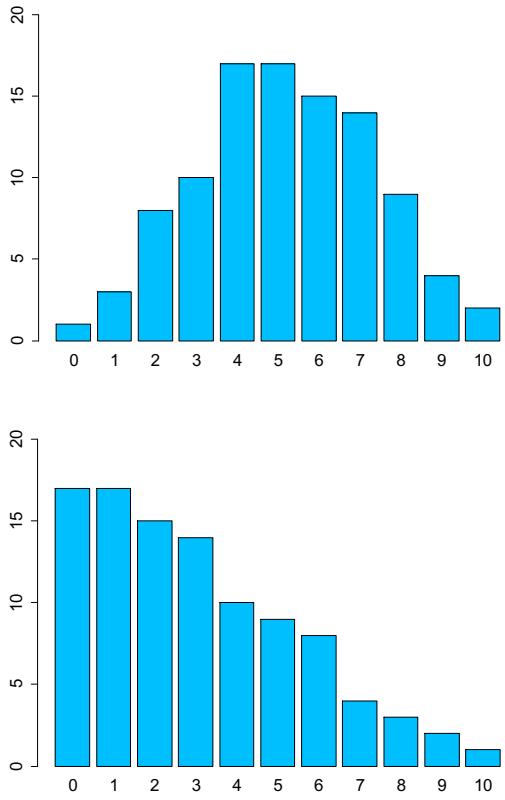
aplicación y a los supuestos teóricos que subyacen al constructo. Por ejemplo, la distribución de frecuencias que aparece en la primera figura, mostraría un ítem en el que la mayor parte de las personas que lo responden eligen puntuaciones altas del mismo. Este perfil gráfico podría ser esperable, por ejemplo, si la muestra de pilotaje fuese una en la que la tendencia a agredir fuese altamente prevalente.

Por su parte, en el tercer gráfico de la aparece un perfil antagónico. Este perfil gráfico podría aparecer si la muestra utilizada hubiese estado compuesta por personas elegidas aleatoriamente de la población. En este gráfico se apreciaría que la mayor parte de la población puntuaría valores muy bajos en el ítem lo que significarían que su probabilidad de agredir a su pareja es nula. Sin embargo, habría un porcentaje residual y muy pequeño de personas que sí podrían manifestar esa tendencia subjetiva a agredir a su pareja.

Por su parte, en el gráfico central aparece una distribución de frecuencias con una forma parecida a la distribución normal. Este tipo de perfiles son deseables en las puntuaciones de una gran variedad de constructos psicológicos.

Figura 1. Algunas distribuciones de frecuencias para tres ítems hipotéticos diseñados para medir el constructo RAP.





Ítem sesgado negativamente (arriba), ítem que se distribuye aproximadamente como una normal (centro) e ítem con sesgo positivo (abajo). Fuente: elaboración propia.

4. Análisis de datos

Como cualquier instrumento que proporciona una medida, los tests psicológicos son susceptibles de ser utilizados para estudiar relaciones cuantitativas o cualitativas entre variables.

Por ejemplo, se pueden utilizar para indagar sobre la relación que se establece entre el nivel de estudios de un preso y su potencialidad para reinsertarse eficientemente en la sociedad, para elaborar un perfil psicológico de un tipo concreto de criminal o víctima o para valorar la probabilidad de que un ladrón vuelva a cometer un delito de las mismas características en los próximos doce meses.

En el último epígrafe de esta sección se darán algunas pinceladas sobre algunas de las técnicas que se pueden utilizar para analizar, en el sentido expresado anteriormente, los datos recogidos con tests mientras que en las dos primeras secciones se abordará el análisis de datos básico que permite valorar la calidad

técnica o psicométrica de los tests: la fiabilidad y la validez.

4.1. Fiabilidad

Como indica Peter (1979), que un test sea fiable es una condición necesaria pero no suficiente para que el instrumento de medida goce de calidad técnica. Por fiabilidad de un test nos estamos refiriendo al grado de error que comente cuando evalúa el constructo o la variable correspondiente.

Cuando un test mide con muy poco error el constructo o la variable bajo estudio decimos que el test es fiable. Por el contrario, cuando el instrumento de medida estima una medida con gran cantidad de error decimos que existe una baja fiabilidad.

El estudio de la fiabilidad de un test está condicionado por el modelo psicométrico que tomemos de base en el diseño del mismo y aquí sólo nos centraremos en abordarla desde el punto de vista de la Teoría Clásica de Test (Muñiz, 2010) por ser la perspectiva más ampliamente usada en el contexto aplicado.

Los procedimientos empíricos destinados a estimar la fiabilidad de un test podrían agruparse en dos grandes grupos (López Puga, 2013):

1. utilizando un único conjunto de puntuaciones del test, y
2. utilizando dos, o más, puntuaciones del test.

Comenzaremos discutiendo los procedimientos que sólo requieren una única administración del test.

Por lo general, recoger información con tests psicológicos suele ser una tarea que consume tiempo y dinero. Por ello, se han desarrollado procedimientos que pretenden estimar la fiabilidad de un instrumento de medida utilizando sólo una administración del mismo. Dado que no tenemos un referente externo para corroborar que el test mide sin error el constructo o variable objeto de estudio se suele decir que estos procedimientos estiman la fiabilidad entendida como consistencia interna del test.

Existe un primer método empírico para estimar fiabilidad que se conoce como «procedimiento de las dos mitades».

Con este procedimiento se dividen los ítems del test que mide un constructo psicológico en dos grupos, se obtienen las puntuaciones totales para cada individuo considerando las dos mitades del test de manera separada y, finalmente, se estima la correlación que se establece entre las puntuaciones generadas por cada una de las partes del test. Este índice de correlación se corrige, para obtener el índice de fiabilidad, aplicando una ecuación y por ello también se denomina como método Spearman-Brown.

El punto clave, y que hace poco práctico el método del procedimiento de las dos mitades, está en dividir los ítems del test. Existen varios procedimientos para dividir los ítems que componen el test. Por ejemplo, se pueden considerar los ítems pares e impares por separado, los de la primera parte del test y los de la segunda o, incluso, podría hacerse una división aleatoria de los ítems. Sin embargo, independientemente de cuál sea el método utilizado para dividir los ítems y lo equiparables que sean las dos mitades, se podría llegar a resultados diferentes si se hubiese utilizado cualquier otra forma de dividir los ítems.

El coeficiente alfa (α) ha sido el procedimiento más usualmente utilizado para contrarrestar este problema que se achaca al procedimiento de las dos mitades. El coeficiente alfa es el equivalente al promedio de todas las posibles índices de fiabilidad que se obtendrían dividiendo el test en dos partes.

Hay que tener en cuenta que el coeficiente alfa es el límite inferior de la fiabilidad de un test y que suele utilizar indiscriminadamente sin comprobar que se satisfacen los supuestos del modelo estadístico-psicométrico que subyacen a su formulación matemática (Elosua y Zumbo, 2008).

Por ello, cada vez son más las voces que abogan por la utilización del coeficiente omega (ω) como una alternativa más robusta

y más apropiada para valorar la fiabilidad entendida como consistencia interna de un test unidimensional (Dunn, Baguley, y Brunsden, 2014; Kelley y Cheng, 2012).

Cuando la fiabilidad es entendida como estabilidad temporal puede calcularse un estadístico que nos informe sobre el grado en que las puntuaciones del test son robustas al paso del tiempo. Lo que suele hacerse en estos casos es medir el mismo constructo en las mismas personas, pero en dos momentos del tiempo diferentes (por ejemplo, en el presente y tras el transcurso de varias semanas).

Como se puede deducir, este procedimiento implica la recolección de dos conjuntos de puntuaciones y, por tanto, supondría una inversión mayor de tiempo y dinero. Una vez recolectados ambos conjuntos de puntuaciones se obtiene el índice de consistencia interna que no es más que una estimación de la correlación que se establece entre las puntuaciones en el test en el momento presente y trascurrido cierto tiempo.

Uno de los aspectos más delicados de este procedimiento para estimar fiabilidad es delimitar la cantidad de tiempo que ha de transcurrir entre una administración del test y la siguiente. Lo que hay que hacer en estos casos es determinar un intervalo temporal de demora entre una aplicación que no favorezca el cambio en el constructo estudiado.

Otro conjunto de procedimientos que podrían ser de gran utilidad en el ámbito de conocimiento que nos ocupa son aquellos que se encuadran bajo la etiqueta de «acuerdo perceptual» (LeBreton y Senter, 2008; James, 1982; James, Demaree, y Wolf, 1984). Por medio de este tipo de procedimientos se estima el grado en que un conjunto de personas, observadores expertos o «jueces» están de acuerdo o coinciden al valorar un objeto, situación o fenómeno.

El índice de acuerdo interjueces kappa (κ) de Cohen es un ejemplo de este tipo de procedimientos destinados a valorar la concordancia entre las respuestas de diferentes observadores (León y Montero, 2003).

Si retomamos el ejemplo principal que vamos siguiendo en el manuscrito podríamos llevar a cabo dos estrategias destinadas a valorar la fiabilidad de los ítems diseñados para evaluar el riesgo de agresión a la pareja. Por un lado, sería recomendable valorar la estabilidad temporal la puntuación que genera el test de RAP. En este caso, como se indicaba anteriormente, se deberían recolectar respuestas al test en dos momentos diferentes para las mismas personas. Además, una medida de la consistencia interna de los ítems también convendría para este caso.

Si el test es unidimensional y no podemos garantizar que la variabilidad de todos los ítems es la misma, como suele suceder en la mayor parte de los casos prácticos (Dunn et al., 2014), sería recomendable estimar el índice omega de consistencia interna.

4.2. Validez

Si anteriormente indicábamos que la fiabilidad es una condición necesaria pero no suficiente para que un instrumento de medida goce de calidad técnica, ahora tenemos que apuntar que sin validez la medida psicológica no es posible. De hecho, algunos autores consideran que la fiabilidad de un test es otra evidencia más de la validez de un test psicológico (p.e., Cook y Beckman, 2006).

La validez se ha entendido tradicionalmente como el «grado en el cual un instrumento mide ciertamente aquello que pretende medir» (Peter, 1979, p. 6). No obstante, el concepto de validez ha sufrido una evolución histórica que, según Elosua (2003), podría caracterizarse por tres fases diferentes.

En una primera etapa histórica, la validez de un test estaba íntimamente relacionada con la operacionalización de las medidas que dejaba entrever la profunda influencia que el positivismo lógico ejerció en la evolución de la ciencia psicológica. Algún tiempo más tarde el foco de atención se orientó a la teoría, dando una relevancia capital a las relaciones teóricas

que se establecen entre los constructos estudiados. Por último, tal y como se entiende contemporáneamente, la validez es entendida de manera más contextual.

Dicho de otro modo, la validez de un test sería una hipótesis momentánea referida al conjunto de evidencias que apoyan las interpretaciones que hacemos de las puntuaciones de un test y su utilidad en un contexto determinado (p.e., Cook y Beckman, 2006; Elosua, 2003).

Así, por ejemplo, un test de medición de la ansiedad que haya mostrado ser válido para diagnosticar a pacientes en una clínica psicológica no sería, necesariamente, válido para evaluar la ansiedad experimentada por un testigo en una rueda de reconocimiento de sospechosos. Más bien, tendríamos que aportar evidencias que justificasen que el test utilizado genera puntuaciones que son legítimamente interpretables y útiles en el contexto concreto en que estamos administrándolo.

Desde el puno de vista empírico, esta definición de la validez genera un estado de incertidumbre sobre la calidad técnica de un test que requiere una aproximación responsable y cautelosa frente a la administración de test en diferentes contextos.

La validez de un test es un aspecto que está indisolublemente ligado, como hemos sugerido anteriormente, al proceso de desarrollo del test. Así, en las fases iniciales de construcción del test hay que ir determinando las posibles evidencias que servirán, a posteriori, para justificar la interpretación y el uso de las puntuaciones que genera el test (Elosua, 2003).

Desde la perspectiva clásica, este conjunto de evidencias estaba relacionada con lo que se denomina validez de contenido, validez criterial y validez de constructo. Cuando se alude a la validez de contenido nos estamos refiriendo al grado en que el test recoge una muestra representativa de los comportamientos, dimensiones o ítems que sirven para evaluar el constructo bajo estudio. La validez criterial está referida al grado en que las puntuaciones del test covarían con un comportamiento observable que se considera un reflejo claro del constructo que estamos valorando. Y, por último, la validez de constructo está referida al grado en que las puntuaciones generadas por el test pueden explicarse aludiendo a la variabilidad del constructo objeto de análisis.

En cualquier caso, el análisis de la validez de un test no se limita, hoy en día, a presentar evidencias sobre la calidad psicométrica del test teniendo en cuenta las tres formas clásicas de entender la validez que han sido expuestas previamente. Más bien, lo que se pretende es aportar el mayor número de evidencias posibles que justifiquen que las puntuaciones que genera el test son útiles y sensatas para un ámbito de aplicación concreto. Así, se deberán proporcionar, entre otras, evidencias sobre la validez diferencial, validez factorial, validez predictiva, validez convergente, validez divergente, validez aparente o validez ecológica de las puntuaciones que genera el test.

En el caso de un test que permite evaluar o predecir el riesgo de agresión en la pareja y que venimos desgranando en este texto podríamos formularnos las siguientes cuestiones: ¿son útiles las puntuaciones que genera el test para valorar el riesgo de que una persona agreda a su pareja en el contexto, por ejemplo, del peritaje judicial?

Para responder a esta pregunta que apela directamente a la validez de las puntuaciones del test podríamos abordar la cuestión desde varios ángulos. Si nos apoyamos en el modelo más clásico, tendríamos que disponer de evidencias sobre el contenido, sobre algún criterio externo y sobre la “verosimilitud” del constructo estudiado.

El estudio de la validez del contenido del test podría haberse iniciado en la fase en la que se solicitó a los expertos que valorasen los ítems. En esa fase podría haberse solicitado al panel de expertos que indicasen si se estaba obviando algún elemento esencial sobre el

riesgo de agresión en la pareja. Esto es un ejemplo que pone de manifiesto que los estudios de validez no se llevan a cabo en las fases finales del desarrollo de tests sino que, más bien, es algo que se viene fraguando desde las fases más iniciales del desarrollo del test.

Para valorar la validez de criterio de las puntuaciones del test habría que comprobar si la puntuación en RAP está relacionada con comportamientos agresivos hacia la pareja manifiestamente observables. Por ejemplo, podríamos administrar el test a una muestra de personas y recabar información sobre detenciones previas o denuncias interpuestas relacionadas con agresiones hacia la pareja. Si se observase una relación positiva entre la puntuación del test y alguna de esas variables dispondríamos de evidencia en favor de que las puntuaciones en el test de RAP evalúan el riesgo de que una persona agreda a su pareja.

Por su parte, para aportar evidencias sobre la validez de constructo podrían realizarse diferentes comprobaciones. En primer lugar, tendríamos que retomar la definición sintáctica del constructo que señalábamos más arriba y testar las relaciones teóricas previamente hipotetizadas.

En la medida en que estas relaciones señaladas en la definición sintáctica del constructo se satisfacen, estaremos encontrando evidencias sobre la validez del constructo "riesgo de agresión a la pareja". Análisis de estructuras de varianzas-covarianzas de diferente sofisticación (Análisis de Componentes Principales o Análisis Factoriales tanto Exploratorios como Confirmatorios) también podrían llevarse a cabo para aportar pruebas de la validez del constructo estudiado.

4.3. Relación entre variables

Si anteriormente indicábamos que la fiabilidad es una condición necesaria pero no suficiente para que un instrumento de medida goce de calidad técnica, ahora tenemos que apuntar que sin validez la medida psicológica no es posible.

De hecho, algunos autores consideran que la fiabilidad de un test es otra evidencia más de la validez de un test psicológico (p.e., Cook y Beckman, 2006).

Al abordar el análisis estadístico de datos utilizando información procedente de tests psicológicos debemos actuar con cautela. Esta aproximación razonada al análisis de datos sobre variables procedentes de tests psicológicos viene auspiciada por la naturaleza de las variables que se registran con este tipo de instrumentos de medida. En nuestra opinión, la sugerencia hecha por Stevens (1946) goza de plena vigencia en nuestros días. Por ello, sugerimos que se utilicen los cómputos estadísticos apropiados y legítimos que corresponda realizar al reconocer el nivel de medida de las variables registradas por los tests psicológicos.

Aparte de la matización sobre las escalas de medida anteriormente introducida, los datos registrados con un test psicológico pueden ser analizados tanto descriptiva (Solanas, Salafranca, Fauquet y Núñez, 2005) como inferencialmente (Pagano, 1999).

Desde el punto de vista multivariado, los modelos de ecuaciones estructurales gozan de gran popularidad en el ámbito de estudio de las ciencias sociales (Ruiz, Pardo y San Martín, 2010). Los modelos de ecuaciones estructurales permiten modelar las relaciones estructurales que se establecen entre un conjunto de variables por medio de un conjunto de ecuaciones lineales.

Estas técnicas permiten contrastar hipótesis de causalidad expresadas en términos gráficos, son de utilidad en el ámbito de la validación factorial de constructos así como para caracterizar la relación que se establece entre estos y otras variables.

No obstante, los modelos de ecuaciones estructurales son modelos multivariados paramétricos y los supuestos del modelo imponen estrictas restricciones a los datos susceptibles de ser analizados. Por ello, algunos autores han apuntado a las redes bayesianas como alternativas plausibles como

herramientas de modelado gráfico multivariadas (Anderson y Vastag, 2004).

Entre el conjunto de ventajas que reporta el uso de redes bayesianas en el ámbito del modelado gráfico-estadístico multivariado se podrían destacar las siguientes (Anderson y Vastag, 2004):

1. Son modelos que soportan tanto estructuras de relaciones lineales como no lineales.
2. Mientras que los modelos de ecuaciones estructurales imponen ciertas restricciones sobre la normalidad multivariada de los datos, las redes bayesianas no pre-suponen ningún tipo de distribución en los datos muestrales.
3. Mientras que los modelos de ecuaciones estructurales están orientados al modelado teórico y conceptual, las redes bayesianas están más destinadas a la práctica, al diagnóstico, a la clasificación y a la predicción.

Además, como han indicado López Puga y García (2011), las redes bayesianas permiten gestionar eficientemente la presencia de datos perdidos, pueden utilizarse combinando el conocimiento experto (previo) sobre un dominio de investigación y tienden a reducir el sobre ajuste de los modelos.

5. Responsabilidad social y ética del uso de tests

Como se indicaba anteriormente, los tests son instrumentos de medida que se pueden utilizar bien o mal (Muñiz y Fernández, 2000) y en España parece existir cierta cultura de desensibilización con relación al uso responsable de tests (Rojas, 2002). Por consiguiente, no estarán de más unas breves notas que enfaticen la necesidad de que los usuarios de test se cercioren de la gran responsabilidad y del compromiso ético que conlleva el uso de este tipo de instrumentos.

Uno de los ejemplos históricos más sobresalientes que puede servir para ejemplificar un uso socialmente irresponsable de los tests lo podemos encontrar en el estudio de la inteligencia.

Sin profundizar en los antecedentes y condicionantes socio-culturales o antropológicos que coadyuvaron en el desarrollo de este fenómeno, podríamos referirnos a cómo el sesgo antropológico que llevó a discriminar racialmente a millones de personas estuvo también apoyado por el uso de test de inteligencia (Gould, 1981).

En esta página de la historia de los tests, según Golud (1981), el problema fue «cosificar» o «reificar» la inteligencia. Es decir, considerar que la inteligencia es una especie de ente físico y estable susceptible de ser medido inequívocamente. Más bien, los test de inteligencia, más allá de su utilidad científica-social, no dejan de ser instrumentos falibles construidos sobre modelos estadísticos que valoran el comportamiento de una persona en comparación con un grupo normativo de referencia. Cuando existe cierto sesgo o discrepancia cultural-antropológica entre el grupo normativo de referencia y la persona evaluada pueden producirse sesgos importantes que podrían tener serias consecuencias a nivel de convivencia social.

Otro aspecto relacionado con el uso ético y responsable de los test es el fenómeno de «etiquetado». Es decir, la tendencia pseudo-automática que experimentan algunos usuarios de tests de categorizar taxativamente la naturaleza de las personas con base en las puntuaciones generadas por un test. Por ejemplo, si un test que mide síntomas depresivos genera una puntuación muy alta, la persona evaluada es etiquetada como «depresiva».

Si el test es de personalidad y la puntuación es muy alta (incluso sobrepasando ciertos límites sugeridos en la literatura científica), la persona es «neurótica» o «narcisista». Nada más lejos de la realidad. Los tests no deberían de ser utilizados para encasillar a las personas de manera tajante en compartimentos

estancos de nuestro estante particular de las psicopatologías o categorías psicológicas. Proceder de este modo puede ser pernicioso para la persona evaluada y para su entorno social.

Un ejemplo derivado de este tipo de etiquetaje indiscriminado (como no, alentado algunas veces por el uso de tests) lo podemos encontrar en la elevada tasa de diagnósticos de trastornos autistas que recientemente se están observando en el ámbito de la psiquiatría y psicología infantil (Batstra et al., 2012; Frances, 2013a, 2013b).

Los mencionados anteriormente son sólo dos ejemplos de los malos usos que pueden darse a los tests psicológicos. Por la repercusión social que generan son llamativos o extremos y, en cierto modo, fáciles de detectar. Sin embargo, existen otros muchos aspectos que han de ser tenidos en cuenta para que la utilización de tests produzca beneficios para las personas evaluadas y para la sociedad en su conjunto.

El trabajo de Lyman (1977) apunta a este sentido y sugiere algunos aspectos que podrían tenerse en cuenta para que la interpretación de las puntuaciones que generan los tests estén dentro de un cauce sensato, éticamente aceptable y socialmente responsable.

Por su parte, el Consejo General de Colegios Oficiales de Psicólgos aborda el asunto de la responsabilidad en el uso de tests dentro del Código Deontológico del Psicólogo. Por ejemplo, en los artículos 40 y 46 del citado documento se alude explícitamente a cómo se ha de manejar responsablemente la información obtenida por medio de tests. Del mismo modo, la Asociación Americana de Psicología (APA, 2003/2010), un referente mundial para la psicología, también se hace eco de la necesidad de un buen uso de los tests. Además de alentar a la utilización de tests fiables y válidos la APA alude en el punto 9.11 a la gestión responsable de las puntuaciones generadas por los tests psicológicos.

Por su parte, los puntos 9.08 y 9.09(c) alertan sobre la necesidad de interpretar con cautela las puntuaciones que producen los tests psicológicos.

El ejemplo que venimos desglosando en este manuscrito es uno que requiere una gran dosis de responsabilidad social. En primer lugar, la violencia en la pareja es un fenómeno que recibe gran atención mediática en nuestros días y que, mayoritariamente, asume la forma de violencia del hombre hacia la mujer.

Supongamos que conseguimos diseñar un test cuyas puntuaciones son válidas y fiables para cribar a un conjunto de hombres acusados de haber perpetrado acciones violentas contra sus parejas mujeres. Supongamos que hemos encontrado, utilizando la tecnología psicométrica contemporánea a nuestro alcance, un umbral que nos permite predecir si un hombre agredirá violentamente, o no, a su pareja mujer en el futuro. Esta herramienta nos permitiría valorar y emitir informes forenses sobre el riesgo estimado de que un hombre agrediese a su pareja mujer. Dicho de otro modo, con este test podríamos clasificar a los hombres como agresores o como no-agresores potenciales.

Este procedimiento podría categorizarse como un fenómeno de “etiquetaje”. Es decir, sería como poner una etiqueta a cada persona en función del riesgo percibido de que cometa un acto delictivo sobre su pareja mujer. Sin embargo, pese a que hayamos desarrollado un test fiable y válido siguiendo las directrices más escrupulosas posibles, los test no están carentes de error. Son falibles y, como en este caso, los efectos de etiquetado pueden ser devastadores para las personas. Por ejemplo, que se tilde a un hombre como agresor de una mujer puede ser deletéreo para su estado anímico y social. Además, también pueden producirse efectos indeseados en las personas cuyo riesgo de agresión es moderado o moderadamente alto.

En estos casos, los hombres podrían “apegarse” en exceso a la etiqueta colocada y

comportarse en esa dirección de manera artificial o inducida.

Por tanto, cuando se lleva a cabo el diseño de tests se ha de tener especial sensibilidad a las consecuencias sociales que éstos pueden generar. Sobre todo, cuando, como en el caso ficticio que nos incumbe, el constructo estudiado es especialmente sensible, tabú o implica consecuencias sociales destacables.

6. Conclusiones

Como hemos visto a lo largo de este trabajo, el desarrollo y la utilización de tests en el ámbito de la investigación criminológica requiere que se adopten ciertos modos de trabajar que garanticen la calidad técnica y científica de todo el proceso. Dado que los tests son instrumentos de medida que proporcionan información útil y auténtica que puede guiar la toma de decisiones, se recomienda que se adopten modos de trabajar genuinamente sistemáticos y metódicos con el ánimo de optimizar el uso responsable y ético de este tipo de instrumentos de medida.

Se sugiere que se preste especial atención al proceso de desarrollo del test ya que mucho de lo que se hará después (por ejemplo, la validación del mismo) está condicionado por esta fase. Se advierte que, cuando se lleven a cabo adaptaciones de ítems, se sigan las directrices oportunas ya que esto mejorará la calidad del instrumento generado. Se recomienda también llevar a cabo estudios piloto que contengan análisis cuantitativos y cualitativos del funcionamiento del test para poder mejorar los aspectos en los que se detecten debilidades.

Se ha dedicado algún espacio a tratar el tema de la validez y la fiabilidad de los tests desde la óptica de la Teoría Clásica de Test (por ser la más común mente utilizada en la investigación aplicada), dado que estos dos son los elementos clave que justifican que el instrumento de medida goza de calidad científica suficiente para realizar interpretaciones útiles de las puntuaciones que generan.

En este trabajo también se han facilitado algunas orientaciones relativas al abordaje del análisis de datos que se puede llevar a cabo utilizando las puntuaciones de los tests y se ha terminado dando unas pequeñas pinceladas sobre la responsabilidad social y el compromiso ético que aparece indisolublemente asociado al uso de tests.

Esperamos que este trabajo pueda servir a la comunidad científica en este ámbito de conocimiento para utilizar responsable y éticamente los tests en aras de ahondar en el conocimiento de los fenómenos que son estudiados desde esta óptica.

Referencias

ANDERSON, Ronald D., y VASTAG, Gyulia (2004). "Causal modeling alternatives in operations research: Overview and application". *European Journal of Operational Research*, *156*, 92-109. https://doi.org/10.1016/S0377-2217(02)00904-9

BARTOL, Curt R., y BARTOL, Anne M. (2017). *Comportamiento Criminal. Una perspectiva psicológica*. Pearson.

BATSTRA, Laura, HADDERS-ALGRA, Mijna, NIEWEG, Edo, VAN TOL, Donald, PIJL, Sip Jan, y FRANCES, Allen (2012). "Childhood emotional and behavioral problems: reducing overdiagnosis without risking undertreatment". *Developmental Medicine and Child Neurology*, *54*, 492-494. https://doi.org/10.1111/j.1469-8749.2011.04176.x

COOK, David A., y BECKMAN, Thomas J. (2006). "Current concepts in validity and reliability for psychometrics instruments: theory and application". *The American Journal of Medicine*, *119*, 166e7-166e16. https://doi.org/10.1016/j.amjmed.2005.10.036

CROCKER, Linda, y ALGINA, James (1986). *Introduction to classical and modern test theory*. Holt, Rinehart y Winston.

DUNN, Thomas J., BAGULEY, Thom, y BRUNSDEN, Vivienne (2014). "From alpha to omega: A practical solution to the

pervasive problem of internal consistency estimation". *British Journal of Psychology*, *105*, 399-412. https://doi.org/10.1111/bjop.12046

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

ELOSUA, Paula, y ZUMBO, Bruno D. (2008). "Coeficientes de fiabilidad para escalas de respuesta categórica ordenada". *Psicothema*, *20*, 896-901.

FERNÁNDEZ-ABELLÁN, Antonio (Productor), y GÓMEZ, Bernardo (Realización). (2002). *Historia de la estadística* [Documental]. Universidad Nacional de Educación a Distancia.

FRANCES, Allen (2013a). The new crisis of confidence in psychiatric diagnosis. "*Annals of Internal Medicine"*, *59*, 221-222. https://doi.org/10.7326/0003-4819-159-3-201308060-00655

FRANCES, Allen (2013b). "Past, present and future of psychiatric diagnosis". *World Psychiatry*, *12*, 111-112. https://doi.org/10.1002/wps.20027

GARCÍA, Juan (2000). *Adaptación del cuestionario de actitudes legales para la definición de perfiles psicosociales en la selección de jurados*. Servicio de Publicaciones de la Universidad de Almería.

GOULD, Stephen Jay (1981). *The mismeasure of man*. Norton. https://doi.org/10.2307/2801521

HAMBLETON, Ronald K. (1996). Advances in assessment models, methods, and practices. En Berliner, David C. y Calfee, Robert C. (Eds.), *Handbook of educational psychology* (pp. 899-925). Macmillan.

HOTHERSALL, David (1997). *Historia de la Psicología*. McGraw-Hill.

JAMES, Lawrence R. (1982). "Aggregation bias in estimates of perceptual agreement". *Journal of Applied Psychology*, *67*, 219-229. https://doi.org/10.1037/0021-9010.67.2.219

JAMES, Lawrence R., DEMAREE, Robert G., y Wolf, Gerrit (1984). "Estimating within-group interrater reliability with and without response bias". *Journal of Applied Psychology*, *69*, 1984, 85-98. https://doi.org/10.1037/0021-9010.69.1.85

KELLEY, Ken, y CHENG, Ying (2012). "Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments". *Methodology*, *8*(2), 39-50. https://doi.org/10.1027/1614-2241/a000036

KROPP, P. Randal, y HART, Stephen, D. (2000). "Law and Human Behavior, Vol. 24, No. 1, 2000The Spousal Assault Risk Assessment (SARA) Guide:Reliability and Validity in Adult Male Offenders". *Law and Human Behavior*, *24*, 101-118. https://doi.org/10.1023/A:1005430904495

LARA, Mario R., y MOLINA, Jesús (2014). "Impacto de una intervención psico-educativa en una situación de violencia entre escolares (bullying)". *Ciencias Jurídicas y Victimológicas, 2*, 317-338.

LEBRETON, James M., y SENTER, Jenell L. (2008). "Answers to 20 questions about interrater reliability and interrater agreement". *Organizational Research Methods*, *11*, 815-852. https://doi.org/10.1177/1094428106296642

LEÓN, Orfelio G. (1994). *Análisis de decisiones. Técnicas situacionales aplicables a directivos y profesionales*. McGraw-Hill.

LEÓN, Orfelio G., y MONTERO, Ignacio (2003). *Métodos de Investigación en Psicología y Educación* (3ª ed.). McGraw-Hill.

LÓPEZ PUGA, Jorge (2013). *Psicometría esencial*. Murcia: Universidad Católica San Antonio.

LÓPEZ PUGA, Jorge, y GARCÍA, Juan (2011). *Utilidad de las redes bayesianas en psicología*. Editorial Universidad de Almería.

LYMAN, Howard B. (1977). *Las puntuaciones de los tests y sus significados*. Madrid: Manual Moderno.

MARTÍN, María Concepción (2004). "Diseño y validación de cuestionarios". *Matronas Profesión*, *5*(17), 23-29.

MUÑIZ, José (1992). *Teoría clásica de los tests*. Pirámide.

MUÑIZ, José (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.

MUÑIZ, José (1998). "La medición de lo psicológico". *Psicothema, 10*, 1-21.

MUÑIZ, José (2010). "La teoría de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems". *Papeles del Psicólogo*, *31*, 57-66.

MUÑIZ, José, ELOSUA, Paula, y HAMBLETON, Ronald K. (2013). "Directrices para la traducción y adaptación de los tests: segunda edición". *Psicothema*, *25*, 151-157. https://doi.org/10.7334/psicothema2013.24

MUÑIZ, José, y FERNÁNDEZ, José Ramón (2000). "Utilización de los tests en España". *Papeles del Psicólogo*, *76*, 41-49. https://doi.org/10.23923/pap.psicol2020.2921

MUÑIZ, José, y FONSECA-PEDRERO, Eduardo (2008). "Construcción de instrumentos de medida para la evaluación universitaria". *Revista de Investigación en Educación*, *5*, 13-25.

MUÑIZ, José, y HAMBLETON, Ronald K. (1996). "Directrices para la traducción y adaptación de los tests". *Papeles del Psicólogo*, *66*(1), 63-70.

PAGANO, Robert R. (1999). *Estadística para las ciencias del comportamiento* (5ª ed.). Thomson. (Trabajo original publicado en 1998)

PÉREZ, Meritxel, SÁIZ, Milagros, y SÁIZ, Dolores (2012). Aspectos generales de la evaluación en el ámbito jurídico-criminal. En Soria, Miguel Ángel y Sáiz, Dolores (Coord.), *Psicología Criminal* (pp. 431-464). Pearson.

PETER, J. Paul (1979). "Reliability: a review of psychometric basics and recent marketing practices". *Journal of Marketing Research*, *16*, 6-17. https://doi.org/10.2307/3150868

ROJAS, Antonio José (2002). "Reflexiones críticas sobre la investigación en medición mediante tests en España". *Apuntes de Psicología*, *20*, 81-96. https://doi.org/10.55414/q7vj1x21

RUIZ, Miguel A., PARDO, Antonio, y SAN MARTÍN, Rafael (2010). "Modelos de ecuaciones estructurales". *Papeles del Psicólogo*, *31*, 34-45.

SCHEIER, Michael F., y CARVER, Charles S. (1985). "Optimism, coping and health: Assessment and implications of generalized outcome expectancies". *Health Psychology*, *4*, 219-247. https://doi.org/10.1037//0278-6133.4.3.219

SOLANAS, Antonio, SALAFRANCA, Luis, FAUQUET, Jordi, y NÚÑEZ, María Isabel (2005). *Estadística descriptiva en ciencias del comportamiento*. Thomson.

STEVENS, Stanley Smith (1946, 7 de junio). "On the theory of scales of measurement". *Science*, *103*, 677-680. https://doi.org/10.1126/science.103.2684.677

SUEN, Hoi, y MCCLELLAND, Susan. (2003). Test item construction techniques and principles. En Huang, N. (Ed.), *Encyclopedia of vocational and technological education* (pp. 777-798). Taipei: ROC Ministry of Education.