Validez

Luis Manuel Lozano
Jaume Turbany

PID_00198629

Universitat Oberta de Catalunya

Índice

Introducción

Cuando un psicólogo decide aplicar un cuestionario es para alcanzar un objetivo determinado. Para ello, debe asegurarse de que el cuestionario que va a usar posee unas adecuadas propiedades psicométricas. Entre ellas cabe destacar la que vamos a tratar en este módulo: la validez.

La validez es uno de los aspectos más importantes, quizá el que más, tanto en la elaboración como en la evaluación de cuestionarios psicológicos. A fin de cuentas, se trata de comprobar que la utilización del test está siendo correcta y que los objetivos que desea alcanzar el psicólogo que lo utiliza son factibles.

En el apartado "¿Qué es la validez?" se hace un breve recorrido histórico sobre dicho concepto. Como se puede observar, es un concepto que ha estado evolucionando (y aún lo está) hasta que se alcanza la idea que actualmente está en vigor. Dicho concepto se define oficialmente en los estándares publicados en 1999 conjuntamente por la American Educational Research Methods (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME). Estas entidades defienden que se pueden agrupar los indicios de validez de un test en cinco apartados: evidencia basada en la validez de contenido, basada en el proceso de respuesta, basada en la estructura interna del cuestionario, basada en la relación con otras variables y basada en las consecuencias de la evaluación.

En los siguientes apartados se trata cada uno de los indicios de validez previamente mencionados y se buscan estrategias para poder obtener dichos indicios (a excepción del apartado de las consecuencias de la evaluación, en el que solo se tratan las consecuencias que se pueden esperar de la aplicación de un cuestionario).

En el último apartado, "Factores que afectan a la validez", se tratan diferentes aspectos que afectan a algunas de las técnicas expuestas con anterioridad para determinar los diferentes indicios de validez.

1. Qué es la validez

1.1. Definición

Para comprender el concepto de validez es necesario realizar un pequeño estudio de la evolución histórica que ha sufrido dicho concepto.

La utilización de cuestionarios se vio impulsada por la primera y segunda guerras mundiales. Durante esos momentos se tuvo la necesidad de incorporar al ejército a la población civil, destinándola al puesto más adecuado. Tras rellenar los cuestionarios se comprobaba en el campo de entrenamiento si los sujetos rendían satisfactoriamente o no en el puesto al que se les había destinado. Dado que en primer lugar se hacía la medición y posteriormente se evaluaba el éxito, se hablaba de validez predictiva. Es decir, un test posee **validez predictiva** si sirve para predecir el comportamiento en un constructo que será evaluado posteriormente a la aplicación del cuestionario.

Más tarde se trató de evaluar la relación existente entre las características de las personas que realizaban un trabajo y su éxito en él. De este modo, se trataba de conocer qué características podrían predecir el éxito laboral y buscarlas cuando se realizaba una selección de personal. Dado que el estudio se realizaba sobre personas que ya tenían el puesto y se valoraba su ejecución, se hablaba de validez concurrente, ya que ambas mediciones se hacían a la vez. Es decir, un test posee **validez concurrente** si sirve para predecir el comportamiento en un constructo que es evaluado simultáneamente a la aplicación del cuestionario.

Como se puede observar, inicialmente los tests eran exclusivamente empleados para predecir. Así pues, en un comienzo, se consideraba que un test era válido si servía para predecir alguna variable de interés, denominada **criterio** (Guilford, 1946).

Por lo tanto, se conceptualiza la validez como correlación entre el cuestionario y el criterio de interés (ya sea evaluado con posterioridad o simultáneamente a la aplicación del cuestionario). Así pues, se considera que un test es válido para evaluar cualquier aspecto con el que correlacione (Bingham, 1937; Guilford, 1946; entre otros).

Uno de los problemas de la conceptualización de la validez como correlación es el hecho de que hay que encontrar una medida del criterio adecuada, es decir, se necesitan datos del criterio que hayan sido obtenidos de una manera fiable y válida. Por tanto, si ya se posee una medida válida del criterio, ¿para qué se necesita aplicar un cuestionario?

Otro problema de esta conceptualización es que dejaba fuera a un gran número de tests educativos. En estos no se trata de predecir la conducta, se trata de comprobar cuánto se ha aprendido después de un periodo de formación. En estos cuestionarios la puntuación obtenida es un indicador de lo que el test pretende evaluar (conocimiento en matemáticas, en inglés, etc.) y no un predictor de criterios distintos del test. Desde esta perspectiva, la validez hace referencia a que los ítems que componen el cuestionario sean representativos de aquello que se pretende evaluar. A este concepto se le denominó **validez de contenido** (Anastasi, 1954).

Por otro lado, a lo largo de los años treinta se produce un auge de las teorías que tratan de conocer la estructura factorial de la inteligencia. Con estas teorías comienza a conceptualizarse un test como válido cuando representa de manera fidedigna el constructo psicológico que pretende medir, así como las relaciones esperadas entre los diferentes constructos. De este modo nace la validez de constructo (Cronbach y Meehl, 1955). Las técnicas estadísticas empleadas para poder comprobar dicha validez son, tradicionalmente, el análisis factorial exploratorio y las matrices multirrasgo-multimétodo (Campbell y Fiske, 1959), y más recientemente el análisis factorial confirmatorio. Por ejemplo, si se emplea un test que evalúa la triada cognitiva desde el modelo cognitivo de depresión de Beck (Beck, Rush, Shawn y Emery, 1979) (pensamientos sobre mí mismo, pensamientos sobre el mundo y pensamientos sobre el futuro), el cuestionario tendrá **validez de constructo** si evalúa las tres dimensiones y estas tienen las relaciones que se esperan con, por ejemplo, ansiedad.

Hasta los años ochenta se podía hablar de validez predictiva, validez concurrente, validez de contenido y validez de constructo de un cuestionario, si bien las dos primeras en los estándares de los tests y manuales educativos y psicológicos publicados por la APA, AERA y NCME en 1966 y 1974 se englobaban como **validez de criterio**.

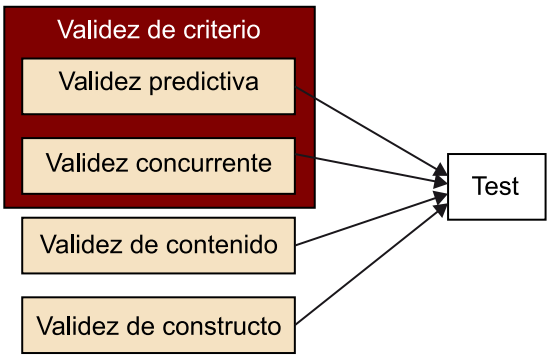

Figura 1

Posteriormente, Cronbach (1971) puntualizó que en un test que pretende medir un rasgo de personalidad no existe solo un criterio relevante que predecir, ni un contenido que muestrear (validez predictiva y de contenido respectivamente). Se dispone, por el contrario, de una teoría acerca del rasgo y de sus relaciones con otros constructos y variables (validez de constructo). Si se hi-

potetiza que la puntuación del test es una manifestación válida del atributo, se puede contrastar la asunción analizando sus relaciones con otras variables. Por tanto, comenzó a existir una tendencia en la que consideraban la validez como algo unitario, siendo la validez de constructo la científicamente más admisible y estando la validez de criterio y de contenido incluidas en esta (Messick, 1989). Así pues, se impone la concepción de que la validación de constructo constituye un marco integral para obtener pruebas de la validez incluyendo las procedentes de la validación de criterio y de contenido. De hecho, deja de hablarse de las diferentes categorías de validez para comenzar a hablar de diferentes evidencias implicadas en los tres tipos tradicionales de validez (criterio, contenido y constructo).

Dado que tanto el estudio de la estructura del constructo como las relaciones de este con otros constructos pasa a ser considerado la principal forma de validez, este proceso puede concebirse como un caso particular de la contrastación de las teorías científicas mediante el método hipotético-deductivo (Prieto y Delgado, 2010).

Notad que en estos momentos, a mediados de los años ochenta, existe un cambio muy relevante: mientras que al comienzo se conceptualiza la validez como una propiedad inherente al test, después se pasa a concebir que lo que realmente se valida no es el test en sí, sino las inferencias que se realizan a partir de este. Por ello, el responsable de asegurar la validez ya no es solo el constructor del test, sino que también lo es el usuario que emplea dicho cuestionario para una finalidad determinada. En muchas ocasiones los problemas que un cuestionario posea en lo referente a la validez se deben no al diseño del cuestionario sino a la utilización que se hace de este.

Actualmente, en la última edición hasta el momento de los *Standards for educational and psychological testing* (AERA, APA y NCME, 1999), muy influenciados por el capítulo escrito por Messick (1989) y el libro de Shepard, Camilli, Linn y Bohrnstedt (1993), se defiende que la validez hace referencia al **grado en el que la evidencia empírica y la teoría apoyan la interpretación de las puntuaciones de los tests relacionada con un uso específico**. Como se puede apreciar, se concibe la validez como un concepto unitario. Para comprobar la validez se debe atender a cinco evidencias de esta:

- **El contenido de test**: Los ítems que constituyen el test son relevantes y representativos del constructo psicológico que se desea medir.

- **El proceso de respuesta**: El proceso que siguen las personas al contestar al test permite extraer respuestas indicadoras de lo que se quiere evaluar.

- **La estructura interna**: Las relaciones de los ítems entre sí son congruentes con el modelo teórico empleado a la hora de definir el constructo que evaluar.

- **La relación con otras variables**: Las relaciones que se establecen entre el constructo que se evalúa y otros constructos son las esperadas según el marco teórico en el que se haya definido el constructo que evaluar.

- **Las consecuencias de la aplicación del cuestionario**: Las consecuencias tanto positivas como negativas que se extraen al emplear un test son las previstas.

Como breve resumen de lo expuesto anteriormente se presenta la siguiente tabla, en la que se puede apreciar la evolución del concepto en los diferentes estándares publicados por la APA.

Tabla 1

Edición	Validez
1954	Constructo, concurrente, predictiva, contenido
1966	Criterio, constructo, contenido
1974	Criterio, constructo, contenido
1985	Unitaria (pero mantienen criterio, constructo y contenido)
1999	Unitaria: 5 fuentes de evidencia

1.2. Importancia de la validez

El concepto de validez es central en psicometría. Tal y como se comentó anteriormente, para comprobar la validez se deben acumular evidencias que proporcionen una base científica para interpretar las puntuaciones de un cuestionario de manera adecuada. Por ello, lo que realmente se valida no es el cuestionario en sí, sino las interpretaciones que se hacen a partir de él. Por tanto, no se puede defender que un test sea válido o que por el contrario carezca de validez. Un test puede ser adecuado para un propósito pero no para otro.

> Si se aplica un cuestionario con el que se pretende medir autoestima, las respuestas pueden ser empleadas con diferentes fines (conocer el nivel de autoestima de una persona para saber si es un problema que tratar en terapia, en selección de personal, como investigación sobre el propio constructo, etc.). Para poder usar el cuestionario con una finalidad determinada, se deben acumular evidencias que indiquen que el uso es correcto ("evidencias de validez"). En caso contrario, se estaría haciendo un mal uso de los tests, principales herramientas en el trabajo psicológico, y las conclusiones que se extrajeran de ellos no serían correctas. En el ejemplo anterior no se sabría si es un aspecto sobre el que se debe intervenir terapéuticamente, no se sabría si la persona seleccionada realmente tiene la autoestima que se desea o no se sabe si realmente se está midiendo autoestima.

Para poder realizar correctamente el trabajo como psicólogos, se debe saber si las conclusiones que se extraen a partir de los tests empleados son adecuadas, ya que en caso contrario se corre el riesgo de no saber exactamente qué se está evaluando o si esa medición realmente es útil para el propósito del psicólogo.

2. Evidencia de validez basada en el contenido

2.1. Concepto

Muchas de las inferencias y asunciones que se derivan de la interpretación de las puntuaciones en un test son más fácilmente evaluables si se examinan los procedimientos empleados para generar las puntuaciones. Por ejemplo, si se quiere inferir a partir de las puntuaciones en un test sobre determinada conducta o constructo psicológico, es de esperar que los ítems que componen el cuestionario sean tanto **relevantes** (que la información que se pregunta esté directamente relacionada con lo que se pretende medir), como **representativos** (las cuestiones que se realicen deben ser una muestra adecuada de todo lo que se pretende medir) de la conducta (Kane, 2006).

La evidencia de la validez de contenido hace referencia a la relación que existe entre los ítems que componen el test y lo que se pretende evaluar con él, prestando atención tanto a la relevancia como a la representatividad de los ítems. Este tipo de evidencia se recoge principalmente en el momento de la elaboración del test.

Supongamos que se desea elaborar un test para evaluar la personalidad. En este caso, se decide trabajar dentro del marco teórico de los cinco factores de la personalidad (extraversión, apertura, responsabilidad, amabilidad y neuroticismo). Dado que se trata de un test que se va a emplear en una selección de personal concreta, solo interesan las dimensiones de responsabilidad (a), amabilidad (b) y neuroticismo (c). En este ejemplo el constructo es la personalidad que está compuesta por las cinco dimensiones. Las dos primeras, para los intereses del test que se está realizando, son información irrelevante. Las otras tres son el dominio que interesa evaluar. A partir de este dominio se construyen ítems destinados a evaluar la responsabilidad (a'), la amabilidad (b') y el neuroticismo (c'). Dichos ítems deben tener relación con el factor que pretenden medir, es decir, los ítems que evalúan responsabilidad están relacionados con la definición que existe en la comunidad científica sobre dicho factor (relevancia). Pero a su vez los ítems deben preguntar por la totalidad del dominio que evaluar (representatividad).

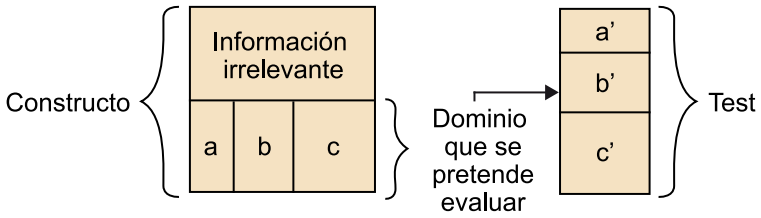

Figura 2

En las pruebas educativas, las evidencias de validez basada en el contenido son fundamentales. Si no se comprueba que el test es consistente con los objetivos curriculares perseguidos (relevancia), es decir, que está libre de material irrelevante y que el que está representa adecuadamente el dominio que se pretende evaluar (representatividad), la utilidad del test se verá seriamente afectada y, por tanto, las conclusiones que se obtengan serán erróneas. En estas situacio-

nes se suele recomendar, dado que el dominio que se quiere evaluar está claramente definido, emplear los diferentes métodos estadísticos de muestreo para obtener una muestra representativa de los contenidos que deben constituir el test (Muñiz, 2003).

El problema surge cuando no se dispone del dominio tan claramente definido.

> Por ejemplo, si se quiere realizar un test que evalúe la inteligencia, lo primero que se debe preguntar el constructor del cuestionario es: ¿qué es la conducta inteligente? En este caso, dado que no existe un dominio perfectamente definido, se deben buscar otras estrategias para obtener el indicador de la validez de contenido.

2.2. Procedimiento

En este apartado se presentará el procedimiento más habitual en la valoración de la evidencia basada en el contenido, si bien existen otros métodos menos empleados. Una recopilación de ellos se puede encontrar en Sireci (1998).

Si se quiere desarrollar un test, lo primero que se debe realizar es definir de manera operativa el dominio que evaluar. Tras realizar o aceptar una definición ya existente, se debe elaborar una **tabla de especificaciones**. Se trata de realizar una descripción detallada del test, determinar la proporción o el número de ítems que evaluarán cada contenido o habilidad del dominio que evaluar; el formato de los ítems y de las respuestas (AERA, APA y NCME, 1999) (usualmente en este paso también se determinan las propiedades psicométricas que se desea que tenga la prueba).

Tras realizar los ítems se debe acudir a un grupo de expertos en la materia, que harán las veces de jueces. Para evitar cualquier sesgo, dichos jueces no deben estar implicados en la elaboración del cuestionario. Estos deben analizar cada uno de los ítems valorando en qué medida son **representativos** y **relevantes** para evaluar el dominio de interés, tomando como definición de este la aportada por los autores del test.

Se puede defender que existen, por tanto, tres aspectos bien diferenciados que se deben tener en cuenta a la hora de comprobar las evidencias de la validez de contenido: la definición del dominio, la representación de los ítems que evalúan el dominio y su relevancia (Sireci, 1998).

Es recomendable que la valoración de los ítems la realice cada juez por separado para, de este modo, evitar posibles sesgos a la hora de responder. Una vez que se poseen las valoraciones de todos los expertos, se deben buscar aquellos ítems en los que haya concordancia, seleccionándolos para formar parte del cuestionario.

Por ejemplo, si 8 de los 10 jueces determinan que un ítem destinado a medir depresión realmente evalúa lo que pretende, dicho ítem tendrá un índice de congruencia de 0,8. Se suelen considerar adecuados aquellos ítems que poseen un índice de congruencia igual o superior a 0,7 (Sireci, 1998).

Los ítems en los que no haya acuerdo (que no alcancen un índice de congruencia de 0,7) no tienen por qué ser eliminados. Es recomendable que con estos ítems se realice un grupo de discusión con los expertos para que comenten las diferencias tratando de llegar a un punto de acuerdo para mejorar dichos ítems.

Este es el procedimiento más habitual a la hora de valorar los indicios de validez de contenido, si bien no está libre de críticas. El principal problema que se pone en la utilización de expertos es que estos son altamente competentes en el contenido que se evalúa, por lo que pueden pasar por alto un texto cuyo nivel no sea adecuado para la comprensión de los sujetos que hay que evaluar o que puede ser fácilmente malinterpretado. Es decir, aunque el experto nos puede proporcionar información muy relevante, lo que realmente importa es cómo percibe y reacciona ante el test o el ítem la persona que está respondiendo (Leighton, 2004).

2.3. Contenido sesgado

El uso de expertos para valorar tanto la relevancia como la representatividad de los ítems tiene como finalidad evitar que el cuestionario tenga contenidos sesgados. Se dice que el contenido de un test está sesgado si los ítems que lo componen evalúan aspectos no relevantes para el dominio (**sesgo por falta de relevancia**) o si no representan de manera adecuada todo el dominio que se pretende evaluar (**sesgo por falta de representatividad**). Como se puede comprobar, un test está sesgado si no cubre adecuadamente el dominio que pretende medir o si incluye cuestiones no necesarias para valorar correctamente el dominio.

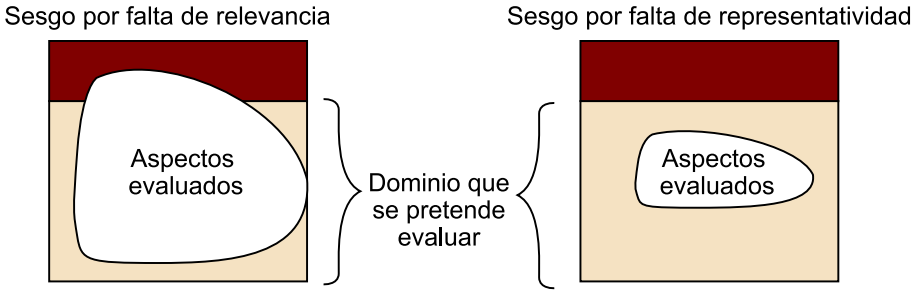

Figura 3

3. Evidencia de validez basada en el proceso de respuesta

3.1. Concepto

Este tipo de evidencia es un concepto que se introdujo como novedoso en los estándares publicados en 1999, si bien había sido previamente mencionado por algunos especialistas en la medida de lo psicológico, como Messick (1989). Los estándares describen este tipo de indicios como el ajuste entre el constructo evaluado y el proceso de respuesta realizada por las personas que responden al test.

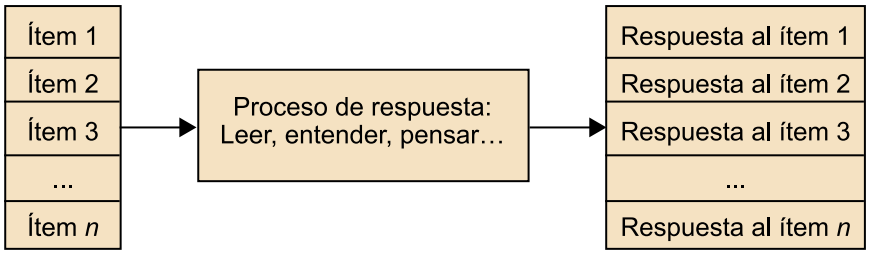

Figura 4

Por **proceso de respuesta** se entienden todas las conductas que se necesitan para poder contestar un ítem, como pueden ser leer las preguntas, comprenderlas, decidir la respuesta que se quiere dar y finalmente responder al ítem.

> Un ejemplo sobre este indicio de validez se puede encontrar en un examen de matemáticas que se realice a niños que están aprendiendo a sumar. Si el enunciado del ítem es: "3 + 2 = ", probablemente, si han adquirido el conocimiento necesario, pueden dar la respuesta correcta. Dicho ítem también puede tener un enunciado como: "Calcule el valor resultante de realizar una operación aditiva entre los valores 3 y 2". Evidentemente, un niño que esté aprendiendo a sumar puede responder al primer enunciado pero no al segundo (no tiene el vocabulario necesario, su capacidad lectora no le permitirá comprender la pregunta, etc.). Como resultado del primer enunciado se concluirá que ya tiene adquirido cierto nivel del dominio evaluado, pero con el segundo se concluirá que no. Es decir, dado que el segundo enunciado carece de validez de respuesta, llevará a los evaluadores a conclusiones erróneas sobre el nivel de habilidad del niño a la hora de realizar sumas.

A la hora de responder a un test se deben combinar las características de las preguntas realizadas, las de las respuestas que se pueden dar y las de la persona que responde. Por ello, existen diferentes factores que pueden afectar a la respuesta:

- **Factores relacionados con los ítems**. En este apartado se deben tener en cuenta varios factores.
  - Contenido de los ítems. Se debe asegurar que el contenido es adecuado a la población que se quiere evaluar. Por ejemplo, no se puede usar un

test para evaluar depresión infantil si este fue construido para hacerlo en adultos. Si se hace, se pueden encontrar preguntas del tipo "¿Ha sentido cambios en su deseo sexual?", que evidentemente son inapropiadas para evaluar a niños.

– Redacción de los ítems. El lenguaje empleado para realizar el ítem no debe superar la capacidad comprensiva de las personas que van a responder. Un ejemplo de esto puede observarse en el ejemplo de la suma expuesto anteriormente.

– Validez aparente del ítem. Cuando se evalúan conocimientos es deseable que las personas que responden al cuestionario piensen que es adecuado. Si se evalúa el conocimiento en psicometría mediante un test, es de esperar que los alumnos que rellenen el test piensen que las preguntas que se les realizan son adecuadas para medir el conocimiento en psicometría. Esto es lo que se denomina validez aparente. Por el contrario, en los tests de personalidad hay que tratar de que la persona que responde no sepa exactamente lo que se evalúa. De este modo se intenta evitar que responda lo que más le favorece o lo que piensa que se espera de él.

- **Factores relacionados con la respuesta a los ítems**.
  – El número de alternativas que se ofrezcan como respuesta. Los tests de actitudes suelen responderse en un formato tipo Likert. En estas escalas se les pide a las personas en qué grado están de acuerdo con la afirmación que se les presenta teniendo que responder en una escala donde, por ejemplo, 0 significa totalmente en desacuerdo y 5, totalmente de acuerdo. El problema surge cuando se emplea una escala que supera la capacidad discriminativa de las personas. En estudiantes universitarios una escala de 0 a 10 es perfectamente comprensible pero, por el contrario, esa misma escala empleada en personas sin estudios puede ser excesiva. Un universitario comprende perfectamente la diferencia entre un 4 y un 5, pero esa diferencia puede estar menos clara en una persona sin estudios, con lo que se introduce, de este modo, error en la evaluación.

  – Las instrucciones a la hora de rellenar el cuestionario. A la hora de rellenar un cuestionario las instrucciones que se den para hacerlo deben ser claras y comprensibles. Estas deben adaptarse al grupo que se desee evaluar para que el criterio empleado a la hora de responder esté claro.

- **Factores relacionados con las personas**. En este apartado entrarían todas las características personales de aquellos que van a responder al cuestionario (capacidad lectora, capacidad intelectual, capacidad discriminativa, estado emocional, etc.). Es necesario hacer especial mención a las situaciones en las que la persona está en un proceso de selección, ya que tratará

de dar una imagen distorsionada de sí misma, tratando de adaptarse a lo que piensa que el seleccionador busca.

3.2. Procedimiento

Aunque en los estándares de la APA (1999) aparece este indicio de validez, apenas aportan información sobre cómo determinar si un test tiene indicios de validez basados en el proceso de respuesta. Dentro de las alternativas que proponen se encuentran técnicas como:

- **La entrevista**. En ella se les preguntará a las personas que responden al test por las diferentes estrategias empleadas para responder a cada uno de los ítems. El conocimiento de dichas estrategias puede conducir incluso al enriquecimiento de la definición del constructo estudiado.

- **Técnicas de pensamiento en voz alta**. Se les pide a las personas que rellenen el cuestionario diciendo en voz alta los diferentes procesos por los que pasan a medida que deben contestar el test.

- **Entrevistas cognitivas**. Están diseñadas para comprender cómo las personas que responden a un test comprenden la pregunta, recuperan la información relevante para responder, evalúan la relevancia de lo recordado y responden a la pregunta. Empleando esta información se pueden identificar potenciales errores de respuesta así como patrones de interpretación de las preguntas. También puede aportar información sobre los factores socioculturales que afectan al modo de responder (Czaja y Blair, 1996).

4. Evidencia de validez basada en la estructura interna

4.1. Concepto

Para la elaboración de un test se utilizarán distintos ítems o preguntas; es posible que se considere que todos los ítems son igual de relevantes para medir la característica estudiada, en cuyo caso obtendremos una puntuación total del test a partir de la simple suma de las puntuaciones obtenidas por el sujeto en los diferentes ítems.

La situación puede no ser tan sencilla cuando suponemos que no todos los ítems tienen la misma importancia en la medida del constructo, y por tanto será necesario ponderar las puntuaciones de los ítems antes de proceder a la suma; hablamos en este caso de puntuaciones compuestas. En esta situación la estructura del test que tendremos que determinar es unidimensional, ya que suponemos que todos los ítems, aunque de distinta manera, contribuyen a la medida de un único aspecto de la variable criterio.

Un test también puede presentar una estructura interna multidimensional, esto es, que las diferentes preguntas no miden un solo aspecto o dimensión sino dos o más.

La técnica estadística del análisis factorial nos servirá para el estudio de la contribución de los diferentes ítems a un solo factor (estructura unidimensional) o a varios factores (estructura multidimensional).

La técnica del análisis factorial nos permitirá determinar k factores subyacentes, a partir de una serie p de puntuaciones determinadas por los ítems iniciales del test. La idea es la búsqueda de un modelo parsimonioso (simple) a partir de un conjunto complejo de datos.

A partir de los trabajos de Spearman a principios del siglo XX, y sobre todo de Thurstone en los años cuarenta, el análisis factorial se evidencia como una buena herramienta en psicología para tratar de identificar los factores que intervienen en la inteligencia. Thurstone propuso la utilización del análisis factorial para dar explicación a las correlaciones que observaba entre diferentes ítems de los tests de inteligencia. Así, el empleo de esta técnica le permitió la identificación y diferenciación de las capacidades espacial, verbal y numérica, como factores de la inteligencia.

El problema de esta técnica estriba en las dificultades del cálculo, sobre todo a partir de un número elevado de ítems (variables); sin embargo, el desarrollo y la popularización actual de los programas estadísticos han permitido la difusión de esta y otras técnicas de análisis de datos multivariables.

4.2. Procedimientos

El término de *análisis factorial* no es un concepto unitario, sino que reúne diferentes procedimientos que persiguen la reducción inicial de múltiples variables en un menor número de factores. En procesos exploratorios, la más utilizada es la técnica del análisis en componentes principales, pero existen otras formas de extracción de los factores o componentes.

4.2.1. Unidimensionalidad

El análisis en componentes principales parte inicialmente de la matriz de correlaciones entre las diferentes variables. Disponemos de la matriz de correlaciones obtenida a partir de la administración de un test a una muestra de 52 individuos, y compuesto por ocho preguntas o ítems que intentan medir un único constructo, en este caso la autoestima de los sujetos.

Tabla 2. *Correlation matrix*

	ítem1	ítem2	ítem3	ítem4	ítem5	ítem6	ítem7	ítem8
ítem1	1,000	,447	,411	,444	,533	,337	,365	,442
ítem2	,447	1,000	,561	,662	,707	,528	,333	,522
ítem3	,411	,561	1,000	,665	,699	,462	,572	,540
ítem4	,444	,662	,665	1,000	,682	,518	,560	,564
ítem5	,533	,707	,699	,682	1,000	,467	,592	,488
ítem6	,337	,528	,462	,518	,467	1,000	,424	,418
ítem7	,365	,333	,572	,560	,592	,424	1,000	,422
ítem8	,442	,522	,540	,564	,488	,418	,422	1,000

La matriz de correlaciones presenta, en esta situación, una distribución de valores suficientemente uniforme, y no se detectan en este caso agrupaciones de variables con correlaciones altas entre ellas y bajas con las demás.

La varianza de cada variable (ítem) es posible descomponerla en tres fuentes de variación: la varianza factorial común, que comparten las variables en común, la varianza específica, o no compartida por otras variables, y la varianza del error. La varianza común, también denominada comunalidad (h^2), interesa que sea suficientemente alta una vez que hemos seleccionado los factores relevantes.

En el inicio del análisis la comunalidad de las variables es la unidad; después del análisis, cuanto más próxima esté a uno, más relación habrá con el factor o factores extraídos.

Tabla 3. Comunalidades

	Inicial	Extracción
ítem1	1,000	,410
ítem2	1,000	,628
ítem3	1,000	,670
ítem4	1,000	,722
ítem5	1,000	,743
ítem6	1,000	,455
ítem7	1,000	,489
ítem8	1,000	,518

Método de extracción: análisis de componentes principales

La comunalidad de un ítem *j* viene representada por:

$$h_j^2 = a_j^2 + b_j^2 + \cdots + k_j^2$$

Donde a_j^2, b_j^2,..., k_j^2 representan el cuadrado de los coeficientes de saturación de cada ítem con cada factor *A*, *B*,..., *K*, extraídos, siendo el coeficiente de saturación la correlación de cada ítem con los factores extraídos.

En el ejemplo, las variables que presentan mayor comunalidad son los ítems 4 y 5.

Tabla 4. Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,636	57,947	57,947	4,636	57,947	57,947
2	,718	8,969	66,916			
3	,681	8,513	75,429			
4	,572	7,155	82,584			
5	,558	6,969	89,553			
6	,344	4,303	93,856			
7	,304	3,803	97,659			

Método de extracción: análisis de componentes principales

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
8	,187	2,341	100,000			

Método de extracción: análisis de componentes principales

A partir de *p* variables, el análisis factorial extrae el mismo número de factores, no relacionados entre sí, y cada uno de los factores se define como combinación lineal de las *p* variables originales. Estos *p* factores se ordenan por orden de importancia; en efecto, el primer componente o factor es el mejor resumen de las relaciones lineales que presentan los datos. El segundo factor es la segunda mejor combinación de las variables, con la condición de que sea ortogonal (sin relación) con el primero, y así sucesivamente con el resto de los *p* factores o componentes.

Un criterio muy extendido, el más utilizado a la hora de extraer los componentes es el del valor propio o autovalor superior a 1. Otro sería retener los factores necesarios hasta conseguir un porcentaje adecuado de variabilidad explicada por los componentes.

El valor propio o autovalor (λ) se define como la suma de los cuadrados de las saturaciones o correlaciones de cada ítem con el componente en cuestión. Representa, por tanto, una medida de la variabilidad explicada en las variables por parte del componente o factor.

En la tabla de varianza total explicada, donde se desglosan los diferentes componentes, vemos que solo hay un componente con valor propio (4,63) superior a 1. Por tanto, confirmaría una estructura unidimensional del test, ya que todas las preguntas confluyen en un solo componente, identificable como el constructo subyacente que se pretende medir. En nuestro ejemplo, las diferentes preguntas de la escala elaborada contribuirían a la medida de un único constructo psicológico que identificaríamos con la autoestima.

Tabla 5. Matriz de componentes[a]

	Componente
	1
ítem1	,640
ítem2	,793
ítem3	,819
ítem4	,850
ítem5	,862

Método de extracción: análisis de componentes principales. a. 1 componentes extraídos

	Componente
	1
ítem6	,675
ítem7	,699
ítem8	,720

Método de extracción: análisis de componentes principales. a. 1 componentes extraídos

La matriz de componentes indica las correlaciones entre cada ítem con el componente, lo que hemos denominado anteriormente saturaciones (*factor loadings*). En el ejemplo vemos que los valores son altos, y además con poca fluctuación, lo cual indicaría que todos los ítems tienen similar importancia en la medida del constructo.

Con todos los indicadores mencionados podemos comprobar que la comunalidad de cada ítem, al haber seleccionado un solo factor, simplemente es el cuadrado de la saturación entre ítem y factor.

Así, para el ítem 1 la comunalidad final es $h_1^2 = (0{,}64)^2 = 0{,}41$. Un 41% de la variabilidad del primer ítem viene explicada por el componente.

El valor propio del componente 1 se obtiene de la suma de los cuadrados de las saturaciones de cada ítem con el factor.

Así, en el primer componente, $\lambda_1 = (0{,}64)^2 + (0{,}793)^2 + \ldots + (0{,}72)^2 = 4{,}63$.

Como tenemos 8 ítems, el máximo serían 8 componentes. Si hacemos el cociente 4,63 / 8 = 0,5795. Un 57,95% de la variabilidad total es explicada por el primer componente.

Una vez hemos extraído los componentes, dispondremos de la matriz de puntuaciones factoriales que nos proporciona las ponderaciones de cada variable para el cálculo de la puntuación de cada sujeto en los factores extraídos. Las puntuaciones factoriales (*factor scores*) para los datos individuales se calculan a partir de la matriz de coeficientes de puntuaciones factoriales

$$F_i = a_1 \cdot Z_1 + a_2 \cdot Z_2 + \cdots + a_p \cdot Z_p$$

Lectura de la fórmula

a_i: Coeficientes de ponderación de cada variable para cada factor.

Z_i: Puntuaciones tipificadas de los valores de cada variable, obtenidos por cada individuo.

Con los datos del ejemplo la matriz de ponderaciones:

Tabla 6. Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente
	1
ítem1	,138

Método de extracción: análisis de componentes principales

	Componente
	1
ítem2	,171
ítem3	,177
ítem4	,183
ítem5	,186
ítem6	,146
ítem7	,151
ítem8	,155

Método de extracción: análisis de componentes principales

Para cada sujeto es posible calcular una puntuación factorial del índice de autoestima:

$$F_1 = 0,138 \cdot Z_{ítem1} + 0,171 \cdot Z_{ítem2} + \cdots + 0,155 \cdot Z_{ítem8}$$

De todos modos, en este ejemplo se observa que las ponderaciones de todos los ítems son muy similares, por lo que la contribución de todas las preguntas es muy similar en la medida del constructo de interés; por tanto, sería adecuado optar por una puntuación simple, únicamente sumando las puntuaciones obtenidas en cada ítem, frente a una puntuación compuesta ponderando los valores de cada ítem.

4.2.2. Multidimensionalidad

En muchas ocasiones, aunque de entrada intentemos elaborar una escala para la medida de un solo constructo psicológico, es posible que después de la primera administración del test, en la prueba piloto, observemos que en realidad los ítems se agrupan entre ellos y afectan a diferentes constructos subyacentes.

Presentamos otro ejemplo en el que se ha realizado un cuestionario con la intención de medir las actitudes sobre ideas religiosas en una muestra de 870 sujetos. Los ítems se han identificado con el concepto principal que implicaba la pregunta. La matriz de correlaciones de Pearson entre los ítems se presenta a continuación:

Tabla 7. *Correlation matrix*

	Sent. vida	Religión	Obe-diencia	Más allá	Experi.	Inseguridad	Influencia	Independ.
Sentido de la vida	1,000	,295	,220	,253	,226	,134	,178	,027
Religión	,295	1,000	,440	,507	,243	,099	,241	,103
Obediencia	,220	,440	1,000	,339	,292	,063	,336	,054

	Sent. vida	Religión	Obe-diencia	Más allá	Experi.	Inseguridad	Influencia	Independ.
Más allá	,253	,507	,339	1,000	,309	,113	,278	,121
Experiencia	,226	,243	,292	,309	1,000	,078	,204	,049
Inseguridad	,134	,099	,063	,113	,078	1,000	,117	,169
Influencia	,178	,241	,336	,278	,204	,117	1,000	,102
Independencia	,027	,103	,054	,121	,049	,169	,102	1,000

Las correlaciones fluctúan en un rango similar entre las diferentes preguntas; quizá las que presentan correlaciones inferiores son las preguntas referidas a inseguridad e independencia.

Al aplicar el correspondiente análisis en componentes principales, y utilizando el criterio de valor propio superior a 1, para la extracción de los componentes, obtenemos el siguiente listado:

Tabla 8. Comunalidades

	Inicial	Extracción
Sentido de la vida	1,000	,279
Religión	1,000	,554
Obediencia	1,000	,513
Más allá	1,000	,525
Experiencia	1,000	,333
Inseguridad	1,000	,562
Influencia	1,000	,315
Independencia	1,000	,588

Método de extracción: análisis de componentes principales

Tabla 9. Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,554	31,926	31,926	2,554	31,926	31,926
2	1,115	13,936	45,862	1,115	13,936	45,862
3	,901	11,258	57,120			
4	,826	10,329	67,448			
5	,787	9,840	77,288			

Método de extracción: análisis de componentes principales

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
6	,739	9,234	86,522			
7	,632	7,906	94,428			
8	,446	5,572	100,000			

Método de extracción: análisis de componentes principales

Tabla 10. Matriz de componentes[a]

	Componente	
	1	2
Sentido de la vida	,526	-,046
Religión	,735	-,115
Obediencia	,685	-,208
Más allá	,723	-,056
Experiencia	,558	-,146
Inseguridad	,267	,701
Influencia	,560	,044
Independencia	,221	,734

Método de extracción: análisis de componentes principales. a. 2 componentes extraídos

De los ocho componentes posibles, solamente los dos primeros cumplen el criterio de autovalor superior a 1, aunque un tercer factor está a punto de llegar a este límite ($\lambda_3 = 0{,}901$). En todo caso, se extraen dos componentes.

La última de las tablas presentadas (matriz de componentes) nos muestra las saturaciones o correlaciones entre los diferentes ítems y los dos componentes.

Recordemos que la comunalidad de los ítems representa la variabilidad explicada del ítem por los factores extraídos. Así, en el ítem 1 (sentido de la vida):

$$h_1^2 = (0{,}526)^2 + (-0{,}046)^2 = 0{,}279$$

Explicaría un 27,9% de la variabilidad del ítem. Es la más baja de todos los ítems del test, quizá este ítem saturaría con un tercer componente.

El valor propio (autovalor) de cada componente se calcula con la suma de cuadrados de las correlaciones (saturaciones) entre ítems y componente.

Componente 1: $\lambda_1 = (0,526)^2 + (0,735)^2 + \ldots + (0,221)^2 = 2,554$

Componente 2: $\lambda_2 = (-0,046)^2 + (-0,115)^2 + \ldots + (0,734)^2 = 1,115$

El primer componente (2,554 / 8 = 0,319) explicaría un 31,9% de la variabilidad presentada por los ítems, mientras que el segundo (1,115 / 8 = 0,139) explicaría un 13,9%. Un total de un 45,8% de la varianza total es explicada por la combinación de los dos factores o componentes.

El análisis de la matriz de saturaciones nos permitirá intentar buscar interpretación a los dos componentes, en función de los ítems que correlacionen de forma más alta.

En el ejemplo observamos que el primer componente tiene altas saturaciones en los ítems 1, 2, 3, 4, 5, y 7; mientras que las correlaciones son bajas en los ítems 6 y 8. En el segundo componente ocurre justo al contrario: tiene altas correlaciones con los ítems 6 y 8, y en cambio bajas en los demás.

En ocasiones la solución final no presenta, a simple vista, una tan fácil interpretación; en estos casos es posible utilizar una rotación de los ejes para conseguir que las correlaciones sean fuertes en un eje o componente y bajas en los demás. Recordemos que los ejes son ortogonales y no relacionados entre ellos. Una de las técnicas matemáticas de rotación de los ejes es la rotación varimax, la más utilizada en procesos exploratorios, aunque los programas estadísticos incorporan otras, como las rotaciones quartimax, equimax, etc.

La matriz de saturaciones con la solución rotada con el método varimax, con los datos del ejemplo, quedaría como sigue:

Tabla 11. Matriz de componentes rotados[a]

	Componente	
	1	**2**
Sentido de la vida	,522	,080
Religión	,742	,062
Obediencia	,715	-,040
Más allá	,715	,117
Experiencia	,577	-,010
Inseguridad	,093	,744
Influencia	,533	,175
Independencia	,041	,765

Método de extracción: análisis de componentes principales. Método de rotación: normalización Varimax con Kaiser. a. La rotación ha convergido en 3 iteraciones.

Vemos cómo la solución rotada confirma en este caso la conclusión previa, seis de las preguntas saturan en el primer componente. Viendo los seis ítems que saturan este componente –sentido de la vida, religión, obediencia, más allá, experiencia, influencia–, podemos interpretar este componente como la medida de la actitud sobre ideas religiosas, que era el motivo inicial de la elaboración del test.

El segundo componente, se supone que no esperado inicialmente en la elaboración del cuestionario, satura los ítems inseguridad e independencia. Este segundo componente se puede interpretar como una medida de la actitud hacia la dependencia personal.

Al tratarse de solo dos componentes es posible representar gráficamente en un espacio bidimensional los dos ejes que representan los componentes y la situación de los ítems respecto a ellos, en función de las saturaciones obtenidas.

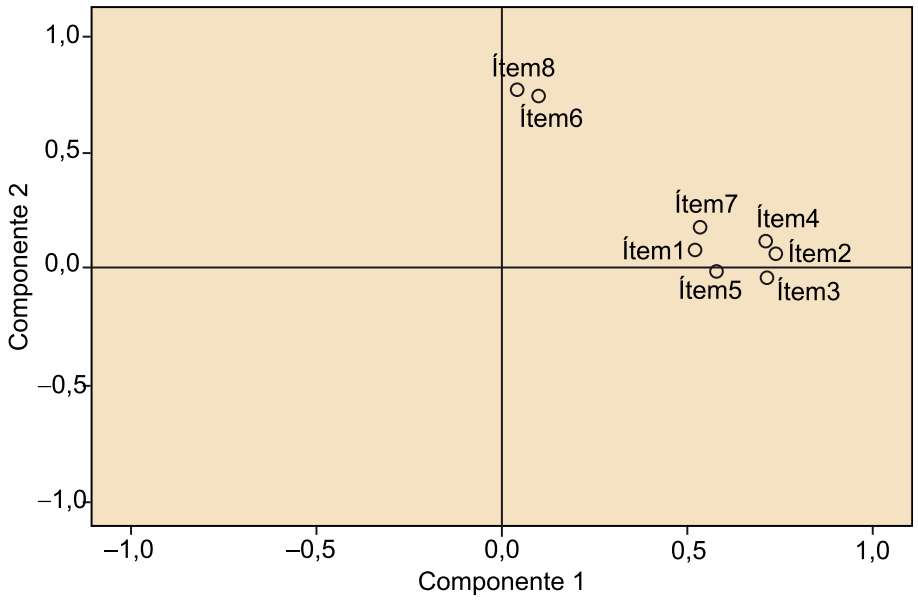

Figura 5. Gráfico de componentes en espacio rotado

El gráfico da una rápida información visual de la agrupación de los ítems en los dos componentes.

Una vez obtenidos los factores terminales, podremos calcular los valores o las puntuaciones factoriales de las dimensiones teóricas que hemos asociado a los respectivos factores, ideas religiosas (componente 1) y dependencia (componente 2).

Tabla 12. Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente	
	1	**2**
Sentido de la vida	,210	,009

Método de extracción: análisis de componentes principales. Método de rotación: normalización Varimax con Kaiser.

	Componente	
	1	**2**
Religión	,304	-,032
Obediencia	,305	-,118
Más allá	,287	,018
Experiencia	,243	-,076
Inseguridad	-,047	,635
Influencia	,204	,090
Independencia	-,072	,660

Método de extracción: análisis de componentes principales. Método de rotación: normalización Varimax con Kaiser.

Para cada sujeto *i* es posible calcular las puntuaciones factoriales de las nuevas variables ideas religiosas y dependencia:

$$Ideas\ religiosas_i = 0,21 \cdot Zi_{ítem1} + 0,304 \cdot Zi_{ítem2} + \cdots + 0,204 \cdot Zi_{ítem7} - 0,072 \cdot Zi_{ítem8}$$

$$Dependencia_i = 0,009 \cdot Zi_{ítem1} - 0,032 \cdot Zi_{ítem2} + \cdots + 0.,09 \cdot Zi_{ítem7} + 0,66 \cdot Zi_{ítem8}$$

Los diferentes paquetes estadísticos permiten el cálculo automático y la generación de estas nuevas variables generadas, y que representarían la medida de los constructos subyacentes.

5. Evidencia de validez basada en la relación con otras variables

5.1. Concepto

En el proceso de validación de una nueva prueba psicológica podemos ayudarnos de la existencia de otros instrumentos de medida del constructo de interés, que estén contrastados como válidos y fiables. En este proceso hablaremos de validez convergente o correlación entre puntuaciones del test con otras medidas del mismo constructo realizadas a partir de diferentes técnicas o indicadores.

Las diferentes técnicas estadísticas de relación entre variables nos servirán para determinar el coeficiente de validación entre las dos variables. Así, el más utilizado será el coeficiente de correlación de Pearson, en el caso de que las dos variables sean cuantitativas, pero también cualquiera de sus variaciones, como los coeficientes de Spearman, Biserial puntual, Biserial, phi, Tetracórica, etc., en función de cómo sean las dos variables que hay que relacionar.

Tabla 13

Variable A	Variable B	Coeficiente correlación
Numérica (intervalo o razón)	Numérica (intervalo o razón)	*r* de Pearson
Numérica (intervalo o razón)	Numérica (ordinal)	r_s Spearman o τ de Kendall
Numérica (ordinal)	Numérica (ordinal)	r_s Spearman o τ de Kendall
Cualitativa	Cualitativa	V de Cramer
Cualitativa (dicotómica)	Numérica (intervalo o razón)	r_b Biserial o r_{bp} Biserial puntual
Cualitativa (dicotómica)	Cualitativa (dicotómica)	Φ Phi o r_t Tretacórica

Para poder conocer el nivel de bienestar físico y psicológico en personas mayores, nos interesa validar un nuevo cuestionario que hemos elaborado para determinar el grado de independencia en las actividades básicas de la vida diaria (ABVD). A tal objeto, podemos utilizar algunas de las pruebas que ya existen en el mercado y que se encuentran suficientemente contrastadas. En una muestra de 300 sujetos mayores de 70 años, e ingresados en centros geriátricos, administramos la nueva prueba elaborada junto con la escala de medida de independencia funcional (FIM) (Keith, Granger, Hamilton y Sherwin, 1987) y la escala de grado de autonomía de Barthel (Mahoney y Barthel, 1965).

En la tabla siguiente se muestra la matriz de correlaciones de Pearson (datos simulados) entre las tres pruebas administradas:

Tabla 14

	ABVD	FIM	Barthel
ABVD	1		
FIM	0,69	1	
Barthel	0,67	0,77	1

Los valores de la correlación entre la nueva prueba (ABVD) y las escalas FIM y Barthel presentan valores suficientemente altos (0,69 y 0,67, respectivamente), lo cual indica una alta validez concurrente de la nueva prueba elaborada con las técnicas previas para la medida del grado de autonomía de las personas mayores analizadas.

5.2. Evidencia de decisión (sensibilidad y especificidad)

En situaciones en las que la prueba realizada tenga como objetivo el diagnóstico o la clasificación de los sujetos en dos grupos (diagnóstico negativo-diagnóstico positivo) hablaremos de la validez de decisión cuando comparemos esta nueva prueba con otro método de diagnóstico anterior suficientemente contrastado. En la validez de decisión podemos distinguir dos procesos: por una parte, la sensibilidad de la prueba, definida como la capacidad de esta en la detección de verdaderos positivos, y por otra parte, la especificidad, definida como la capacidad de determinación de diagnósticos negativos verdaderos.

Tabla 15

		Diagnóstico prueba inicial		
		Positivo	**Negativo**	**Total**
Diagnóstico nueva prueba	**Positivo**	Decisión correcta (f_{11})	Falso positivo (f_{12})	$f_{1.}$
	Negativo	Falso negativo (f_{21})	Decisión correcta (f_{22})	$f_{2.}$
	Total	$f_{.1}$	$f_{.2}$	n

Una medida del acuerdo logrado a través de las dos pruebas diagnósticas consistirá en calcular el porcentaje de acuerdo (P_c) entre ambas técnicas a partir de la razón entre la suma de decisiones correctas y el total de decisiones.

$$P_c = \frac{f_{11} + f_{22}}{n}$$

La sensibilidad de la nueva prueba la obtendremos a partir de la proporción de sujetos clasificados correctamente como verdaderos positivos.

$$Sensibilidad = \frac{Diagnósticos\ \ positivos\ \ de\ \ la\ \ prueba}{Total\ \ diagnósticos\ \ positivos} = \frac{f_{11}}{f_{.1}}$$

Mientras que la especificidad se obtiene mediante el cociente de los diagnosticados sin trastorno por la prueba entre el total de diagnósticos negativos.

$$Especificidad = \frac{Diagnósticos\ \ negativos\ \ de\ \ la\ \ prueba}{Total\ \ diagnósticos\ \ negativos} = \frac{f_{22}}{f_{.2}}$$

Un índice global para valorar la validez lo proporciona el cálculo del coeficiente kappa, establecido inicialmente como indicador del acuerdo entre dos observadores. La ventaja que presenta consiste en su fácil interpretación, similar a la de otros indicadores de relación entre variables. En efecto, su valor fluctúa entre 0 (ningún acuerdo) a valor 1 (máximo acuerdo). Su fórmula de cálculo es sencilla:

$$K = \frac{F_c - F_a}{n - F_a}$$

Donde

$$F_c = f_{11} + f_{22} \qquad y \qquad F_a = \frac{f_{1.} \cdot f_{.1} + f_{2.} \cdot f_{.2}}{n}$$

Tabla 16. Criterios Alman para interpretar kappa

Valor	Relación
0-0,20	Inexistente
0,21-0,40	Muy baja
0,41-0,60	Moderada
0,61-0,80	Buena
0,81-1	Intensa

En una consulta psicológica se pretende validar una nueva prueba, más simple que las tradicionales, para el diagnóstico de trastorno de depresión de los pacientes atendidos. En una muestra de 500 pacientes atendidos en el centro se administran dos pruebas (tradicional y versión breve) para el diagnóstico del trastorno de depresión.

Tabla 17

		Diagnóstico tradicional depresión		
		Positivo	**Negativo**	**Total**
Versión breve Escala Hamilton	**Positivo**	125	50	175
	Negativo	25	300	325
	Total	150	350	500

$$P_c = \frac{125+300}{500} = 0,85$$

Las dos pruebas presentan un porcentaje de acuerdo (85%) elevado.

Asimismo, los valores de sensibilidad y especificidad indican buena capacidad de la nueva prueba en la detección de sujetos con trastorno depresivo (sensibilidad = 0,83), como en la detección de los sujetos sin depresión (especificidad = 0,86).

$$Sensibilidad = \frac{125}{150} = 0,83$$

$$Especificidad = \frac{300}{350} = 0,86$$

$$F_c = 125+300 = 425 \qquad \text{y} \qquad F_a = \frac{175 \cdot 150 + 325 \cdot 350}{500} = 280$$

$$K = \frac{425-280}{500-280} = 0,66$$

El cálculo del índice kappa de acuerdo entre las dos pruebas (K = 0,66) indica una buena relación entre las dos pruebas. Por tanto, parece adecuado que la economía de tiempo y esfuerzo (tanto para el paciente como para el terapeuta) justificaría la nueva escala de diagnóstico de depresión en función de los resultados obtenidos en la validez de decisión.

5.3. Evidencias convergentes y/o discriminantes

Hasta ahora, en este apartado, nos hemos referido a escalas que pretenden mesurar un solo constructo psicológico. Si nos referimos a pruebas formadas por ítems que miden diferentes constructos (múltiples rasgos), podremos diferenciar entre dos tipos de validez. Por una parte, la validez convergente (enunciada anteriormente), es decir, la validez que determinan diferentes pruebas que miden el mismo constructo, y por otra parte, la validez discriminante, que viene determinada por la medida de diferentes constructos dentro de la misma prueba.

La matriz de correlaciones entre las puntuaciones de los diferentes rasgos obtenidos a partir de las diferentes escalas (matriz multirrasgo-multimétodo; Campbell y Fiske, 1959) nos servirá para determinar los diferentes valores de validez convergente y discriminante.

Imaginemos la situación en la que disponemos de tres escalas diferentes, formadas por ítems que miden los mismos dos constructos subyacentes. Escalas A, B y C, que miden los constructos 1 y 2. Para cada sujeto analizado tendremos por tanto seis puntuaciones diferentes obtenidas por la combinación de cada prueba y cada rasgo analizado.

Tabla 18

	A1	**A2**	**B1**	**B2**	**C1**	**C2**
A1	**Fi**					
A2	Vd	**Fi**				
B1	Vc		**Fi**			
B2		Vc	Vd	**Fi**		
C1	Vc		Vc		**Fi**	
C2		Vc		Vc	Vd	**Fi**

En la tabla anterior las diferentes escalas se representan por las letras A, B y C; dentro de cada escala 1 y 2 representan los dos constructos que analizar.

Si observamos la matriz multirrasgo-multimétodo (MRMM), encontramos en la diagonal principal valores de fiabilidad de las escalas (valores 1 si son obtenidas en una única administración).

Los valores de validez convergente se encuentran en las combinaciones de los mismos rasgos y diferentes escalas (por ejemplo, casilla A1 y B1), esperando que estos valores de validez sean suficientemente altos, lo cual indica convergencia de las diferentes maneras de medir el constructo, aportando evidencia real de la existencia del constructo.

Los valores de validez discriminante serán los coeficientes de correlación obtenidos dentro de la misma escala por las puntuaciones de los diferentes rasgos. En este caso, esperamos que los diferentes constructos sean suficientemente independientes entre sí para que las correlaciones sean próximas a cero.

En una muestra de 600 sujetos se han utilizado tres pruebas diferentes de personalidad (tests 1, 2 y 3), formadas cada una de ellas por ítems referidos a los tres mismos constructos de la personalidad (rasgos A, B y C).

La tabla siguiente muestra los valores de la matriz de correlaciones entre las 9 variables obtenidas de la combinación de las tres pruebas y los tres rasgos (3 x 3).

Taula 19

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	1								
B1	*0,03*	1							
C1	*0,28*	*0,17*	1						
A2	0,73	0,14	0,22	1					
B2	0,15	0,69	0,03	*0,18*	1				
C2	0,12	0,04	0,81	*0,03*	*0,15*	1			
A3	0,77	0,11	0,09	0,77	0,21	0,16	1		
B3	0,21	0,75	0,18	0,21	0,68	0,03	*0,15*	1	
C3	0,19	0,05	0,78	0,22	0,09	0,72	*0,14*	*0,09*	1

Los valores de validez convergente son los valores resaltados por el subrayado. Los valores de validez discriminante se resaltan por la cursiva.

Si nos fijamos en el rasgo B, se detecta convergencia a partir de la medida a través de diferentes pruebas:

$$r\,(B1 - B2) = 0{,}69$$

$$r\,(B1 - B3) = 0{,}75$$

$$r\,(B2 - B3) = 0{,}68$$

Si nos fijamos en la escala 1, observamos que existe suficiente independencia entre las diferentes medidas de los tres rasgos medidos:

$$r\,(A1 - B1) = 0{,}03$$

$$r\,(A1 - C1) = 0{,}28$$

$$r\,(B1 - C1) = 0{,}17$$

5.4. Evidencias basadas en las relaciones test-criterio

En ocasiones, un test o prueba psicológica construida para la medida de determinado constructo psicológico puede encontrarse relacionada con otra variable de interés, que se denomina criterio.

Por ejemplo, imaginemos que hemos elaborado una prueba válida que nos permite medir la capacidad de razonamiento numérico de las personas (test), y observamos que presenta una muy alta correlación con los resultados que obtienen los sujetos en una determinada prueba de matemáticas (criterio).

Podemos distinguir tres tipos de situaciones:

- Validez concurrente o simultánea.

- Validez predictiva.

- Validez retrospectiva.

5.4.1. Validez concurrente o simultánea

En este caso el test y el criterio se miden de manera simultánea. Obtendremos validez concurrente al obtener valores altos de coeficientes de correlación entre las puntuaciones del test y del criterio. Por tanto, nos permite validar el test, inicialmente elaborado para la medida de otra variable, para la medida del criterio.

En función del tipo de escala de medida utilizado tanto para las puntuaciones del test como del criterio, utilizaremos un tipo u otro de coeficiente para la medida de la correlación.

Como ejemplo veamos la siguiente situación, en la que a un grupo de 20 sujetos se les ha administrado un test de razonamiento numérico, justo antes de realizar determinada prueba de matemáticas.

Tabla 20

Test razonamiento	Nota matemáticas
100	10
100	9
100	9
90	8
90	8
80	9
70	7
70	7
70	7
70	5
70	4
70	4
50	3
40	3
40	2
40	2
30	1

Test razonamiento	Nota matemáticas
20	1
20	3
20	2

Una visión del gráfico de dispersión nos dará una idea de si se observa relación lineal entre ambas pruebas o no:

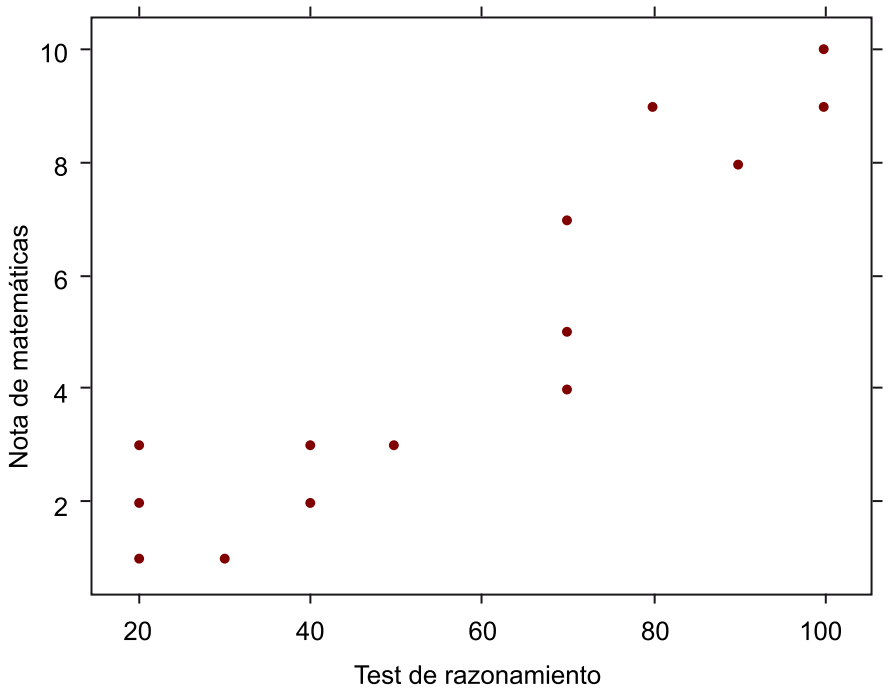

Figura 6. Gráfico de dispersión (nube de puntos)

Al calcular el coeficiente de correlación de Pearson:

```
> rcorr.adjust(Datos[,c("Nota.Matemáticas","Test.Razonamiento")], type="pearson")
                 Nota.Matemáticas     Test.Razonamiento
Nota.Matemáticas          1.00              0.92291
Test.Razonamiento         0.92291           1.00


n= 20
```

Se obtiene un valor de validez concurrente igual a 0,92291, que indica una fuerte relación directa y próxima a 1.

Una medida de la bondad de ajuste lineal entre las dos variables se define por r^2, es decir, el valor del cuadrado de la correlación, en nuestro ejemplo $r^2 = (0{,}92291)^2 = 0{,}8518$. Este valor, que se denomina coeficiente de determinación, multiplicado por 100, indica el porcentaje de variabilidad en la variable criterio, que viene explicado por la relación con la variable independiente. Por

tanto, el 85,18% de la variabilidad que presentan las puntuaciones obtenidas en la nota de matemáticas estaría explicada por la relación que presenta con los valores obtenidos en el test de razonamiento numérico.

5.4.2. Validez predictiva

Si conocemos que un determinado test y una variable criterio se encuentran altamente relacionados, será posible utilizar los valores obtenidos en el test para la predicción o el pronóstico de los valores que se obtendrán en el criterio. Hablaremos en este caso de la validez predictiva que tiene el test respecto a la variable criterio.

> Por ejemplo, si queremos seleccionar un candidato para un determinado puesto de trabajo, podemos utilizar determinadas pruebas que tengan alta validez predictiva con el futuro rendimiento de los candidatos en el puesto de trabajo. O, utilizando el ejemplo mencionado anteriormente, podemos hacer un pronóstico de la nota que obtendrán los sujetos en una determinada prueba de matemáticas, a partir de las puntuaciones que obtuvieron en su momento, en el test de razonamiento numérico.

Cuando tanto las puntuaciones del test como el criterio son puntuaciones numéricas y hemos calculado el correspondiente coeficiente de correlación de Pearson, siendo este estadísticamente significativo, es posible establecer un modelo de regresión lineal para poder realizar el pronóstico de los valores del criterio.

Las puntuaciones en el test constituyen la variable independiente del modelo (variable predictora), mientras que el criterio representa la variable dependiente.

Regresión lineal simple

Es el caso más sencillo: solo disponemos de una variable independiente (X) y una variable dependiente (Y).

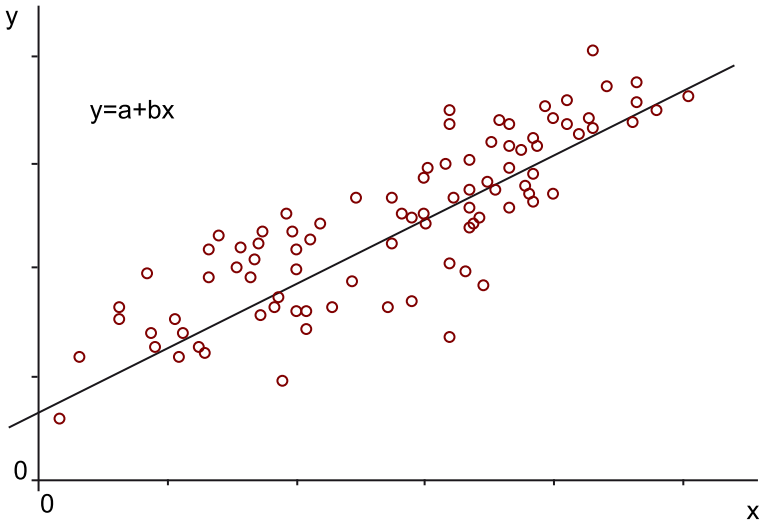

Figura 7. Gráfico de dispersión con recta de regresión

La regresión lineal describe una relación lineal entre Y e X, esto es, representar una recta en el gráfico de dispersión, que mejor ajuste a la nube de puntos.

La expresión de una línea recta es $y = a + bx$, donde b representa la pendiente de la recta, o sea, el cambio que se produce en y a partir del cambio que se produzca en x; y a se denomina intersección o intercepta, y es el valor que toma y cuando x es igual a cero.

Para encontrar los coeficientes de la regresión, a y b, usamos un método de estimación muy conocido en estadística, el método de mínimos cuadrados, que minimiza la suma de los cuadrados de las diferencias (o residuos) entre los valores y_i y los valores estimados según la recta de regresión $y'_i = a + bx_i$.

A partir de los datos (x_i, y_i), $i = 1, ..., n$, estimamos los coeficientes a y b de la recta de regresión. Así pues, tenemos:

- Pendiente:

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{s_x^2}$$

Lectura de la fórmula

S_{xy}: Covarianza entre x e y.

s^2_x: Varianza de x.

- Intersección:

$$a = \overline{y} - b \cdot \overline{x}$$

Comparando las fórmulas de la pendiente b y del coeficiente de correlación r, tenemos la relación siguiente:

$$b = r_{xy} \cdot \frac{s_y}{s_x}$$

Es posible verificar o validar el modelo de regresión a partir del coeficiente de determinación r^2. Recordemos que lo hemos definido anteriormente como una medida de bondad de ajuste, o medida de la proximidad de los puntos a la recta estimada. Representa la proporción de varianza de la variable dependiente explicada por la recta de regresión.

El valor $1 - r^2$ cuantifica la proporción de varianza que no es explicada por la regresión. A partir de estos dos valores podemos calcular un estadístico de contraste:

$$F_{EC} = \frac{r2}{(1-r2)/n-2}$$

Este estadístico de contraste F se distribuye siguiendo una distribución F de Snedecor, con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador.

Las hipótesis que contrastar serán:

- H_0: El modelo no es válido, no hay relación.

- H_1: Sí existe relación, por tanto el modelo sí es válido.

Siguiendo con el ejemplo del test de razonamiento numérico y el criterio de la nota de matemáticas, el resultado con el programa R, es el siguiente:

```
> summary(RegModel.1)

Call:
lm(formula = Nota.Matemáticas ~ Test.Razonamiento, data = Datos)

Residuals:
   Min       1Q    Median    3Q      Max
-2.01102 -0.96970 -0.03857 0.98898 2.05785

Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)      -1.085399  0.673899   -1.611    0.125
Test.Razonamiento 0.101377  0.009969   10.170    6.89e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.201 on 18 degrees of freedom

Multiple R-squared: 0.8518, Adjusted R-squared: 0.8435
F-statistic: 103.4 on 1 and 18 DF, p-value: 6.89e-09
```

Consultado el listado, podemos especificar el modelo de regresión:

$$\textit{Nota matemáticas} = -1{,}085 + 0{,}1014 \times \textit{Test} + \textit{Residual}$$

También observamos que el modelo se encuentra verificado, ya que el valor de p (grado de significación) que acompaña al valor del estadístico de contraste (F = 103,4) es tendente a cero. Por tanto, nada se opone a rechazar la hipótesis nula, y el modelo se encuentra validado.

Si representamos en el gráfico de dispersión la recta de regresión:

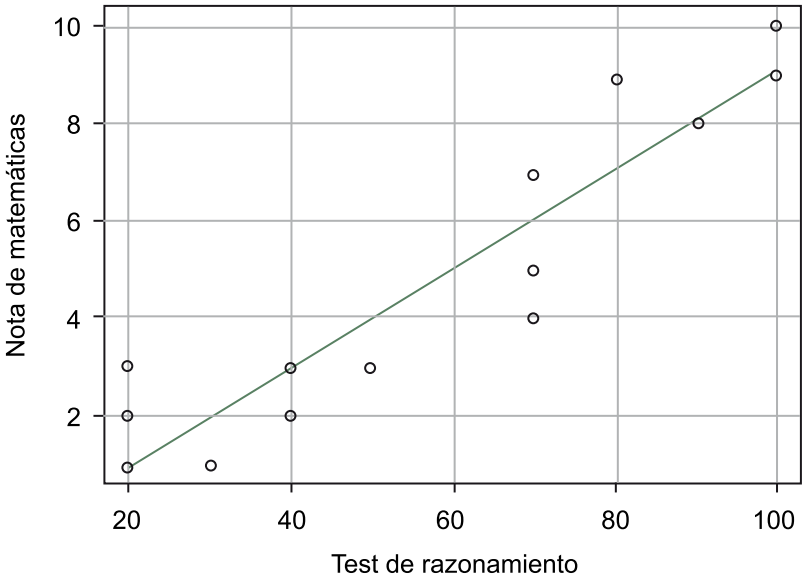

Figura 8. Gráfico de dispersión con ajuste de la recta

A partir de la expresión de la recta de regresión, podemos realizar el pronóstico para cada sujeto del valor de nota de matemáticas en función del test, así como el cálculo del residual, calculado mediante la diferencia entre la puntuación real obtenida y la puntuación pronosticada.

Tabla 21

Test razonamiento	Nota matemáticas	Pronóstico	Residual
100	10	9,0546	0,9454
100	9	9,0546	–0,0546
100	9	9,0546	–0,0546
90	8	8,0406	–0,0406
90	8	8,0406	–0,0406
80	9	7,0266	1,9734
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	7	6,0126	0,9874
70	5	6,0126	–1,0126
70	4	6,0126	–2,0126
70	4	6,0126	–2,0126
50	3	3,9846	–0,9846
40	3	2,9706	0,0294
40	2	2,9706	–0,9706
40	2	2,9706	–0,9706
30	1	1,9566	–0,9566

Test razonamiento	Nota matemáticas	Pronóstico	Residual
20	1	0,9426	0,0574
20	3	0,9426	2,0574
20	2	0,9426	1,0574

```
> numSummary(Datos[,c("Nota.Matemáticas", "Pronóstico", "Residual")],
+  statistics=c("mean", "sd","var"))
                 mean     sd        var     n
Nota.Matemáticas 5.2000  3.036619  9,221    20
Pronóstico       5.2014  2.803138  7,855    20
Residual        -0.0014  1.169176  1,367    20
```

Tal como hemos indicado anteriormente, se define el coeficiente de determinación como el cociente entre la varianza explicada por la regresión y la varianza total de la variable criterio:

$$r^2 = \frac{s^2_{y'}}{s^2_y} = 1 - \frac{s^2_{y-y'}}{s^2_y}$$

Lectura de la fórmula

s^2_y: Varianza de la variable dependiente *Y*.

$s^2_{y'}$: Varianza de los pronósticos obtenidos mediante la ecuación de regresión.

$s^2_{y-y'}$: Varianza de los errores producidos.

Con los datos de nuestro ejemplo:

$$r^2 = \frac{7,855}{9,221} = 1 - \frac{1,367}{9,221} = 0,8518$$

Por tanto, a partir de los valores del coeficiente de determinación y de la varianza de la variable criterio, es posible, despejando de la expresión, obtener el valor de la varianza de los errores:

$$s^2_{y-y'} = s^2_y(1 - r^2)$$

La desviación típica de los errores o error típico o estándar del error nos ayudará en la estimación por intervalo de nuevos valores desconocidos.

$$s_{y-y'} = s_y\sqrt{1 - r^2}$$

En efecto, si necesitamos realizar un pronóstico en el criterio a partir de un nuevo valor en el test, o variable predictora, es posible realizarlo de forma puntual, pero conseguiremos mejores estimaciones, dada una probabilidad, si se realiza por intervalo.

$$IC^{1-\alpha} \rightarrow y' \pm t_{n-1;\alpha/2} \cdot s_{y-y'}$$

Lectura de la fórmula

1 – α: Nivel de confianza del intervalo construido.

t: Valor de la distribución t de Student-Fisher tabulado en función de α y de los grados de libertad (*n*-1).

Imaginemos que un nuevo sujeto obtiene una puntuación igual a 60 en el test de razonamiento numérico. La estimación puntual del valor en la nota de matemáticas será:

$$Nota\ matemáticas = -1{,}085 + 0{,}1014 \cdot 60 = 4{,}999$$

Si realizamos una estimación por intervalo, con un nivel de confianza del 95%:

$$IC^{0{,}95} \rightarrow 4{,}999 \pm 2{,}093 \cdot 1{,}169 = [2{,}552 \quad 7{,}446]$$

Con una probabilidad de 0,95 el valor en el criterio de un sujeto que obtenga una puntuación 60 en el test estará entre 2,552 y 7,446 puntos.

Regresión lineal múltiple

El modelo lineal general plantea que una variable dependiente (criterio) sea función de varias variables independientes, situación por otra parte bastante más habitual. En el caso de dos variables independientes la expresión que relaciona las tres variables será la fórmula de un plano. En las situaciones en las que existan más de dos variables independientes, situaciones multivariantes, hablaremos del hiperplano de regresión.

$$Criterio = a + b_1 \cdot Test_1 + b_2 \cdot Test_2 + b_3 \cdot Test_3 + \cdots + b_p \cdot Test_p + Residual$$

Recuperando el ejemplo anterior, imaginemos que a los 20 sujetos se les ha administrado un test de razonamiento junto con un test de cálculo mental, previamente a la realización de una prueba de matemáticas, que representa el criterio que más adelante queremos pronosticar.

Tabla 22

Test razonamiento	Test cálculo	Nota matemáticas
100	9	10
100	8	9
100	8	9
90	8	8
90	7	8
80	9	9
70	6	7
70	5	7
70	6	7

Test razonamiento	Test cálculo	Nota matemáticas
70	6	5
70	4	4
70	4	4
50	5	3
40	4	3
40	4	2
40	5	2
30	3	1
20	2	1
20	3	3
20	3	2

Al calcular la matriz de correlaciones de Pearson:

Tabla 23

	Test razonamiento	Test cálculo	Nota matemáticas
Test razonamiento	1		
Test cálculo	0,891763351	1	
Nota matemáticas	0,922905961	0,925261006	1

Observamos que el test de cálculo mental también se encuentra altamente correlacionado con la variable criterio (nota de matemáticas). Asimismo, vemos una alta correlación entre las dos variables independientes, test de razonamiento numérico y test de cálculo mental ($r = 0,89176$).

El análisis de regresión múltiple nos ayudará a determinar si la incorporación de esta nueva variable aumenta, significativamente, la variabilidad explicada por la regresión en la variable criterio.

El análisis de regresión se basará en el análisis de la relación conjunta entre la variable criterio y el conjunto de las dos variables independientes. El cuadrado de esta correlación múltiple será el nuevo coeficiente de determinación.

La verificación del modelo se realizará a partir de la expresión:

$$F_{EC} = \frac{r2/p}{(1-r2)/n-p-1}$$

Lectura de la fórmula

p es igual al número de variables independientes en el modelo.

El estadístico de contraste *F* se distribuye siguiendo una distribución *F* de Snedecor, con *p* grados de libertad en el numerador y $(n - p - 1)$ grados de libertad en el denominador.

A continuación se presenta el listado obtenido en este ejemplo mediante el uso del programa R:

```
> RegModel.2 <- lm(Nota.Matemáticas~Test.Numérico+Test.Razonamiento, data=Datos)

> summary(RegModel.2)

Call:
lm(formula = Nota.Matemáticas ~ Test.Numérico + Test.Razonamiento, data = Datos)

Residuals:
   Min      1Q    Median    3Q      Max
-1.72686 -0.64006 -0.00108 0.52167 1.74005

Coefficients:
                    Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)         -1.91579  0.62646    -3.058    0.00711 **
Test.Cálculo        0.70883   0.23719     2.988    0.00825 **
Test.Razonamiento   0.05246   0.01835     2.858    0.01088 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 17 degrees of freedom
Multiple R-squared: 0.9028, Adjusted R-squared: 0.8914
F-statistic: 78.96 on 2 and 17 DF, p-value: 2.481e-09
```

El valor de coeficiente de determinación es 0,9028; por tanto, un 90,28% de la varianza de la variable criterio viene explicada por la regresión entre esta variable y la combinación de las dos variables independientes.

La ecuación del plano de regresión se encuentra verificada globalmente, ya que el valor del estadístico de contraste (F = 78,96) indicia una probabilidad tendente a cero (*p value*) de que no exista relación entre las variables.

Es importante también observar si se encuentran significados los diferentes coeficientes de la regresión (*b*). En este caso, tanto el coeficiente que afecta al test de razonamiento (*p value* = 0,01) como el que afecta al test de cálculo mental (p = 0,008), se encuentran verificados, ya que sus respectivos grados de significación asociados son próximos a cero.

La especificación de la ecuación resultante del modelo de regresión quedará, por tanto, de la siguiente manera:

$$Nota\ matemáticas = -1{,}916 + 0{,}709 \cdot Test\_cálculo + 0{,}052 \cdot Test\_razonamiento + Residual$$

El listado, asimismo, informa del valor del error típico o estándar del error ($s_{y\text{-}y'}$ =1,001), valor necesario para la estimación por intervalo de los valores del criterio. En efecto, para un nuevo sujeto que obtuviera una puntuación igual a 6 en el test de cálculo mental y 60 en el test de razonamiento numérico, la estimación puntual del valor en la nota de matemáticas será:

$$Nota\ matemáticas = -1{,}916 + 0{,}709 \cdot 6 + 0{,}052 \cdot 60 = 5{,}458$$

Si realizamos una estimación por intervalo, con un nivel de confianza del 95%:

$$IC^{0{,}95} \rightarrow 5{,}458 \pm 2{,}093 \cdot 1{,}001 = [3{,}363 \quad 7{,}553]$$

Con una probabilidad de 0,95, el valor en el criterio de un sujeto que obtenga una puntuación 6 en la prueba de cálculo y 60 en el test de razonamiento estará entre 3,363 y 7,553 puntos.

Otras técnicas estadísticas de análisis multivariable

En función del tipo de escala de medida utilizada para las variables criterio y las variables predictoras, será necesario aplicar alguna de las diferentes técnicas de análisis de datos multivariables.

Por ejemplo, si disponemos de una variable criterio medida en escala nominal, y por tanto de tipo cualitativo o categórico, mientras que las variables independientes son de tipo cuantitativo o numérico, podemos utilizar una técnica de clasificación, como el análisis discriminante. En efecto, supongamos que la variable criterio nota de matemáticas, del ejemplo utilizado, estuviera codificada en suspenso, aprobado, notable, excelente, o simplemente divididos los sujetos entre aprobados y suspendidos.

Aplicar un análisis discriminante nos permitiría determinar la mejor función discriminante, que consiga la clasificación de los sujetos en función de las puntuaciones obtenidas en las pruebas predictoras del cálculo mental y el razonamiento numérico. La función discriminante establecerá la estimación de los pesos (coeficientes) y la combinación lineal de las variables independientes (discriminantes), de modo que los grupos sean, desde el punto de vista estadístico, lo más diferentes posible.

Otra opción puede ser que tanto la variable criterio como las variables independientes se encuentren codificadas en categorías. En este caso, sería aplicable la técnica del modelo *logit*. Este modelo, basado en los modelos lineales

logarítmicos, pretende –siguiendo el enfoque de la regresión múltiple– encontrar la expresión de asociación entre la variable criterio y las variables independientes, teniendo en cuenta también la interacción entre las variables independientes.

5.4.3. Validez retrospectiva

La validez concurrente entre uno o varios tests y el criterio, que puede ser útil para la predicción futura de la variable criterio, también en ciertas situaciones puede servir para, dadas ciertas consecuencias medidas a través del criterio, encontrar las causas a los valores obtenidos.

En este caso la variable criterio ha sido registrada anteriormente a las variables predictoras. Por ejemplo, en psicología es habitual la aplicación de diferentes pruebas psicológicas que permitan dar una explicación a determinada conducta de un sujeto.

5.5. Generalización de la validez

El concepto de generalización de la validez se refiere al hecho de extender la validez establecida entre test y criterio a otras situaciones o a grupos de sujetos diferentes a los utilizados inicialmente en el cálculo.

Por ejemplo, imaginemos que se encuentran validadas determinadas pruebas dirigidas a la correcta selección de personal para determinados puestos de administrativo en un banco A. Si es posible utilizar las mismas pruebas para la selección de administrativos en otro banco B, podremos considerar que la validez se ha generalizado. En este caso se trataría de sujetos diferentes a los utilizados inicialmente.

Otro caso podría ser que las pruebas de selección para administrativos en banca se utilicen para la selección de personal administrativo en compañías de seguros. En este caso hablaríamos de generalización también a situaciones diferentes.

6. Evidencia de validez basada en las consecuencias de la aplicación

Cuando se toman decisiones a partir de la aplicación de un test y no se trata solo de describir o interpretar sin que existan acciones que se deriven de ello, se debe pensar en las consecuencias que tiene aplicar dicho cuestionario (Shepard, 1997). Los tests deben usarse cuando se maximicen las consecuencias positivas (beneficios) y se minimicen las negativas (costes) derivadas de su aplicación.

Los tests se aplican esperando que de la información obtenida se extraiga algún tipo de beneficio (poder seleccionar el mejor tratamiento terapéutico, ubicar a los trabajadores de una empresa en el puesto más adecuado, mejorar las técnicas didácticas empleadas, etc.). Uno de los propósitos fundamentales de la validación es indicar en qué casos se pueden obtener estos beneficios.

Dentro de este concepto hay que diferenciar entre evidencias que son relevantes para la validez y evidencias que son importantes para las políticas sociales pero que se sitúan fuera del concepto de validez. Esta diferencia se hace más importante cuando las consecuencias que se derivan del test son diferentes para distintos grupos. Por ejemplo, si se sabe que existen diferencias entre hombres y mujeres en las puntuaciones de un test empleado para la selección de personal, esto va a afectar al uso del test pero no se podría afirmar nada sobre las evidencias de validez basadas en las consecuencias de la aplicación. Para ello, se debe realizar un estudio más pormenorizado de las consecuencias. Si las diferencias se deben a que el aspecto evaluado se distribuye de manera diferente entre los grupos en la población, las diferencias obtenidas no implican que las decisiones que se extraigan de la aplicación del test carezcan de validez. El problema surge cuando dichas diferencias se deben a que se están valorando habilidades que no están relacionadas con la labor que van a realizar los seleccionados o cuando el test es sensible a algunas características de los candidatos que no se pretende que estén relacionadas con el constructo que se va a medir. En el primero de los casos no se puede concluir la falta de indicios de validez respecto a las consecuencias, pero en las dos últimas situaciones sí. No obstante, evidentemente, las tres situaciones serían inadecuadas dentro de las políticas sociales de igualdad de género (APA, 1999).

7. Factores que afectan a la validez

Tal y como se ha comentado anteriormente, uno de los indicios de validez que se pueden (o deben) calcular es la correlación existente entre el test y un criterio ajeno a este. Dicha correlación se puede ver afectada por múltiples factores, como son la fiabilidad de ambas medidas, la longitud del test y la variabilidad (dispersión) de la muestra empleada para obtener las puntuaciones. A continuación se tratarán estos aspectos.

7.1. Fórmulas de atenuación

Cuando se trata de calcular la correlación entre un test y un criterio se parte de las puntuaciones empíricas que se han obtenido en ambos cuestionarios. Dichas puntuaciones están compuestas, según el modelo lineal de Spearman, por la puntuación verdadera y el error de medida, que es aleatorio. Por tanto:

$$X = V_x + e_x$$
$$Y = V_y + e_y$$

Lectura de la fórmula

X: Puntuación obtenida en el test.
V_x: Puntuación verdadera en el test.
e_x: Error de medida (aleatorio) en el test.
Y: Puntuación obtenida en el criterio.
V_y: Puntuación verdadera en el criterio.
e_y: Error de medida (aleatorio) en el criterio.

Como se puede comprobar, al correlacionar las puntuaciones X e Y también se están correlacionando los dos errores de medida entre sí. Dichos errores son aleatorios, lo que significa que la correlación entre ambos debe ser igual a 0. Por ello, cuanta mayor importancia tengan los errores en la puntuación obtenida (o lo que es lo mismo, cuanto más baja sea la fiabilidad del test y el criterio empleados), menor será la correlación entre X e Y. En definitiva, se puede encontrar una regla según la cual a mayor fiabilidad del test y el criterio, la correlación entre ambos aumentará. Para saber en qué grado lo hace se emplean las fórmulas de atenuación (Spearman, 1907).

7.1.1. Estimación del coeficiente de validez en el supuesto de que el test y el criterio tengan una fiabilidad perfecta

En el caso en el que se supone que el test y el criterio poseen una fiabilidad perfecta se asume que los errores de medida son iguales a 0. En esta situación (en la que el coeficiente de fiabilidad es igual a 1) la puntuación empírica (X o Y según se trate del test o del criterio) es igual a la verdadera.

En este caso la nueva correlación puede calcularse mediante:

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

Lectura de la fórmula

ρ_{xy}: Coeficiente de validez obtenido al correlacionar las puntuaciones del test y el criterio.

$\rho_{xx'}$: Coeficiente de fiabilidad del test.

$\rho_{yy'}$: Coeficiente de fiabilidad del criterio.

Ejemplo

La correlación entre un test de ansiedad y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79 y la del criterio de 0,90. ¿Cuál es la estimación de dicha correlación si se supone que ambos tienen una fiabilidad perfecta?

$$\rho_{v_x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0,63}{\sqrt{0,79}\sqrt{0,90}} = 0,75$$

La correlación que se estima entre el test y el criterio si ambos tuviesen una fiabilidad perfecta pasa de 0,63 a 0,75.

7.1.2. Estimación del coeficiente de validez en el supuesto de que el test tenga una fiabilidad perfecta

En el caso de que solo el test tenga una fiabilidad perfecta (igual a 1) la estimación de la nueva correlación viene dada por:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}}$$

En el ejemplo anterior si solo el test tuviese la fiabilidad perfecta, el resultado sería:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,79}} = 0,71$$

La correlación que se estima entre el test y el criterio al suponer que el test tiene una fiabilidad perfecta cambia de 0,63 a 0,71.

7.1.3. Estimación del coeficiente de validez en el supuesto de que el criterio tenga una fiabilidad perfecta

Si es el criterio el que se supone que tiene una fiabilidad perfecta, la estimación de la correlación viene dada por:

$$\rho_{x v_y} = \frac{\rho_{xy}}{\sqrt{\rho_{yy'}}}$$

En el ejemplo anterior:

$$\rho_{v_x y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}} = \frac{0,63}{\sqrt{0,90}} = 0,66$$

La correlación que se estima entre el test y el criterio al suponer que el test tiene una fiabilidad perfecta cambia de 0,63 a 0,66.

7.1.4. Estimación del coeficiente de validez en el supuesto de que se ha mejorado tanto la fiabilidad del test como la del criterio

La situación más frecuente es en la que se mejora la fiabilidad del test, la del criterio o la de ambos pero sin llegar a 1 (este suceso es más teórico que práctico). A continuación se verá cómo estimar la correlación entre test y criterio cuando se mejoran las fiabilidades de ambos. Posteriormente se tratarán las otras opciones.

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}}$$

Lectura de la fórmula

$\rho_{XX'}$: Es la fiabilidad mejorada del test.

$\rho_{YY'}$: Es la fiabilidad mejorada del criterio.

Ejemplo

La correlación entre un test de ansiedad y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79 y la del criterio de 0,90. Añadiendo ítems se consigue incrementar la fiabilidad del test hasta 0,83 y la del criterio hasta 0,92 ¿Cuál es la estimación de dicha correlación tras haber mejorado la fiabilidad de ambos?

$$\rho_{XY} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}} = \frac{0{,}63\sqrt{0{,}83}\sqrt{0{,}92}}{\sqrt{0{,}79}\sqrt{0{,}90}} = 0{,}65$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad de ambos pasa de 0,63 a 0,65.

7.1.5. Estimación del coeficiente de validez en el supuesto de que se ha mejorado la fiabilidad del test

Cuando solo se mejora la fiabilidad del test, la estimación del nuevo coeficiente de correlación viene dado por:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}}$$

En el ejemplo anterior, si solo se mejora la fiabilidad del test:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{\rho_{XX'}}}{\sqrt{\rho_{xx'}}} = \frac{0{,}63\sqrt{0{,}83}}{\sqrt{0{,}79}} = 0{,}645$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad del test pasa de 0,63 a 0,645.

7.1.6. Estimación del coeficiente de validez en el supuesto de que se ha mejorado la fiabilidad del criterio

La estimación del coeficiente de correlación, cuando solo se mejora la fiabilidad del criterio, viene dada por:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}}$$

En el ejemplo anterior, si solo se mejora la fiabilidad del criterio la respuesta sería:

$$\rho_{xY} = \frac{\rho_{xy}\sqrt{\rho_{YY'}}}{\sqrt{\rho_{yy'}}} = \frac{0.63\sqrt{0,92}}{\sqrt{0,90}} = 0,637$$

La estimación de la correlación entre el test y el criterio al haber mejorado la fiabilidad del criterio pasa de 0,63 a 0,637.

7.1.7. Valor máximo que puede alcanzar el coeficiente de correlación entre test y criterio

Como se puede apreciar, a medida que incrementamos el coeficiente de correlación del test, del criterio o de ambos, el coeficiente de correlación aumenta. Esto solo ocurre hasta cierto punto, ya que el coeficiente de correlación entre test y criterio siempre es menor o igual que su índice de fiabilidad ($\rho_{xv} = \sqrt{\rho_{xx'}}$).

Matemáticamente puede representarse como:

$$\rho_{xy} \leq \rho_{xv}$$

Por tanto, en el ejemplo anterior, el máximo coeficiente de correlación que se puede obtener entre el test y el criterio es:

$$\rho_{xv} = \sqrt{\rho_{xx'}} = \sqrt{0,79} = 0,89$$
$$\rho_{xy} \leq 0,89$$

Así pues, el máximo valor del coeficiente de correlación que se puede obtener en ese test es de 0,89.

7.2. Efecto de la longitud del test sobre el coeficiente de correlación test-criterio

Uno de los medios por el que se puede incrementar el coeficiente de correlación test-criterio es aumentando el número de ítems que componen el test. La relación entre el número de ítems y dicha correlación es directa, es decir, a medida que se incremente el número de ítems la correlación aumentará (y se reducirá si se quitan ítems).

La relación entre ambos viene dada por:

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}}$$

Lectura de la fórmula

ρ_{xy}: Es el valor inicial de la correlación test-criterio.

$\rho_{x'}$: Es el coeficiente de fiabilidad del test.

n: Es el número de veces que se aumenta el test.

Ejemplo

La correlación entre un test de ansiedad de 20 ítems y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79. Estima el valor de la correlación test-criterio si se añaden 10 ítems más.

$$n = \frac{20+10}{20} = 1,5$$

$$\rho_{Xy} = \frac{\rho_{xy}\sqrt{n}}{\sqrt{1+(n-1)\rho_{xx'}}} = \frac{0,63\sqrt{1,5}}{\sqrt{1+(1,5-1)\,0,79}} = 0,65$$

Al añadir 10 ítems a los 20 originales, el coeficiente de correlación incrementa desde 0,63 hasta 0,65.

Otra posibilidad es el hecho de querer llegar hasta un coeficiente de correlación que se desee y por tanto haya que calcular el número de ítems que se deben añadir al cuestionario para poder alcanzarlo. Para ello:

$$n = \frac{\left(1-\rho_{xx'}\right)\rho^2_{Xy}}{\rho^2_{xy} - \rho^2_{Xy}\rho_{xx'}}$$

Lectura de la fórmula

ρ^2_{Xy}: Es el cuadrado del coeficiente de correlación deseado.

Ejemplo

La correlación entre un test de ansiedad de 20 ítems y un criterio (depresión) es de 0,63. La fiabilidad del test es de 0,79. ¿Cuántos ítems habrá que añadir al test si se pretende alcanzar un coeficiente de correlación test-criterio de 0,67?

$$n = \frac{\left(1-\rho_{xx'}\right)\rho^2_{Xy}}{\rho^2_{xy}-\rho^2_{Xy}\rho_{xx'}} = \frac{(1-0,79)0,67^2}{0,63^2-0,67^2 0,79} = 2,23$$

El test debería ser incrementado 2,23 veces. Dado que el test original tiene 20 ítems: 20 × 2,23 = 44,6. Evidentemente no se pueden tener decimales (estamos hablando de ítems, no se puede tener una porción de ítem en el test), por lo que lo debemos ajustar. Cuando estemos añadiendo ítems, el ajuste siempre se debe hacer hacia el entero superior, es decir, en este caso tendría que haber 45 ítems. En el caso de que el test sea excesivamente largo y no nos importe eliminar ítems hasta llegar a un coeficiente de correlación test-criterio menor (pasar de 0,63 a 0,50 por ejemplo), el ajuste se deberá hacer al entero inferior.

7.3. Efecto de la variabilidad de la muestra en la correlación test-criterio

El coeficiente de correlación se ve muy afectado por la dispersión de la muestra en la que esté calculado. La relación entre la dispersión y la correlación es directa: a mayor dispersión se obtendrá una mayor correlación.

En algunos campos de la psicología es muy frecuente que solo se pueda calcular la correlación entre el test y el criterio en una pequeña muestra de personas. El ejemplo más claro es el de la selección de personal. Tras haber empleado un test para seleccionar de entre los candidatos a los más adecuados al puesto, solo

se puede correlacionar la puntuación obtenida en el test con un criterio como el de rendimiento laboral. En este caso, la dispersión de los seleccionados será menor que la del total de candidatos, ya que se selecciona a las personas que tienen características muy similares (y que más se ajustan a las buscadas para el puesto).

Tras calcular el coeficiente de correlación en la muestra de seleccionados puede interesar estimar cuál sería si se hubiese calculado sobre la totalidad de aspirantes. Para ello, se debe partir de dos supuestos: la pendiente de la recta de regresión es la misma para los dos grupos (admitidos y aspirantes) y el error típico de estimación también es el mismo para ambos grupos. Matemáticamente:

$$a)\ \frac{\rho_{xy}\sigma_y}{\sigma_x} = \frac{\rho_{XY}\sigma_Y}{\sigma_X}$$

$$b)\ \sigma_y\sqrt{1-\rho_{xy}^2} = \sigma_Y\sqrt{1-\rho_{XY}^2}$$

Donde las letras mayúsculas hacen referencia al grupo de admitidos y las minúsculas al total de aspirantes.

Para estimar el valor de la correlación test-criterio en el total de aspirantes tras haberla calculado en el de admitidos solo es necesario aplicar:

$$\rho_{xy} = \frac{\sigma_x\rho_{XY}}{\sqrt{\sigma_x^2\rho_{XY}^2 + \sigma_X^2 - \sigma_X^2\rho_{XY}^2}}$$

Lectura de la fórmula

ρ_{XY}: Coeficiente de correlación test-criterio en la muestra de admitidos.

σ_x^2: Varianza en el test del total de aspirantes.

σ_X^2: Varianza en el test de los admitidos.

Ejemplo

Se aplicó un test de asertividad a 1.000 personas para seleccionar a 10 sobrecargos de vuelo. La desviación típica obtenida en el test por el total de aspirantes fue de 15 y en la muestra de admitidos de 4. Tras un tiempo trabajando se comprobó que la correlación entre las puntuaciones en el test y la ejecución laboral (valorada por sus superiores) fue de 0,36. ¿Cuál sería el coeficiente de correlación test-criterio si se hubiese calculado sobre el total de los aspirantes?

$$\rho_{xy} = \frac{\sigma_x\rho_{XY}}{\sqrt{\sigma_x^2\rho_{XY}^2 + \sigma_X^2 - \sigma_X^2\rho_{XY}^2}} = \frac{15\times 0{,}36}{\sqrt{15^2\times 0{,}36^2 + 4^2 - (4^2\times 0{,}36^2)}} = 0{,}82$$

Como se puede observar, hay un incremento notable en el valor del coeficiente de correlación debido a la diferente dispersión que tienen los grupos.

Bibliografía

AERA, A. N. (1974). *Standards for educational and psychological tests*. Whashington, DC: American Psychological Association.

AERA, APA, y NCME (1966). *Standards for educational and psychological test and manuals*. Washington, DC: AERA.

AERA, APA, y NCME (1999). *Standards for educational and psychological testing*. Washington DC: AERA.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Beck, A., Rush, A. J., Shawn, B. F., y Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford Press.

Bingham, W. V. (1937). *Aptitudes and aptitude testing*. Oxford, England: Harpers.

Campbell, D. T. y Fiske, A. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike. *Educational measurement* (pp. 443-507). Washington DC: American Council on Education.

Cronbach, L. J. y Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Czaja, R. y Blair, J. (1996). *Designing Surveys*. Thousand Oaks, CA: Sage Publications.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, pp. 427-439.

Hamilton, M. (1960). A rating scale for depression. *Journal Neurol. Neurosurg. Psychiatry*, 23, 56-62.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. En S. M. Downing y T. M. Haladyna. *Handbook of test development* (pp. 131-153). New Jersey: Erlbaum.

Leighton, J. P. (2004). Avoiding misconception, missed use, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23, 6-15.

Messick, S. (1989). Validity. En R. Linn. *Educational Measurement*. 3.ª ed. (pp. 13-104). New York: Macmillan.

Muñiz, J. (2003). *Teoría clásica de los tests*. Madrid: Pirámide.

Prieto, G. y Delgado, A. R. (2010). Fiabilidad y Validez. *Papeles del Psicólogo*, 31, 67-74.

Ramos-Brieva, J. C. (1986). Validación de la versión castellana de la escala de Hamilton para la depresión. *Actas Luso-Esp. Neurol. Psiquiatr.*, 22, 21-28.

Scott, W. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology*, pp. 61-66.

Shepard, L. (1993). Evaluating test validity. En L. Darling-Hammond. *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5-13.

Shepard, L. A., Camilli, G., Linn, R., y Bohrnstedt, G. (1993). *Setting performance standards for achievement tests*. Stanford, CA: National Academy of Education.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, 45, 83-117.

Sireci, S. G. (1998). Gathering and evaluating content validity data. *Educational Assessment*, 5 (4), 299-321.

Spearman, C. (1907). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.