Psicología, tercera época ▪ Revista digital arbitrada
Vol. 35, N° 2-2016 | pp. 15-38 | ISSN: 1316- 0923

Análisis cuantitativo para items multipunto

Aysbel Gonzáles Álvarez
Escuela de Psicología, Universidad Central de Venezuela
aysbel.gonzalez@gmail.com

Resumen

Cuando el objetivo del instrumento es medir el comportamiento típico de los examinados, se suele utilizar items multipunto o tipo Likert, los cuales deben ser sometidos a una exhaustiva revisión cualitativa y cuantitativa para poder conformar la versión definitiva del instrumento psicométrico. En tanto se disponga de items de mayor calidad, el instrumento será mejor. El análisis cuantitativo de los items se realiza mediante la aplicación de estadísticos, y a partir del resultado obtenido en cada procedimiento, se propone un modelo de selección para aquellos items que se aproximen al valor ideal establecido teóricamente para cada estadístico, considerando el objetivo del instrumento, su población y la tabla de especificaciones. Se recomienda utilizar estadísticos que describan individualmente al item (media, desviación típica), que reporten su capacidad discriminativa (correlación inter-item) y su nivel de homogeneidad (consistencia interna).

Palabras clave: Análisis de items, Items tipo likert, Correlación inter-item, Consistencia interna.

Recibido: 10 de abril de 2016
Aceptado: 15 de julio de 2016
Publicado: 10 de enero de 2017

Psicología ▪ Refereed journal
Volume 35, Issue 2-2016 | Pages 15-38 | ISSN: 1316- 0923

Quantitative analysis of multi-point scale

Aysbel Gonzáles Álvarez
Escuela de Psicología, Universidad Central de Venezuela
aysbel.gonzalez@gmail.com

Abstract

When the objective of an instrument is to measure the typical behavior of examinees, multi-point scales or Likert-type items are usually used, which must be subjected to an exhaustive qualitative and quantitative review to be able to shape the final version of the psychometric instrument. As higher the quality of the items is, the better the instrument. The quantitative analysis of the items is done through statistical analyses, and from the result obtained in each procedure, a selection model is proposed for those items that demonstrate being close to the ideal value theoretically established for each statistic, taking into consideration the objective of the instrument, its population, and the table of specification. It is recommended to use statistics that describe items individually (mean, standard deviation), that report their discriminative capacity (inter-item correlation) and their level of homogeneity (internal consistency).

Keywords: Item analysis, Likert-type items, Inter-item correlation, Internal consistency.

Received: Apr 10, 2016
Accepted: Jul 15, 2016
Published: Jan 10, 2017

Un *Instrumento Psicométrico*, es una medida objetiva y estandarizada, compuesto por un conjunto de tareas, preguntas, estímulos, frases o situaciones que son presentadas a un examinado, con el objetivo de describirlo a partir de sus respuestas. El puntaje que se obtiene intenta poner de relieve una muestra de comportamiento representativa de la variable que se requiere medir (Anastasi y Urbina, 1998; Cortada de Kohan, 1999).

Independientemente de la naturaleza de la variable que se pretenda medir, los instrumentos psicométricos deben cumplir con ciertos requisitos que permitan, a partir de las puntuaciones derivadas de su medición, representar con mayor fidelidad el nivel del constructo que posee la persona evaluada. Explica Tavella (1978), que por ello los constructores se preocupan en demostrar que sus instrumentos miden realmente aquello para lo que ha sido concebido, que, aunque las puntuaciones contienen un margen de error tolerable, efectivamente los resultados permiten, describir o clasificar a los examinados según la variable evaluada. En este sentido, uno de los principales aspectos que debe cuidar el constructor de un instrumento psicométrico es el diseño de los indicadores empíricos (de los items) que son reflejo del constructo o variable a medir.

Los *items* o *reactivos* son las unidades básicas que componen el contenido fundamental de un instrumento de medición psicológica, independientemente del área al cual este dirigido. Los items, deben ser analizados con el propósito de conocer cómo se comporta cada uno de ellos dentro del instrumento psicométrico en construcción. Mientras se tengan reactivos de mayor calidad, el instrumento como totalidad estará dotado de mayor fuerza y será óptimo. Un item va a ser bueno dependiendo del marco de interpretación que dan: la conceptualización teórica de la variable a medir, el objetivo del instrumento y de la población a la cual está dirigido.

El primer paso a seguir para elaborar un instrumento psicométrico consiste en diseñar su tabla de especificaciones o esquema descriptivo. En este punto, se define la estructura teórica de la variable a medir, a través del establecimiento de las dimensiones y del número de items necesarios para evaluarla en la versión definitiva. Se debe decidir cuál es el tipo de item que es más apropiado para ser usado; en este sentido, la revisión de las fuentes documentales especializadas ha permitido determinar que existen diferentes tipologías para la clasificación de los reactivos. Uno de los tipos más utilizados en psicología es denominado *Item Multipunto, Politómico o Tipo Likert*.

Este modelo presenta al examinado un enunciado general en forma de frases, afirmaciones o situaciones, acompañado de una serie ordenada de categorías (como una escala). Posteriormente se solicita al evaluado que, entre las categorías presentes, escoja la opción que mejor lo identifique o describa, realizando así una valoración graduada de la conducta o rasgo medido en los enunciados. Morales, Urosa y Blanco (2003) plantean los reactivos que miden el mismo rasgo y, a partir de las respuestas dadas a los items, se obtiene una sumatoria que permite ubicar a los examinados a lo largo del continuo de la variable medida. Es por lo que los instrumentos que se construyen a partir de estos items se denominan *Escalas Sumativas*.

En este sentido, Domínguez (2013) explica que se asume que existe una equivalencia en la distancia entre las opciones que plantea la escala y la manifestación del constructo en la persona; es decir, si hay cuatro opciones de respuesta en la escala (*nunca, pocas veces, a menudo y siempre*) para cualquier enunciado, la distancia psicológica entre *nunca* y *pocas veces* sería la misma que entre *pocas veces* y *a menudo*.

Este tipo de items es utilizado comúnmente en aquellos instrumentos psicométricos cuyas variables evalúan el *comportamiento típico* de los individuos, como rasgos de personalidad, intereses, actitudes, etc. En estos casos, no existen respuestas correctas ni incorrectas, ni hay una connotación de rendimiento o de aprendizaje, solo se busca describir cómo es la persona "normalmente", saber cuál es su opinión, cómo son los sentimientos que predominan, etc.

Luego de haber seleccionado el tipo de items a diseñar, se debe construir un mayor número de reactivos de los necesarios para la versión definitiva del protocolo (el doble o triple). Inicialmente son analizados cualitativamente por un conjunto de expertos teóricos en la variable, los cuales van a valorar qué tan congruentes son los reactivos con las especificaciones teóricas propuestas, es decir, si los mismos pertenecen o no a la dimensión asignada por el constructor. Aquellos items que sean considerados congruentes por la mayoría de los expertos, pasan a la fase de *análisis cuantitativo*.

Con el número de items obtenido anteriormente, se procede a realizar una aplicación piloto de la versión experimental del instrumento. Martínez, Hernández y Hernández (2014) explican que, para que los resultados de los análisis de los items del estudio piloto sean útiles, la muestra debe ser representativa de la población a la que se destina el instrumento psicométrico,

ya que todos los estadísticos de los items dependen en gran medida de las características de la muestra.

El análisis cuantitativo consiste en la utilización de estadísticos que permitan describir cómo se comportan los items a partir del resultado obtenido en cada procedimiento. Se seleccionan aquellos reactivos que se aproximen al valor ideal establecido teóricamente para cada estadístico, de acuerdo al objetivo del instrumento, la población a la cual se dirige y a la tabla de especificaciones.

Análisis cuantitativo de items multipunto

A continuación, se presentan los pasos a seguir para realizar un analisis cuantitativo para un instrumento psicométrico compuesto por items multipuntos.

Revisión de los protocolos de evaluación:

Luego de haber aplicado la versión experimental del instrumento se debe contar con los protocolos contestados por los examinados. En este momento, se debe verificar que el proceso de estandarización durante la aplicación haya sido el correcto, que se cuenten con formularios completados de forma adecuada, es decir, que tengan datos de identificación y que los items hayan sido contestados de forma correcta. Esta revisión inicial permite constatar que los datos que se van a cargar son de calidad y cumplen con las condiciones mínimas para poder proceder a la revisión estadística. Se recomienda numerar los formularios en físico y que dicha numeración corresponda con la utilizada dentro de la matriz de datos, lo cual es oportuno de ser necesario regresar a revisar el protocolo para la verificación de algún dato.

Construcción de la matriz datos:

Para poder llevar a cabo los diferentes análisis estadísticos, se debe construir una matriz de datos con las respuestas de los examinados para cada uno de los items del instrumento psicométrico; en las filas se van a colocar a los individuos y en las columnas cada uno de los reactivos. Se puede utilizar para el procesamiento de la información programas como Excel o algunos más especializados como SPSS. Se debe cargar en la matriz la opción seleccionada para cada reactivo, asignando el puntaje que corresponda según su selección.

Por ejemplo, si está evaluando una variable *felicidad*, y se tienen las siguientes opciones de respuesta: *nunca, pocas veces, a menudo y siempre*, y el enunciado es el siguiente "*Estoy alegre constantemente*", claramente la direccionalidad del item es directa, es decir, valores altos van a indicar mayor presencia del atributo felicidad en la persona y van a corresponder con la respuesta, comportamiento o evidencia "típica" del rasgo, mientras que valores bajos corresponden con menor presencia del rasgo. Las opciones se codificarían de la siguiente manera: *nunca*= 1, *pocas veces*= 2, *a menudo*= 3 y *siempre*= 4 (Céspedes y Tristán-López, 2014). Ahora bien, si el reactivo fuera "*Estoy triste constantemente*", la direccionalidad de item sería inversa, lo que implicaría que aquellas puntuaciones más altas se corresponderían con baja presencia del constructo de interés, y por tanto, se codificarían en la matriz de datos de la siguiente manera: *nunca*= 4, *pocas veces*= 3, *a menudo*= 2 y *siempre*= 1.

Si bien es cierto que durante la aplicación se debe promover que los examinados no dejen de contestar algún item o que marquen dos opciones en una de las respuestas, se debe tener una codificación especial para dichas omisiones o errores al contestar y se denonomina con la etiqueta "No Contesta" o "No Aplica".

Procesamiento estadístico de la información:

Según Tavella (1978), los procedimientos para analizar un instrumento psicométrico pueden variar en mayor o menor medida según la naturaleza de la variable a medir (rendimiento, aptitud, actitud, etc.), pero todos suponen un análisis estadístico de parámetros que en la psicometría se utilizan comúnmente, lo cual permite utilizar recursos estadísticos y de interpretación para asegurar la calidad del instrumento psicométrico en construcción. A continuación, se van a referir los estadísticos que son apropiados utilizar cuando se realiza un análisis para items multipunto de los reactivos

Análisis de las opciones de respuesta

Los items multipuntos pueden presentar un recorrido variado en el continuo de las opciones de respuestas, por ejemplo, puede ir de 1 hasta 4 o del 1 hasta 7, la cantidad de valores seleccionada por el constructor del instrumento. En este sentido, explican Morales y cols. (2003) que, en la medida que se tenga más números de opciones, el instrumento va a adquirir más confiabilidad y por tanto se va a lograr una mejor descripción del individuo. Los autores

hacen la salvedad de que el número de opciones no debe superar la capacidad discriminativa de los examinados en el constructo evaluado y recomiendan que sean tres valores el número mínimo de opciones y el número máximo entre seis y siete.

Partiendo de lo anterior resulta oportuno analizar cómo se comportan las opciones de respuesta. Uno de los primeros aspectos a revisar es cuántas personas seleccionan cada una de las opciones de respuesta y este reporte se hace en términos *porcentuales*. Cabría esperar que esta inspección inicial permita conocer cuáles son las opciones que están siendo escogidas por la mayoría de las personas.

Un item ideal sería aquel cuyas opciones centrales sean seleccionadas por la mayoría de las personas, y las extrema, por un menor número de examinados. Como se observa en la Tabla 1, las opciones *Nunca* y *Siempre*, por ser las extremas, han sido seleccionadas por un menor porcentaje de personas (10% cada una), mientras que, las opciones centrales *Pocas Veces* y *A menudo*, han sido escogidas de forma equitativa por la misma cantidad de personas (40% cada una).

Tabla 1
Análisis de las opciones de respuestas del item 1.

	Item 1	
	Frecuencia	**Porcentaje**
Nunca	20	10
Pocas Veces	80	40
A Menudo	80	40
Siempre	20	10
Total	100	100

Nota: Tabla de elaboración propia.

Un item susceptible de ser rechazado sería aquel cuyas opciones de respuestas no son escogidas por el mismo porcentaje de personas, en especial cuando son los valores extremos los que presentan un mayor porcentaje de selección, ya que este patrón de respuestas no permite diferenciar adecuadamente a los examinados.

Igualmente, se puede utilizar la *Media* del item ($\bar{x}$) como un estadístico que permite describir las opciones de respuestas. La media aritmética hace referencia a cuál es el número de opciones seleccionado en promedio por los examinado; se espera tener una media cercana a los rasgos centrales de puntuaciones. Siguiendo el ejemplo anterior, se tienen cuatro opciones de respuesta que van desde 1, que significa *Nunca*, hasta 4, que significa *Siempre*. Un item cuyos valores se encuentren entre 2 y 3 (cercano al centro teórico que es 2,5) sería ideal, mientras que, cuando el reactivo presenta una media cercana a las opciones extremas, el item podría estar fallando al detectar ciertos valores del constructo (DeVellis, 2003). Normalmente, los items con medias muy cercanas a alguno de los extremos suelen tener valores bajos en la varianza y correlaciones bajas con otros items.

Otra manera de analizar cómo se comportan las opciones de respuestas para items multipunto es a través de la asimetría (α_3). Hernández, Fernández y Baptista (2010) exponen que ésta, se utiliza para conocer cuánto se parece la distribución de datos a la distribución teórica curva normal; asimismo, plantean que constituye un indicador del lado de la curva donde se agrupan las frecuencias. Cuando adquiere valores cercanos a cero, implica que es un item que tiene una distribución simétrica, donde los valores centrales son los que han sido seleccionado por la mayoría de las personas. Cuando la forma de la distribución es marcadamente asimétrica, la media suele ubicarse hacia las opciones extremas.

La distribución del item puede presentar asimetría negativa cuando la mayoría de las observaciones se ubican hacia la derecha de la curva, es decir, por encima de la media (Ver Figura 1). Mientras que, será asimetría positiva, cuando la mayoría de las opciones de respuestas se encuentran hacia el extremo izquierdo de la curva, es decir, por debajo de la media (Ver Figura 2).

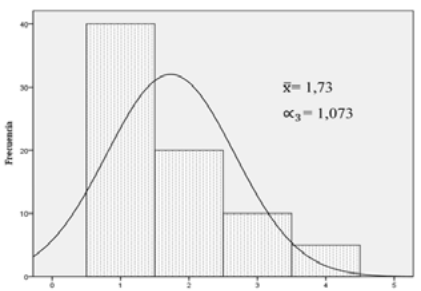

Figura 1. Asimetría Positiva

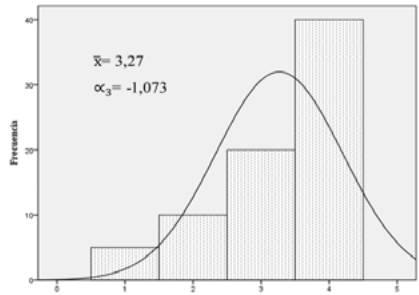

Figura 2. Asimetría Negativa

Como se puede observar, a través de la utilización de un histograma, se recoge los puntos de análisis previamente explicados. Puede apreciarse que, un item óptimo será aquel donde la forma de la distribución se asemeje a una curva normal (Ver Figura 3), es decir, aquél cuyas opciones centrales sean las que presenten mayor porcentaje de selección, su media gire en torno al centro teórico y, por tanto, tenga una asimetría cercana a cero. Cuando las opciones de respuestas son escogidas de forma desigual, la distribución del reactivo se aleja de esta representación y se observa un item que está fallando para este tipo instrumento.

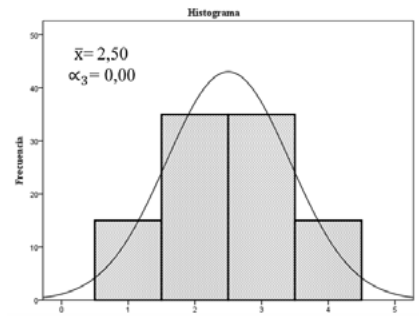

Figura 3. Asimetría de un Item Ideal

Capacidad Discriminativa

Un instrumento psicométrico tiene como objetivo principal poder identificar las diferencias entre los individuos evaluados, es decir, poder saber con certeza objetiva la cuantía del rasgo presente en los examinados. Para lograrlo es necesario que el instrumento esté compuesto por items que diferencien entre las personas; en la medida que los indicadores logren este cometido, se tendrá un instrumento psicométrico que también lo alcance. Lo anterior implica que se deben seleccionar reactivos que tengan la capacidad

de diferenciar o discriminar, en términos estadísticos, entre los grupos de examinados. El item ideal será aquel que distinga entre los individuos que tienen mayor presencia del rasgo en comparación con aquéllos que tienen menor presencia.

Anastasi y Urbina (1998) explican que existen alrededor de cincuenta índices que proporcionan información sobre la capacidad discriminativa de los items. La selección de alguno de ellos depende principalmente del nivel de medida de la variable (dicotómica o continua) y de qué aspecto estadístico del item el constructor está interesado en describir (la capacidad discriminativa del item como elemento, o en conjunto con otras variables). No obstante, a pesar de las diferencias en el procedimiento y las suposiciones de cada uno, la mayor parte los índices que miden la capacidad discriminativa de los items arrojan información similar en cuanto a cuáles son los mejores reactivos para seleccionar (Oosterhof, 1976). Cuando se quiere analizar si los examinados tienden a responder coherentemente, de manera que se pueda deducir que todos los reactivos expresan el mismo rasgo, los procedimientos que se suelen utilizar son la correlación inter-item (r_{ii}), la correlación item-total (r_{iT}) y el contraste de medias de los grupos extremos.

A continuación, se van a presentar aquellos indicadores de capacidad discriminativa que son reportados en la literatura como los más frecuentes para analizar los reactivos multipunto.

a) Capacidad discriminativa de un item

En primer lugar, resulta oportuno hacer una revisión de la capacidad discriminativa del item como elemento único, y a tal efecto se puede calcular la desviación típica, la varianza o la kurtosis. Estos estadísticos informan el grado en que un reactivo diferencia correctamente entre los examinados con relación al rasgo del instrumento que se pretende medir.

La varianza (s^2) y desviación típica (s), estadísticamente, se definen como una medida de dispersión de los puntajes en torno a la media de los reactivos. Ambos estadísticos se interpretan de la misma manera, aunque se expresan en magnitudes distintas. La varianza se obtiene al elevar al cuadrado la desviación típica, por lo cual el valor obtenido siempre es positivo, haciendo más difícil interpretar el significado del valor de la varianza. La desviación típica está expresada en la misma unidad de la variable (segundos, respuestas correctas, grado de acuerdo), por lo cual suele existir cierta preferencia por los autores en utilizar la desviación típica para analizar los items. Morales

(2008), plantea que la desviación típica describe el grado de homogeneidad de los reactivos; añade que será más baja en la medida en que la respuesta al item esté más próximo a la media, es decir, se disperse menos del promedio y aumentará si hay puntuaciones extremas muy alejadas de la media. Cuando todas las personas seleccionen la misma opción de respuesta, la desviación típica será de 0.

En la literatura, no existe un valor referencial que permita analizar la calidad del item en términos de la deviación típica. En general, se interpreta cuando se compara con el resultado obtenido entre los reactivos. Dos desviaciones típicas pueden compararse entre sí directamente si provienen de datos comparables (unidades comparables, es decir, la misma escala métrica).

Items con grandes desviaciones estándar son deseables, ya que indica que los examinados usaron la escala de respuesta completa, mientras que una más pequeña (cerca de cero) indica que los evaluados respondieron el item de manera similar. Por ejemplo, una gran mayoría seleccionó "*Nunca*", en este caso, los items del instrumento psicométrico no colocan a los examinados a lo largo del continuo del constructo, por tanto, no permite identificar las diferencias individuales.

Morales (2008), explica que en ocasiones puede ser de utilidad conocer cuál es el valor máximo posible de la desviación típica; este valor máximo posible es igual a:

$$s_{i(\max)} = \frac{puntuación\ máxima\ posible - puntuación\ más\ baja\ posible}{2}$$

Por ejemplo, si se tiene un item multipunto con cuatro opciones de respuesta, que van desde 1, que significa *Nunca*, hasta 4, que significa *Siempre*, la puntuación máxima posible es 4 y la puntuación más baja posible es 1; en este caso, la desviación típica máxima posible es:

$$s_{i(\max)} = \frac{4-1}{2} = 1{,}5$$

Se plantea una progresión aritmética entre el número de opciones y la variabilidad máxima, partiendo de que el mínimo de opciones que puede tener un item es dos y su variabilidad máxima es $s_{i(max)} = 0{,}5$. Se considera que, por cada unidad que aumenten las opciones de respuesta, va a aumentar media unidad la variabilidad máxima, como se plantea en la Tabla 2.

Tabla 2
Variabilidad Máxima según el número de opciones de los items.

Número de Opciones	2	3	4	5	6	7
Variabilidad Máxima	0,5	1	1,5	2	2,5	3

Nota: Tabla de elaboración propia.

Morales (2008) considera que esta referencia suele ser poco útil, ya que este valor máximo es difícilmente alcanzable en la mayoría de las situaciones, aunque hace la salvedad que en el caso de reactivos binarios, sí es más frecuente que la desviación típica obtenida sea la mayor posible o se aproxime mucho a la mayor posible.

Otra medida de dispersión, menos utilizada, es el *coeficiente de variación* (CV), también denominado variabilidad relativa. Consiste en dividir la desviación típica entre la media. Es independiente de la unidad de medida y dicho coeficiente es habitual que sea multiplicado por 100. Mientras más cercano a 100, mayor es el nivel de variabilidad y viceversa.

$$CV = \frac{S_i}{\bar{x}} \times 100$$

Mide los cambios en la desviación típica en comparación con la media. Se debe tener cuidado cuando la media esté influida por la variable (como pruebas de rendimiento máximo). Podría utilizarse como un indicador para realizar comparaciones y verificar cuales items tienen mayor capacidad discriminativa (Morales, 2008).

La capacidad discriminativa de un item también puede verse reflejada en el cálculo de la Kurtosis (α_4), la cual hace referencia al grado de curvatura (o achatamiento) de la distribución en comparación con la distribución normal (Lezama y Urdanibia, 2009). Se analiza la kurtosis de la distribución del item según ésta sea más o menos apuntada que la distribución de la curva normal. Se dice que la distribución es mesokúrtica cuando la distribución del item se comporta como una curva normal, en este caso, el valor de la kurtosis será de cero. Cuando la distribución del item está más apuntada que la de la curva normal, se denomina leptokúrtica y los valores de la kurtosis son positivos. Mientras que, cuando es aplanada, tendiendo a ser menos apuntada que la distribución normal, se denomina platikúrtica y los valores de la kurtosis son negativos. En el caso de los items multipunto, un item ideal será aquél cuyo valor de kurtosis esté cercano a cero, es decir, que su distribución sea cercana a la curva normal. Por ejemplo, al analizar los

items presentes en la Tabla 3, se puede observar que el reactivo 20 presenta una mejor capacidad discriminativa dado que su kurtosis se acerca más a 0.

Tabla 3
Ejemplo de los tipos de Kurtosis.

	Kurtosis	
	Estadístico	**Tipo de Kurtosis**
Item 1	-,533	Platikúrtica
Item 14	,726	Leptokúrtica
Item 20	,025	Tiende a Mesokúrtica

Nota: Tabla de elaboración propia.

Para probar la calidad de los items, se suele seleccionar la media y la desviación típica o la asimetría y la kurtosis. Resulta poco parsimonioso utilizar todos los estadísticos para describir el comportamiento del item de forma individual. Suele existir cierta correspondencia en los resultados, puesto que, en general, si todos los individuos responden a un item dado de manera idéntica, éste no discriminará entre los individuos con diferentes niveles del rasgo que se mide, es decir, no existirá variabilidad. En contraste, si las respuestas son diversas respecto al atributo de interés, entonces el rango de puntajes obtenidos para un reactivo también debe ser diverso, lo cual implica alta variabilidad (DeVellis, 2003).

b) Correlación inter-item

Una forma de conocer qué tanto discriminan los items, consiste en verificar cómo es la relación entre un par de reactivos, a través del análisis de sus inter-correlaciones. Las correlaciones entre reactivos examinan el grado en que el puntaje de un item está relacionado con la puntuación obtenida en otro de los reactivos que componen la escala. Para Cohen y Swerdlik (2005) esta propiedad proporciona una evaluación de la redundancia de items, es decir, en qué medida los reactivos de un instrumento psicométrico evalúan el mismo contenido. Asimismo, DeVellis (2003) expresa que, para alcanzar items altamente intercorrelacionados, cada elemento individual debe correlacionarse con la colección de reactivos restantes. Si todos los items están midiendo el mismo dominio, se esperaría que todos correlacionen

bien. Aquellos items que tengan correlaciones consistentemente bajas podrían requerir ser eliminados para disponer de un cuestionario óptimo.

Cuando se calcula la correlación inter-item, se debe considerar el tipo de reactivo y el nivel de medida que le corresponde. Si se van a analizar items binarios, lo correcto es utilizar la correlación *Phi* (ϕ_{ij}), mientras que, para los items multipunto, lo común es aplicar la *Correlación Producto Momento de Pearson* (r_{ij}). En ambos casos se obtiene un rango de valores que va entre -1 y +1, el cual es un indicador de la variación conjunta del par de items, es decir, cómo es el patrón de comportamiento de ambos reactivos. El resultado esperado va a depender de la asociación teórica, por ejemplo, en el caso de dos items que pertenezcan a una misma dimensión, lo deseable es que su relación sea alta y positiva, mientras que, si los items son de dimensiones teóricamente antagónicas, el resultado debería ser moderado y negativo. Se espera que el comportamiento dado en item *i*, pueda decir algo de cómo se comportará el item *j*, debido a que por ser parte de la misma dimensión comparten información del constructo a medir, en aquellos casos donde las dimensiones son heterogéneas, se puede dar el escenario en el cual el nivel de asociación sea menor.

Rust y Golombok (2013) exponen que mientras mayor sea la correlación, mayor capacidad discriminativa tiene el item. Los autores plantean que es ideal tener una asociación mayor a r_{ij}=.20, descartando aquellas asociaciones que sean: cercanas a cero, negativas cuando sean de una misma dimensión y aquellas que sean 1 o muy cercanas a este valor, ya que implican redundancia en el contenido. Estos autores explican que suele esperarse que la correlación promedio entre los reactivos oscile entre r_{ij}=.20 y r_{ij}=.40 lo que sugiere que, si bien los items son razonablemente homogéneos, contienen una varianza única suficiente como para no ser isomorfos entre sí. Por su parte, Ferketich (1991) plantea que, aunque no hay un criterio único y rápido para establecer el nivel de correlación, una regla de oro es que los reactivos que correlacionen debajo de r_{ij}=.30, no están suficientemente relacionados y, por lo tanto, no contribuyen a la medición de constructo y que los items que se correlacionan r_{ij}=.70 son redundantes y probablemente innecesarios.

Para analizar la correlación promedio entre los items se utilizan todos los reactivos del instrumento que se está diseñando para medir el constructo. Primero, se debe calcular la correlación entre cada par de elementos, como se ilustra en la Tabla 4. Por ejemplo, si se tienen cinco reactivos, se obtendrán diez emparejamientos de diferentes items, es decir, diez inter-correlaciones.

Tabla 4
Matriz de correlaciones inter-items.

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**
Item 1	1	,419*	,640*	,604*	-,410*
Item 2		1	611*	,454*	-,459*
Item 3			1	,580*	-,497*
Item 4				1	-,471*
Item 5					1

*Correlación estadísticamente significativa (p<0,05)

La correlación promedio entre los reactivos, es simplemente el promedio o la media de todas estas correlaciones. En el ejemplo, se obtuvo una correlación promedio entre todos los reactivos de $r_{i\text{-}i}$=.147 . Como se observa, instrumentos con muchas inter-correlaciones negativas van a generar un valor promedio muy bajo, por lo tanto, los reactivos que correlacionen negativamente con sus pares serán aquellos reactivos con muchas inter-correlaciones negativas y, en consecuencia los candidatos principales a ser eliminados. Cuando se desincorpora el item 5, la correlación promedio inter-item mejora pasando a $r_{i\text{-}i}$=.551 y sus correlaciones oscilan entre $r_{i\text{-}i}$=.419 y $r_{i\text{-}i}$=.640.

De igual manera, se puede analizar cómo es el comportamiento promedio de un solo reactivo, por ejemplo, al desincorporar el reactivo 5, para el item 1 se encontró que: r_{12}=.419 ; r_{13}=.640; r_{13}=.604 y por tanto, su comportamiento promedio sería igual $r_{1\text{-}i}$=.554 y todas las inter-correlaciones son estadísticamente significativas (Ver Tabla 5).

Tabla 5
Promedio de las correlaciones inter-items.

	Media de las inter-correlaciones	**Correlación Mínima**	**Correlación Máxima**	**N° de items**
Correlaciones inter-items para la Escala Total	,147	-,497	,640	5
Correlaciones inter-items al Eliminar el item 5	,551	,419	,640	4
Correlaciones inter-item, para el item 1	,554	,419	,640	3

Nota: Tabla de elaboración propia.

El análisis y la aplicación de los baremos para la correlación inter-item se realizará sobre el resultado promedio obtenido de las correlaciones entre los items de todo el instrumento, los que componen una dimensión o de un solo item. Dentro de este análisis, se puede ver qué tanto mejora la relación promedio cuando dicho reactivo se descarta del modelo.

c) Correlación item-puntaje total de la escala (dimensión)

A través del análisis de la correlación item-puntaje total, se puede verificar si los items diferencian entre los examinados. Este procedimiento consiste en asociar las respuestas dadas a un reactivo con la sumatoria de todos los demás items de la escala (o dimensión). Cuando se analiza la relación entre el item y el puntaje total, se espera identificar si puntuar alto en un reactivo se corresponde con la alta presencia del rasgo que mide la escala.

Cuando se tiene un item multipunto y una variable continua que proviene de la sumatoria de todos los demás items, es apropiado utilizar la *Correlación Producto Momento de Pearson* (r_{ij}) . Se espera que el resultado de los coeficientes de correlación sea estadísticamente significativo y que presente valores entre moderado y alto, en especial, cuando la dimensión es homogénea. Puntuar alto en dicho item estará asociado con valores altos en el puntaje total de la escala (o dimensión), es decir, seleccionar las opciones de respuesta que mayor describan a la persona, estará asociado con una mayor presencia del rasgo. Los items con una mayor correlación con el puntaje total son los que tienen más en común y, por tanto, se puede pensar que miden lo mismo que los demás. Los items con correlaciones negativas, no significativas o muy bajas con respecto a las de los otros, suelen ser susceptibles a ser eliminados de la escala (Morales y cols., 2003). Cabe destacar, que no se esperan valores iguales a 1, porque esto implicaría redundancia en la escala. En este sentido, Nunnally y Bernstein (1994), explican que correlaciones menores a r_{ij}=.30 pueden ser considerados como items pobres.

Autores como DeVellis (2003) y Morales y cols. (2003), consideran que se debe realizar el análisis de la correlación eliminando el aporte del propio item, lo cual se logra mediante la *correlación item-puntaje total corregida* ($r_{i(T-i)}$). La inclusión del reactivo en la "escala" puede "inflar" el coeficiente de correlación. Cuanto menor sea el número de reactivos en el conjunto, mayor será la diferencia que hará la inclusión o exclusión del item bajo escrutinio. En general, es aconsejable examinar la correlación corregida entre el item y

el total. En programas como SPSS se puede obtener directamente este valor para su análisis.

Por ejemplo, en la Tabla 6 se observa que el reactivo 3 presenta una correlación alta, positiva y significativa con el puntaje total de la escala ($r_{1\text{-}t}$=.667; p <.05) lo que quiere decir que aquellas personas que se encuentran identificadas con el enunciado del item (*Siempre, A menudo, Totalmente de Acuerdo, etc.*) serán las que presenten en mayor medida el constructo. Mientras que, aquellos que no perciban cuáles items los representan (*Nunca, Pocas veces, Insatisfecho, No me describe, etc.*) serán los que presenten en menor medida el rasgo. Es decir, que el patrón de respuestas del item 3, permite considerarlo como un reactivo con una adecuada capacidad discriminativa. Al ser sus respuestas estadísticamente significativas, se puede considerar que el resultado es producto de la información que comparten ambas variables y que las hacen covariar altamente, y no debido al azar.

Por otra parte, se puede observar que el reactivo 5 presenta una asociación moderadamente alta y negativa, lo cual implica que aquellos examinados que se identificaron con las respuestas de mayor acuerdo en el reactivo (*Siempre, A Menudo*), son quienes presentan en menor medida el rasgo y viceversa. Este escenario es el esperado si se analiza un reactivo inverso (sin recodificar), o un item que sea de una dimensión contraria. Si este escenario se presenta en un item directo de la dimensión o de la escala total, será valorado como un item de poca calidad.

Tabla 6
Correlación Item-total corregida.

	Correlación item-total corregida
Item 1	,607**
Item 2	,489**
Item 3	,667**
Item 4	,567**
Item 5	-,564

**Correlación estadísticamente significativa (p<0,05)

d) Grupos de Contraste

Autores como Anastasi y Urbina (1998), Morales y cols. (2003) y Hogan (2004), plantean que la construcción de grupos de contraste es una técnica válida para analizar la capacidad discriminativa de los items. Partiendo de la premisa de que el instrumento en su totalidad es un indicador válido del rasgo, se construyen grupos que tienen diversidad del rasgo, luego se identifican quienes recibieron una puntuación elevada en el puntaje total en la escala y aquéllos que recibieron una baja puntuación, para después determinar hasta qué punto un reactivo particular diferencia los que obtuvieron en mayor medida el rasgo de los que la presentan en menor medida, es decir, qué tanto el item distingue en función del atributo medido.

El grupo con *Alta Presencia* del rasgo está compuesto por los examinados que se ubican en los porcentajes superiores de la escala (el 25% de los mejores, el 27% de los que tienen en mayor presencia el rasgo, etc.), Mientras que, el grupo con *Baja Presencia* del rasgo serán aquéllos ubicados en el 25%, 27% o 33% del menor puntaje en la escala total (o dimensión). Se suele utilizar *la prueba de contraste de medias t de student*, para comparar la media del grupo con Alta Presencia y Baja Presencia. Si resulta estadísticamente significativa la diferencia, quiere decir que el item efectivamente puede diferenciar en función del atributo, por lo tanto, es un reactivo apto para pertenecer a la escala definitiva. En cambio, si no es estadísticamente significativa la diferencia entre los grupos contrastados, el item no distingue correctamente entre los evaluados, por lo que resulta conveniente eliminarlo de la escala.

Para analizar la capacidad discriminativa de los items, se construyeron dos grupos de comparación Alta Presencia del Rasgo (desde el cuartil 75 hacia los valores mayores) y Baja Presencia del Rasgo (desde el cuartil 25 hacia los valores menores), se espera que como cada grupo tiene variedad de presencia del rasgo, el item discrimine entre cada uno de ellos. Se utilizó la prueba *t de student* para comparar ambas muestras (Ver Tabla 7). Se encontró que todos los reactivos diferencian entre aquéllos que obtuvieron en mayor medida el rasgo y los que lo presentan en menor medida, es decir, que el item distingue en términos del constructo medido. Por ejemplo, el item 1 obtuvo el siguiente resultado: $t_{(174)}$=-15,996; p<.05. Esto quiere decir que existen diferencias estadísticamente significativas entre el grupo que presenta alta presencia y el grupo de baja presencia, por lo tanto, el item discrimina correctamente, lo cual es un comportamiento óptimo para el reactivo.

Tabla 7
Prueba t de student para cada item.

		Prueba T para la igualdad de medias						
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Inferior	Superior
Item 1	Se han asumido varianzas iguales	-15,996	174	,000	-1,646	,103	-1,849	-1,443
Item 2	Se han asumido varianzas iguales	-13,153	174	,000	-1,646	,125	-1,893	-1,399
Item 3	Se han asumido varianzas iguales	-17,078	174	,000	-2,048	,120	-2,284	-1,811
Item 4	Se han asumido varianzas iguales	-22,190	174	,000	-2,293	,103	-2,497	-2,089
Item 5	Se han asumido varianzas iguales	13,477	174	,000	1,525	,113	1,302	1,748

Nota: Tabla de elaboración propia.

Morales y cols. (2003) plantean que tanto la correlación item-total como el contraste de grupos, arrojan información semejante; un item que diferencia adecuadamente a los grupos extremos claramente tiene una alta relación con el total de la escala. Los autores afirman que con los dos procedimientos se llega a la misma selección de items, así que usar uno u otro m~~étodo depende~~ del interés de quien construye el instrumento.

Consistencia Interna

Al definir el dominio del constructo a medir, cada reactivo es una muestra individual del área a evaluar. En este sentido, el análisis de la consistencia interna permite determinar el grado en que los diferentes items miden el mismo rasgo. Esta revisión permite conocer qué tanto los reactivos del instrumento en construcción son iguales en términos de lo que mide, es decir, qué tanto están correlacionados. Se basa en la consistencia de las puntuaciones a todos los reactivos del instrumento. Si el puntaje de los reactivos que constituyen el instrumento tiene correlaciones positivas entre ellos se dice que es un instrumento homogéneo, es decir, que el contenido evalúa el mismo rasgo (Brown, 1980; Kaplan y Saccuzzo, 2006).

Para el caso de los items multipunto, se debe aplicar el *Alfa de Cronbach* (α), el cual adquiere valores entre 0 y 1. Mientras más cercano a 1, implica que el instrumento es más homogéneo y mientras más cercano a 0, el domino

a evaluar resulta más heterogéneo. Si el alfa es negativa, significa que existe algún tipo de falla, probablemente, debido a la presencia de correlaciones negativas (o covarianzas) entre los items. Si esto ocurre, intente recodificar los reactivos inversos o eliminar los items directos que presenten altas inter-correlaciones negativas. Existen varios programas estadísticos que permiten computar el valor del alfa (SPSS, SAS, R). Otra opción para calcular el alfa es hacerlo manualmente a través de la siguiente formula: $\alpha = \frac{n}{n-1} \cdot (\frac{s_t^2 - \sum s_i^2}{s_t^2})$, donde n es el número de items que compone la escala, s^2_t es iguala la varianza del instrumento total y Σs^2_i es la sumatoria de la varianza de los items.

En cuanto a los valores óptimos de consistencia interna, Nunnally (1978) considera que el punto crítico sería a partir de α= 0,70. Por otra parte, DeVellis (2003) desarrolla ciertos parámetros para en análisis de la consistencia interna (Ver Tabla 8). En este orden de ideas, Morales y cols. (2003), añaden que, el valor de la consistencia interna esperado va a depender del uso previsto del instrumento, por ejemplo, para investigaciones básicas, se puede exigir un valor mínimo de α= 0,60, pero para tomar decisiones o selección de personal, recomienda exigir valores mayores de α= 0,80. No obstante, un punto a considerar es el muestreo de contenido que conforma la dimensión o instrumento total, ya que, si originalmente se espera que el mismo mida aspectos similares, de debe apuntar a tener un alfa elevado, mientras que, si el dominio tiende a ser más heterogéneo, lo adecuado es encontrar valores modestos de la consistencia interna.

Tabla 8
Valoración de la consistencia interna según DeVellis (2003).

Valor del α	**Etiqueta**
Mayor a α = 0,90	Acortar la Escala por ser redundante
α= 0,80 y α= 0,90	Muy bueno
α= 0,70 y α= 0,80	Respetable
α= 0,65 y α= 0,70	Mínimamente Aceptable
α= 0,60 y α= 0,65	Indeseable
Menor a α= 0,60	Inaceptable

El análisis presentado anteriormente para revisar la calidad de los items y los posibles comportamientos inadecuados de los mismos (valores de la media no central, poca variabilidad, correlaciones negativas entre los items, correlaciones bajas item-puntaje total y correlaciones débiles entre los items) tenderán a reducir la consistencia interna. Por lo tanto, después de haber seleccionado los mejores reactivos, eliminando aquéllos que son pobres y manteniendo los buenos, el cálculo de la consistencia interna es una forma de evaluar el éxito logrado (DeVellis, 2003).

Algunos programas de computación como SPSS permiten verificar cómo se ve afectado su valor al eliminar un reactivo, por ejemplo, al analizar la consistencia interna de los cinco reactivos analizados en la Tabla 5, se puede observar que su consistencia interna es moderada α= 0,522 debido a que uno de los reactivos presenta altas correlaciones negativas. En la Tabla 9, se observa que ocurría si se desincorpora alguno de los reactivos, por ejemplo, al eliminar el reactivo 5, el valor de la consistencia interna aumenta α= 0,829, mientras que, si se elimina el item 3, su consistencia interna disminuiría considerablemente α= 0,163, es decir, que no sería conveniente eliminarlo de la escala total. Además, a través de este procedimiento, se puede obtener información de cómo se ve afectada la media y la varianza de la escala total cuando se elimina un item, la correlación item-total corregida y la correlación múltiple al cuadrado, la cual indica qué tanto de la variabilidad del reactivo está explicando la variabilidad de la escala total.

Tal como explica Morales y cols. (2003), este no es un procedimiento mecánico. Los items se deben ir suprimiendo de uno en uno o en pequeños bloques, observar qué ocurre con la combinación de reactivos que van quedando y cómo se ve afectada la consistencia interna de la escala. La idea final, es mantener los items que conformen un instrumento excelente.

Tabla 9
Estadísticos de la Escala cuando se desincorpora un reactivo.

	Estadísticos total-item				
	Media de la escala si se elimina el item	Varianza de la escala si se elimina el item	Correlación item-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se elimina el item
Item 1	13,37	5,196	,607	,492	,245
Item 2	13,86	5,335	,489	,412	,320
Item 3	13,49	4,557	,667	,583	,163
Item 4	13,47	5,636	,567	,463	,294
Item 5	13,49	11,773	-,564	,325	**,829**

Nota: Tabla de elaboración propia.

Construcción de baremos:

Para poder decidir cuáles son los reactivos que van a conformar la escala definitiva, el constructor del instrumento debe establecer cuáles son las pautas esperadas del comportamiento de los reactivos en cada uno de los estadísticos seleccionados. Para dicho análisis se va a tomar en cuenta lo propuesto por Cohen y Swerdlik (1999) quienes explican que para la construcción de un instrumento psicométrico, se debe considerar los aspectos internos que lo caracterizan, viendo cómo se comporta la variable, a quién está dirigido y cuál es el objetivo para lo cual se diseña. Por ejemplo, se debe conocer si la variable y sus dimensiones son homogéneas o heterogéneas o si se necesita que los reactivos redunden en consistencia interna.

Como se ha ido observando a lo largo de esta entrega, existen diversos estadísticos, que arrojan información distinta acerca del comportamiento de los reactivos. No obstante, en la bibliografía difícilmente se encontrará información de un comportamiento prototípico que aplique exactamente igual para todos los instrumentos psicométricos y que permitan de antemano tener definido como serán los reactivos que se deben escoger. La selección de los reactivos requiere la construcción de un modelo de análisis particular que refleje el comportamiento ideal de los items, vinculando los números obtenidos en los estadísticos y el comportamiento teórico de la variable.

Lezama y Urdanibia (2009) plantean la pertinencia de construir un baremo que incluya los estadísticos que se van a utilizar y se asignen pesos diferenciales al recorrido de valores que asume cada estadístico, no solo considerando el comportamiento ideal, sino también valorando el desempeño inadecuado de los estadísticos. Por ejemplo, si se tiene una variable homogénea, se espera que valores altos de la correlación item-dimensión sean puntuados con mayor peso, mientras que correlaciones negativas y cercanas a cero recibirán una menor calificación en el baremo.

Es importante resaltar que no se debe elegir un item solo por un estadístico; la decisión de si va a pertenecer o no la escala definitiva depende de cómo se comporte en la mayoría de los estadísticos, del objetivo del instrumento, de la población y de su tabla de especificaciones.

DISEÑO DE LA ESCALA DEFINITIVA

Luego de haber realizado el análisis estadístico de los items, se espera que el producto sea la selección de los mejores reactivos. En ocasiones, se deben hacer ajustes o modificaciones, que pasan por la revisión de redacción o del contenido para lograr items de mayor calidad. Se debe recordar que los items que constituyan el modelo definitivo de la escala deben responder a lo establecido en la tabla de especificaciones. No debe ocurrir que alguna dimensión quede sin item o con menor cantidad de lo establecido, ya que ello afectaría la validez del instrumento.

Cuando los items presentan muchos ajustes, es recomendable realizar una segunda aplicación para verificar que los arreglos realizados efectivamente están generando los efectos deseados. Sobre esta segunda aplicación se puede realizar otros análisis estadísticos multivariantes como el Análisis Factorial, para verificar la calidad de los reactivos (Hair, Tathan y Black, 1999; Richaud, 2005). Cuando los ajustes son mínimos, se puede pasar a la siguiente etapa para comprobar la validez y confiabilidad de la escala.

Referencias bibliográficas

Anastasi, A. & Urbina, S. (1998). *Test Psicológicos*. (7ma Ed.). México: Prentice Hall.

Brown, F. (1980). *Medición en Psicología y Educación* (4ta Ed.). México: El Manual Moderno.

Céspedes, V. & Tristán-López, A. (2014). Influencia de la direccionalidad de los items en los resultados de instrumentos de medición. *Diversitas*, 10(1), 29-43.

Cohen, R. & Swerdlik, M. (2001). *Pruebas y Evaluación Psicológica. Introducción a las pruebas y a la Medición*. (4ta Ed.). México: McGraw Hill.

Cortada de Kohan, N. (1999). *Teorías psicométricas y construcción de tests*. Buenos Aires: Lugar Editorial.

DeVellis, R. (2003). *Scale Development*. (2da Ed.). EE. UU: Sage Publications Inc.

Domínguez, S. (2013). ¿Items politómicos o dicotómicos? Un estudio empírico con una escala unidimensional. *Revista Argentina de Ciencias del Comportamiento*, 5(3), 30-37.

Ferketich, S. (1991). Focus on Psychometrics Aspects of Item Analysis. *Research in Nursing & Health*, 14, 165-168.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1999). *Análisis multivariante*. (5ta Ed.). Madrid: Prentice Hall.

Hernández, R., Fernández, C. & Baptista, P. (2010). *Metodología de Investigación*. (5ta Ed.). México: Trillas.

Hogan, T. (2003). *Pruebas Psicológicas*. México: Manual Moderno.

Kaplan, R. & Saccuzzo, D. (2006). P*ruebas Psicológicas. Principios, aplicaciones y temas*. (6ta Ed.). México: Thomson.

Lezama, L. & Urdanibia, A. (2010). *Análisis de Items y de la Prueba*. Caracas: Fondo Editorial. Facultad de Humanidades. Universidad Central de Venezuela.

Martínez, M., Hernández, M. & Hernández, M. (2014). *Psicometría*. Madrid: Alianza Editorial.

Morales, P. (2008). *Estadística aplicada a las Ciencias Sociales*. Madrid: Universidad Pontificia Comillas.

Morales, P., Urosa, B. & Blanco, A. (2003). *Construcción de Escalas de actitudes tipo Likert*. Madrid: La Muralla.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J. & Bernstein, I. (1996) *Teoría Psicométrica*. (3ra Ed.). México: McGraw Hill.

Oosterhof, A. (1976). Similarity of various items discrimination indices. *Journal of Educational Measurement*, 13(2), 145-150.

Richaud, M. (2005). Desarrollos del análisis factorial para el estudio de: item dicotómicos y ordinales. Interdisciplinaria: *Revista de Psicología y Ciencias Afines*, 22(2), 237-251.

Rust, J. & Golombok, S. (2014). *Modern Psychometrics*. (3ra Ed.). London: Routledge.

Tavella, N. (1978). *Análisis de los itemes en la construcción de instrumentos psicométricos*. México: Trillas.